https://doi.org/10.1093/bioinformatics/bty697

A quantitative map of protein sequence space for the cis-defensin superfamily

Thomas Shafee[1]*, Marilyn Anderson[1]
[1] La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia
* T.Shafee@latrobe.edu.au

Abstract

Motivation: The cis-defensins are a superfamily of small, cationic, cysteine-rich proteins, sharing a common scaffold, but highly divergent sequences and varied functions from host-defence to signalling. Superfamily members are most abundant in plants (with some genomes containing hundreds of members), but are also found across fungi and invertebrates. However, of the thousands of cis-defensin sequences in databases, only have a handful have solved structures or assigned activities. Non-phylogenetic sequence-analysis methods are therefore necessary to use the relationships within the superfamily to classify members, and to predict and engineer functions.

Results: We show that the generation of a quantitative map of sequence space allows these highly divergent sequences to be usefully analysed. This information-rich technique can identify natural groupings of sequences with similar biophysical properties, detect interpretable covarying properties, and provide information on typical or intermediate sequences for each cluster. The cis-defensin superfamily contains clearly-defined groups, identifiable based on their biophysical properties and motifs. The organisation of sequences within this space also provides a foundation of understanding the evolution of the superfamily.

Availability and Implementation: A webtool for exploring and querying this sequence space is hosted at TS404.shinyapps.io/DefSpace.

Contact: T.Shafee@latrobe.edu.au

Supplementary: Supplementary data available at the journal's web site

Keywords

Sequence analysis, non-phylogenetic, innate immunity, antimicrobial peptide, toxin, protein engineering, cysteine-rich protein

https://doi.org/10.1093/bioinformatics/bty697

Introduction

Defensins

Defensins are small, cationic, disulphide-rich proteins found in almost all tissue types across animals, plants, fungi (Shafee *et al.*, 2017). They are best known for their host-defence and antimicrobial activities, but members have also been recruited to a variety of additional specialised functions including ion channel blocking, enzyme inhibition and cell signalling (Zhu *et al.*, 2005; Van der Weerden and Anderson, 2013; Shafee *et al.*, 2017). Furthermore, the same function can be mediated by different mechanisms, even by closely related defensins (Vriens *et al.*, 2014; Bleackley *et al.*, 2016; Gao and Zhu, 2012).

The defensins actually consist of two superfamilies with independent evolutionary origins, the *cis*- and *trans*-defensins (Shafee, Lay, *et al.*, 2016). The two superfamilies are defined by their structural similarity, secondary structure order, and disulphide topology (Shafee, Lay, *et al.*, 2016). The most conserved feature of each superfamily is a pair of characteristic disulphides. In the cis-defensin superfamily, this disulphide pair is cis-oriented (Figure 1A), and sometimes called a cysteine-stabilised αβ fold. The trans-defensins have a trans-oriented pair of disulphides (Figure 1B), and include the defensins from humans and other vertebrates (Shafee, Lay, *et al.*, 2016; Shafee *et al.*, 2017). Extensive divergent evolution within each superfamily has elaborated this scaffold to form multiple families, typically defined by disulphide class (cysteine motif and disulphide connectivity), resulting in an array of markedly different overall structures. Adding to this, the term 'defensin' has sometimes also been used as a general functional descriptor of any protein with host-defence activity (Parisi *et al.*, 2018).

The *cis*-defensin superfamily is most characterised in plants, with thousands of members (sometimes hundreds in a genome) (Silverstein *et al.*, 2005). This number is growing rapidly as more genomes are sequenced (e.g. the 1KP project), and is likely an underestimate, since they are under-annotated by current methods (Silverstein *et al.*, 2007). Additional cis-defensins are also widely distributed across cellular life in animals, fungi and bacteria (Shafee *et al.*, 2017, 2018).

What are all these proteins doing? Few have been functionally characterised and, although they're often assumed to be antimicrobial, even similar sequences can have very different activities (Vriens *et al.*, 2014; Bleackley *et al.*, 2016; Gao and Zhu, 2012; Zhu *et al.*, 2014). Added to this, the sequences and structures of cis-defensins are extremely divergent (pairwise identities frequently <15% and typical length 35-60aa) due to their robust structures combined with strong selection pressures (Shafee, Robinson, *et al.*, 2016). This diversity poses a problem for traditional sequence analysis methods, and specialised bioinformatic techniques often have to be used to address their limitations (Silverstein *et al.*, 2007; Shafee, Robinson, *et al.*, 2016). For example, phylogenetic analysis becomes unreliable for short, divergent, sequences under strong selection pressures, where similarities may be due to either common ancestry or convergence (homoplasy) (Wake, 1991; Rost, 1999). Sequence analysis has therefore

been limited to closely related families (Zhu, 2008; Ma *et al.*, 2012; Takeuchi and Higashiyama, 2012; Cabezas-cruz *et al.*, 2016; Van der Weerden and Anderson, 2013).

Defensins are typically categorised into disulphide classes solely based on their cysteine spacing and connectivity. Although simple, this approach has several drawbacks. Many cysteine patterns contains members with different functions (e.g. the C6 class contains AMPs, toxins, and signalling proteins; The C8 class contains AMPs and sweet tasting proteins (Shafee *et al.*, 2017)). Finally, the relationships between the classes remain ill-defined. Consequently, the limited predictive power and biological meaning of the classes misses a global view of the superfamily's sequence-function relationships.

A better analysis approach for such a diverse superfamily is to build a quantitative map of protein sequence space, based on the biophysical properties of each residue in each sequence. In this work, I describe this analysis procedure, and its application to the cis-defensin superfamily.

Protein sequence space

Protein sequence space is a conceptual arrangement all possible protein sequences in a highly multidimensional space (Smith, 1970). At its simplest, the space has one dimension per amino acid in the sequence, to represent all possible combinations of amino acids. Proteins can then be arranged in this space based on their residues, with functional proteins occupying a vanishingly small fraction of the total space (Keefe and Szostak, 2001; Stemmer, 1995). Protein evolution can be thought of as the movement though sequence space as mutations explore new regions of it, with selection retaining only fit variants (Romero and Arnold, 2009). It is often used in a qualitative sense to aid discussion (Harms and Thornton, 2013), but is difficult to directly analyse in a quantitative manner.

Two key difficulties in generating quantitative maps of protein sequence space are a) how to ascribe meaningful, quantitative axes for the dimensions, and b) how to visualise and interpret such a high-dimensional space. Various metrics have been proposed for describing protein sequences numerically, from BLOSUM matrices (Casari *et al.*, 1995), binary descriptors for each possible residue (Wang and Kennedy, 2014), and biophysical properties of residues or whole proteins (Rackovsky, 2009; Du *et al.*, 2006). We favour the use of biophysical properties for each residue, since the resulting positions of sequences within sequence space more informative. The most divergent biophysical properties of the 20 naturally occurring amino acids were used in order to be easily interpretable (Atchley *et al.*, 2005).

The map of cis-defensin sequence space contains clear groups of sequences. Importantly, the main functions of the superfamily (such as antimicrobial, immunogenic, or neurotoxic activities) are clearly separated, enabling identification of the key biophysical properties that determine those functions. Regions of sequence space that are densely or sparsely populated are clearly identifiable, as are sequences with intermediate properties between groups. It is therefore a useful method of classification and analysis for sequence-diverse superfamilies like the defensins.

https://doi.org/10.1093/bioinformatics/bty697

https://doi.org/10.1093/bioinformatics/bty697

Results

A set of evolutionarily related cis-defensin sequences were gathered by a combination of structural and sequence similarity searches, as previously published (Shafee, Lay, *et al.*, 2016). Sequences were aligned using methods specialised for CRP sequences, which constrain homologous cysteines (and therefore inter-cysteine regions) based on solved structures (Shafee, Robinson, *et al.*, 2016). The resulting Multiple Sequence Alignment (MSA) contains indels mainly in solvent-exposed loop regions and is consistent with structure alignments and homologous secondary structure (Fig S1).

To generate meaningful sequence space axes, residues must be placed along each axis in a meaningful order. In this work, proteins were placed in a quantitative sequence space using the biophysical properties of each residue in the MSA (charge, hydrophobicity, molecular weight, disorder propensity, disulphide potential, MSA column occupancy; Table S3). In this case, each defensin sequence was represented by a vector of its residues' biophysical properties in a 1176-dimensional sequence space. This high-dimensional space was then rotated and projected into a smaller number of dimensions that summarised the main sources of covarying properties by Principal Component Analysis (PCA). This process captured the important features of the sequence space in a human-comprehensible number of dimensions, analogous to a 3D shadow of a mega-dimensional object. Within the projected sequence space, unsupervised learning was then used to identify naturally-occurring clusters of sequences with similar biophysical properties (Figure 1C).

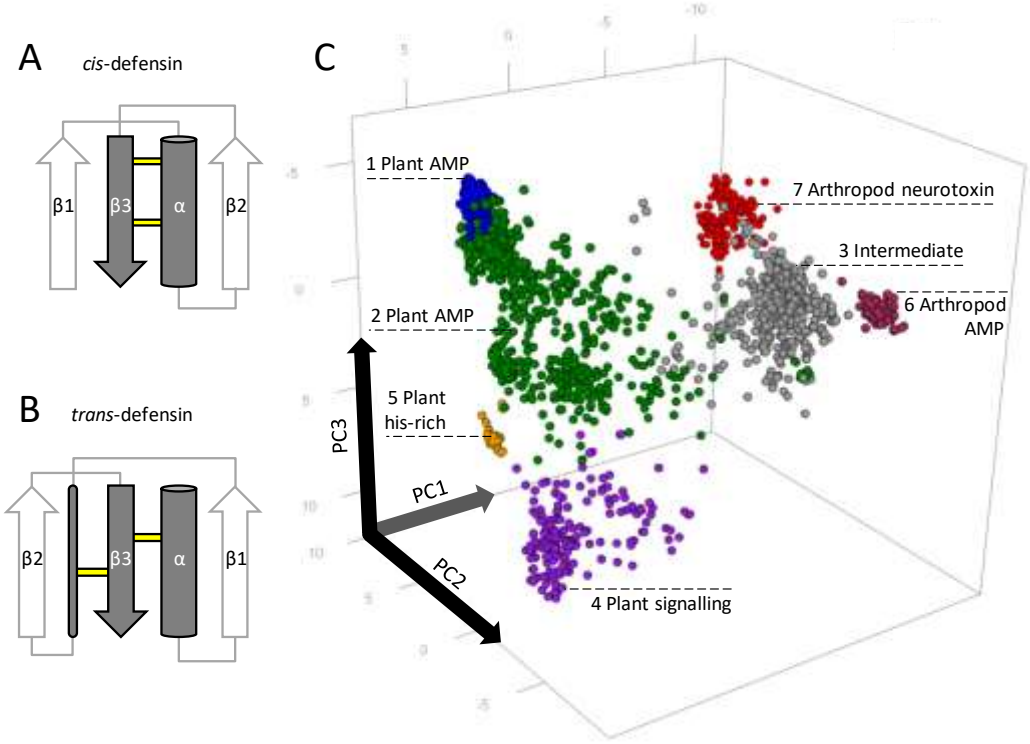


Figure 1 | Quantitative map of sequence space for the cis-defensin superfamily. Secondary structure and disulphide topology that define the (A) cis-defensin and (B) trans-defensin superfamilies with conserved disulphides in yellow. (C) Each point represents a sequence. Axes are principal components 1, 2, and 3 of the mega-dimensional sequence space. Sequences are coloured according to the identified Bayesian model-based groups and named as

described in the text (Group 1 in blue, 2 in green, 3 in grey, 4 in purple, 5 in orange, 6 in maroon, 7 in red). Rotatable image at TS404.shinyapps.io/DefSpace. AMP = antimicrobial protein, His-rich = histidine-rich.

Key functions reside in well-defined regions of sequence space

The axes of the projected sequence space (Principal Components, or PCs) summarise the most significant covarying sequence properties. The first three PC dimensions of the sequence space separate out three key features of the cis-defensin superfamily (Fig S2). PC1 separates the plant and animal sequences. PC2 separates the antimicrobial proteins from the neurotoxins. PC3 separates those defensins with signalling functions. The fourth PC further separates a group of histidine-rich defensins (Fig S3). Further variance is captured by subsequent PCs, however higher PCs do not give much additional separation into clusters (Fig S4).

The sequences form seven well-separated clusters in the sequence space (Table 1), with voids between them sparsely populated by sequences with intermediate properties. Groups were automatically assigned by Bayesian model-based clustering to identify naturally occurring groups of proteins. Each group was additionally analysed separately by repeating the PCA and clustering process to identify subclusters (Fig S5, Fig S6). This produced a broad hierarchy within the sequence space. Phylogenies could be resolved for some subclusters and illustrated how the subcluster had evolved through the sequence space (Fig S3E). This greatly extends on the standard classification system based on cysteine motifs, since sequences with the same motif are further organised within the clusters in sequence space and clear sets of covarying biophysical properties can be identified.

Table 1 | The 7 main groups of cis-defensins in sequence space

	Description	Kingdom	Colour in Fig. 1	Number in group	Plant	Fungal	Animal	Toxin	Signal
1	Plant AMP	Plant	blue	223	100%	-	-	-	1%
2	Plant AMP	Plant	green	677	93%	-	7%	-	3%
3	Intermediate	Mixed	grey	495	39%	5%	57%	31%	3%
4	Plant signalling	Plant	purple	190	100%	-	-	-	100%
5	Plant his-rich	Plant	orange	56	100%	-	-	-	-
6	Arthropod AMP	Animal	maroon	134	-	-	100%	-	-
7	Arthropod toxin	Animal	red	244	-	-	100%	100%	-

Groups 1 and 2 contain most of the plant antimicrobial proteins. They are the only two groups that have a continuous set of intermediate sequences along the spectrum of properties between them (with no void). Group 1 consists of more hydrophilic, cationic proteins (e.g. MtDEF4), whilst group 2 is the largest cluster and contains most of the characterised plant defensins (e.g. NaD1). They contain mostly 8-cysteine defensins, but also include the 10-cysteine defensins from petunia (e.g. PhD1). The class II defensins are tightly clustered, whereas class I are distributed through groups 1 and 2 (Figure 2A), in line with their greater variety of mechanisms and efficacies (Payne *et al.*, 2016; Parisi *et al.*, 2018). Additionally, defensins that are known to enter the target cell during their

https://doi.org/10.1093/bioinformatics/bty697

mechanism of action (e.g. NaD1 and MtDef4) tend to be high in the PC1 direction, whereas defensins that remain outside the cell during their mechanism are lower in PC1 direction (e.g. RsAFP1) (Parisi *et al.*, 2018).

Group 3, the intermediate cluster, contains a mixture of plant, animal and fungal proteins and a mixture of functions. When the sequence space is analysed for just the group 3 proteins alone, clear subclusters are evident, such as the C6 defensins, macin defensins, fungal defensins, plant fusion defensins, and various classes of scorpion toxin (Figure 2C). The defensins with a C6 structures are similarly segregated into toxic and non-toxic subclusters (Figure 2C). Taxa are clearly organised within group 3, and the non-toxic C6 defensins are the only subcluster to contain a mixture of plant, fungal and animal members.

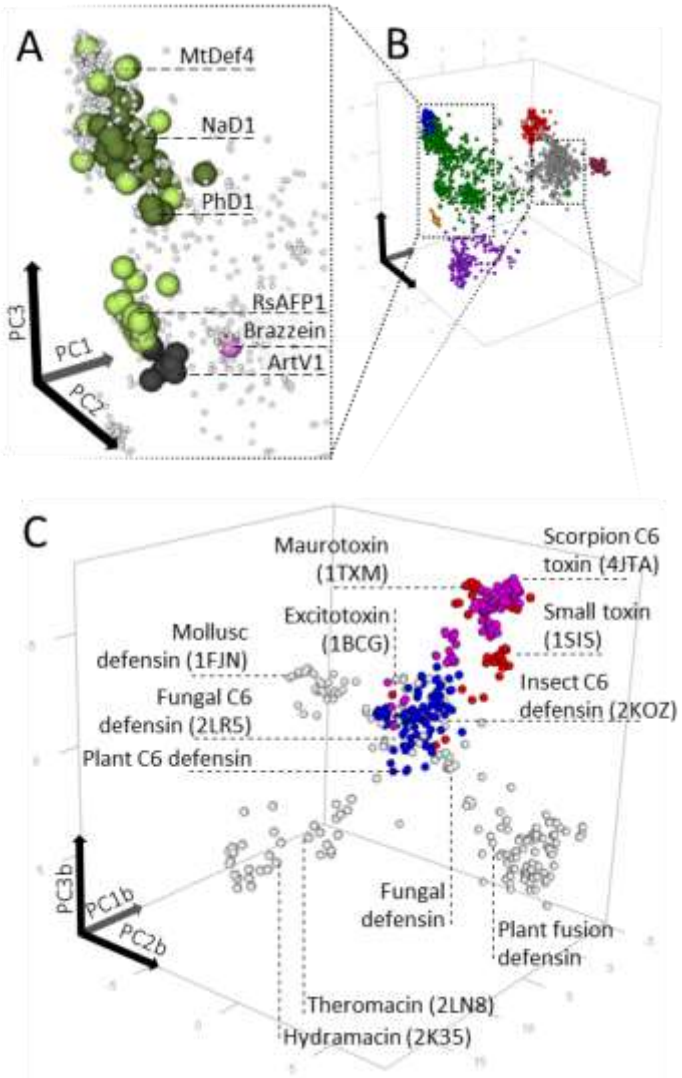


Figure 2 | Groups 1, 2 and 3 separate different disulphide classes, taxa and functions.

(A) Close-up of groups 1 and 2, with antimicrobial proteins in green (antifungal class I in light green, class II in dark green), allergen pollen proteins in grey, sweet tasting protein in pink, un-tested function in white. Example sequences named. (B) Sequence space as in Figure 1 for

context. (C) Group 3 ('intermediates') isolated and re-analysed. C6 toxins in purple, non-C6 toxins in red, and non-toxic C6 defensins in blue. PDB codes for representative structures indicated where available. Axes are labelled as PC1b, 2b, 3b to emphasise that they are new principal component axes for group 3, distinct from those PCs found the overall sequence space.

Group 4 contains the majority of the plant signalling proteins involved in self/non-self recognition, which mediate sporophytic self-incompatibility during fertilisation. The cluster has a broad boundary with group 2 plant defensins, with the interface populated by sequences with intermediate properties, including the unusually disulphide bonded sex-locus g class (Shafee, Lay, *et al.*, 2016).

Group 5 contains a small cluster of plant defensins that have a significant enrichment of histidines (Fig S5E). It also contains two subclusters with intermediate numbers of histidines that they appear to have evolved from (Fig S3E). Currently none of these members are characterised, but given their unusual residue composition, it is likely that they will display interesting properties, possibly in addition to antimicrobial activity (Mirouze *et al.*, 2006).

Group 6 contains C6 defensins with long n-loops (Koehbach, 2017). These are primarily antibacterial insect defensins (such as sapecin) with a few mollusc, tick and lancelet variants. The short n-loop, primarily antifungal, C6 defensins (such as heliomicin) are distributed through subclusters of group 3 (Fig S5F).

Group 7 contains the scorpion α-toxins and β-toxins. This group is particularly well-separated from the others, with relatively few sequences in between. It has two main sub-clusters that separate the known α- and β-toxins (Gopalakrishnakone *et al.*, 2015). The sequences that do fall in the relative void between groups 7 and 3 are the excitotoxins, which show similarities to the α- and β-toxins with elongated N-terminal loops (Oren *et al.*, 1998).

Clusters indicate viable combinations of biophysical properties

The axes of the quantitative sequence space map are determined by covarying sets of residue biophysical properties (Table S1). The sequence and biophysical properties that most strongly determine a sequence's position in the projected sequence space are summarised by the PC loadings that define each axis.

Because the sequence space is built from an MSA, the data can be mapped onto known tertiary structures to give some spatiochemical information (Figure 3). For example, loadings indicate that antimicrobial activity is strongly influenced by having a hydrophobic within loop 5 with charge at either end (Table S1, MSA columns 108-149), a region involved in lipid binding in plant defensins (Poon *et al.*, 2014; Payne *et al.*, 2016). These covarying residue property sets correlate well with regions of biological importance. Variation in this region is also correlated with features on loop 1b and 4a the opposite face of the structure (MSA columns 40-42 and 77-80, Figure 3A), and extend previous work describing roles for these loops in overall antimicrobial function (Bleackley *et al.*, 2016).

https://doi.org/10.1093/bioinformatics/bty697

Disulphide potential and residue occupancy were amongst the most highly loaded properties for the first PC axes (Table S1), in agreement with the known disulphide classes. The additional biophysical properties are necessary to identify well-defined groups, since repeating the analysis using only occupancy and disulphide potential is insufficient to find clear clusters (Fig S4C,D). Similarly, properties that are highly conserved across the whole superfamily (such as the four cysteines that define the cis-defensins) have low loading for the axes (Fig S8). The most useful information is contributed by residues that are not too constrained (e.g. the fully conserved cysteines) nor too unconstrained (e.g. positions where any residue is viable) (Fig S8).

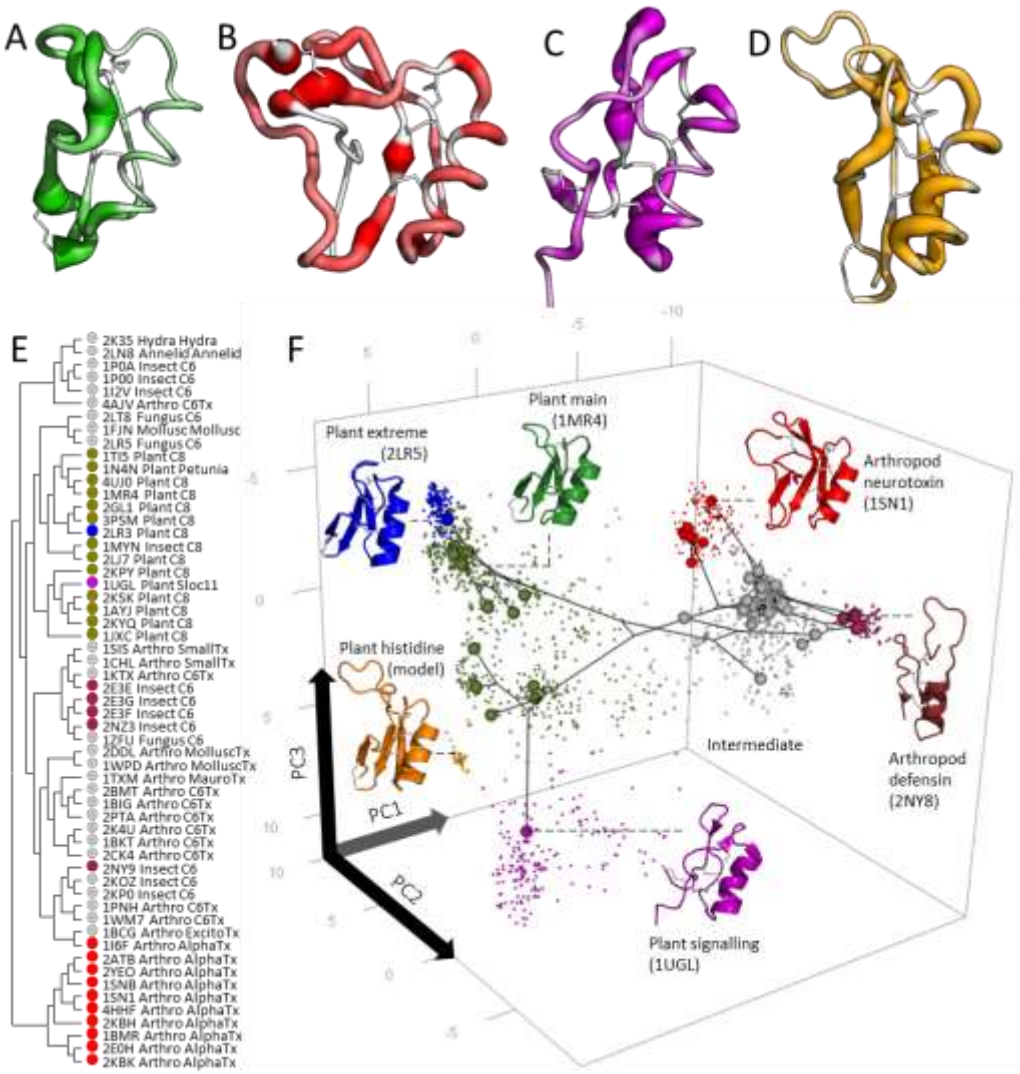


Figure 3 | Structures are well organised in sequence space.

Representative structures with backbone ribbon thickness scaled by residue property loadings (summed across all biophysical properties) for the first four PC axes. The ribbon is thicker in sequence regions which are more important in determining the protein's location in sequence space. (A) Property loadings for PC1 mapped onto a plant defensin (PDB:1MR4). (B) Property loadings for PC2 mapped onto scorpion toxin (PDB:1SN1). (C) Property loadings for PC3 mapped onto plant S-locus protein (PDB:1UGL). (D) Property loadings for PC4 mapped onto a histidine-rich plant DLP (model). (E) Hierarchical clustering of superfamily members with solved structures based on tertiary structural similarity. Colours as in Figure 1. (F) Sequence space diagram with

structurally characterised defensins as larger spheres connected by the same structural dendrogram. Representative structures shown except for intermediate (grey) group, which contains multiple structural classes.

Structural classes are well ordered in sequence space

Structure is well known to be more evolutionarily conserved than sequence and is therefore extremely informative for long evolutionary timescales (Shafee, Lay, *et al.*, 2016; Undheim *et al.*, 2016; Orengo and Thornton, 2005). Although the cis-defensin superfamily has a conserved overall fold, there are a diverse range of variations on the structure (including changes in disulphide number, loop length, and secondary structure elements). The sequence space analysis was performed using only a sequence alignment as an input, i.e. without using structural information (except for guiding the MSA) (Shafee, Robinson, *et al.*, 2016). To test whether these similar structures cluster together into particular regions of sequence space or are distributed throughout, proteins with known structures were hierarchically clustered by their structural similarity. The resulting structural similarity dendrogram was threaded through the sequence space.

Defensins with similar biophysical properties also shared more similar structures (Figure 3E,F), even for proteins of the same cysteine class (e.g. C8 defensins). Within proteins of the same cysteine motif, there are therefore distinct identifiable subclusters with shared biophysical and structural properties. Therefore disulphide classes, and groups of similar structures within classes, are inhabit distinct regions of sequence space. Although each group shares many biophysical properties, intermediate forms do exist between the more populated clusters.

Comparisons with other sequence and structure analyses

A maximum likelihood phylogeny using the MSA gave low bootstrap values, with an average bootstrap of 0.22 across all nodes of the tree (Fig S9a), and with deeper branches showing negligible bootstrap support (Fig S9b). These bootstrap values are in line with previously published defensin phylogenies (Van der Weerden and Anderson, 2013; Zhu *et al.*, 2005). Collapsing the tree at nodes with bootstrap >0.2 produced over 100 separate families (median 9 members). This stems from well-established limitations of phylogenetics for short, divergent, strongly-selected sequences (Shafee, Robinson, *et al.*, 2016). Indeed, for the seven largest phylogenetic clades, only 43% sequences fall into the same clade in each bootstrap repeat (Fig S10A).

Conversely, in the sequence space map of each of the alternative MSAs, 86% of sequences remain in the same cluster in each bootstrap repeat (Fig S10B). Variation in the positions of sequences within the space upon bootstrapping is small enough that cluster prediction is perturbed for only a few sequences near the edges of clusters (Fig 4A). This is in line with related multivariate analysis methods, which can be more accurate for more divergent sequences, robust to changes in dataset, and not as perturbed by intermediate or hybrid sequences (Higgins, 1992; Wallace and Higgins, 2007).

https://doi.org/10.1093/bioinformatics/bty697

Both phylogenetics and the methods presented here rely on an accurate MSA. A set of 100 alternative MSAs using the same sequences were generated to test how variation on the starting MSA perturbs downstream analyses. The average bootstrap across all nodes of the phylogeny fell to 0.11, and the sequences placed into the same major clades fell from 43% to 32% (Fig S10D). The sequence space method proved more robust to alterations in the MSA fall from 86% to 83% (Fig S10C). This is largely because the properties that dominate the separations in sequence space are also those that are most consistently aligned in the alternative MSA.

Sequence similarity networks (SSNs) represent an alternative method that does not require an MSA and is often informative for diverse protein superfamilies (Atkinson *et al.*, 2009; Cheng *et al.*, 2014). However SSNs struggle to organise defensins due to their low sequence similarities, causing the network to fracture into smaller sub-networks, even at a permissive expect value of 10^{-4} (Shafee, Lay, *et al.*, 2016). These sub-networks agree well with the clustering found in the sequence space analysis (Fig S11), allowing them to be arranged relative to one another. However, the sequence space analysis also gives information on the biophysical determinants of the identified sequence groups. Finally, because the layout is deterministic, rather than stochastic, bootstrap and jackknife replicates can be used to check the sensitivity of the sequence locations or number of clusters to the dataset used (Figure 4).

https://doi.org/10.1093/bioinformatics/bty697

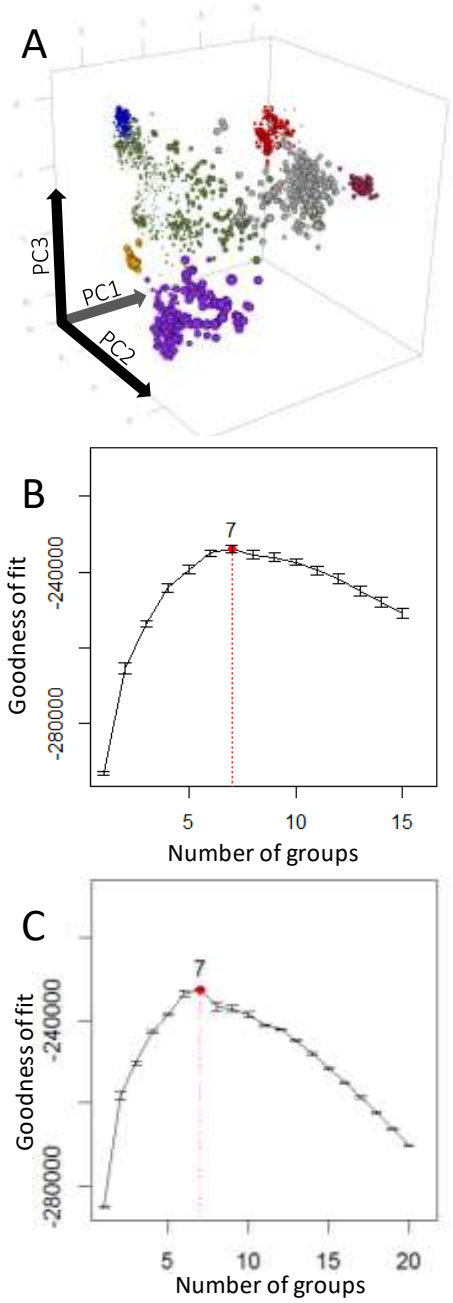


Figure 4 | Repeatability of coordinates within the sequence space.

(A) Sequence space diagram with spheres scaled to the variance in their coordinates from 100 bootstrap replicates. Colours indicate groups as in Figure 1C. (B) The goodness of fit for the number of groups based on Bayesian model based clustering. Error bars = stdev of 100 bootstrap replicates. (C) As in B, but with jackknife replicates.

Defensin sequence space can be explored using a webtool

To facilitate interactive viewing and query of the sequence space, a simple webtool is available at TS404.shinyapps.io/DefSpace (source code at github.com/TS404/DefSpace). This tool can interactively display the sequence space, and calculate the locations of query sequences (Fig S12). It also identifies whether a sequence is a cis- or trans-defensin and which cysteine motifs the sequence contains (if any). Cysteine motifs are based on regular expressions that encompass 90% of the variation in known sequences (Table S2). This allows defensin sequences of interest to be quickly checked for their position in order to give some expectation of their function.

https://doi.org/10.1093/bioinformatics/bty697

Discussion

Defensins are found in most transcriptomes and genomes, yet predicting their function has remained elusive due to their highly diverse sequences and activities.

The broader sequence-function relationships and evolutionary history of the cis-defensins are only beginning to be understood. For example, mutagenesis studies have identified some of the sequence determinants that differentiate antimicrobial from neurotoxic activity in arthropod cis-defensins (Zhu *et al.*, 2014). Conversely, conservation of antiplasmodal activity in a hypothetical ancestral defensin from the base of the chelicerates was demonstrated by ancestral sequence reconstruction (Cabezas-cruz *et al.*, 2016). This work allows for more ancient evolutionarily events in the superfamily's history to be analysed.

Implications for early defensin evolution

The ancient evolution of the defensins is highly unclear. The recent separation into two separate superfamilies indicates independent origins for the cis-defensins and trans-defensins (Shafee, Lay, *et al.*, 2016). However, the early evolution within each superfamily remains obscured by their extreme sequence divergence. Some additional insight can be gained from the organisation of functions within the sequence space, which captures the limits of the viable sequence diversity. As mutations have generated sequences in new regions of sequence space, selection has constrained them to clustered regions where they retain or gain useful functions.

The superfamily displays great sequence diversity and many clusters have proteins between them with intermediate properties. Nevertheless, there are clearly preferred regions of sequence space with several clear voids indicating unfavorable sequence property combinations. At the same time, some activities are achieved by proteins in distinct regions of sequence space, highlighting how the defensins have evolved multiple viable ways of achieving similar functions. It is also consistent with several independent recruitments of the fold to toxic function.

The analysis also gives clues to the origins of the superfamily. The C6 defensins are clearly more similar to one another than they are to other defensins, even when present across multiple taxa (Shafee *et al.*, 2017). This supports a scenario where the C6 defensin scaffold is more ancient than the other disulphide classes, which represent elaborations of this more ancient fold (Figure 5). It is also in line with the possibility of a yet more ancient ancestral structure with only two disulphides (Shafee *et al.*, 2018). The clustering of the C8 defensin fold is consistent with divergence from the C6 defensins before the separation of plants and animals, as opposed to multiple origins of the C8 fold.

https://doi.org/10.1093/bioinformatics/bty697

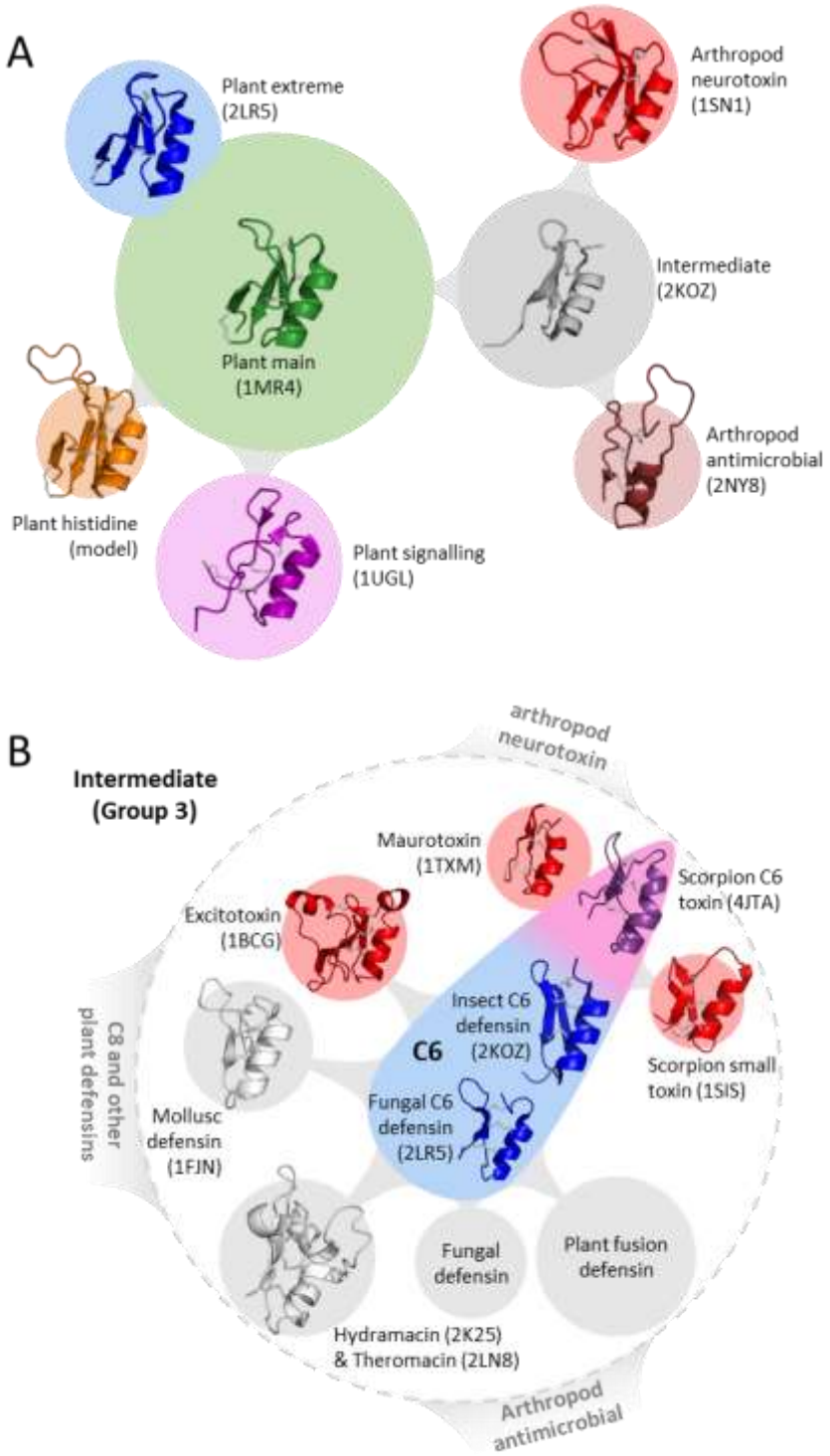


Figure 5 | Hypothetical evolutionary history of the cis-defensin superfamily.

Flattened schematics of (A) the full superfamily's sequence space, coloured as in Figure 1, and (B) just the intermediate group, coloured as in Figure 2. Structures shown where available. The most ancient scaffolds are likely the C6 defensins involved in host defence and antimicrobial activity, followed by the C8 defensins. Subsequent elaborations of those scaffolds gave rise to the variety of folds and functions now present in the superfamily.

https://doi.org/10.1093/bioinformatics/bty697

Prediction of function from sequence

Defensin sequences are not randomly distributed on sequence space, despite their extreme sequence diversity. Instead, the sequence space contains well-defined clusters of sequences with similar biophysical properties, separated by sparsely-populated voids. However, there remain plenty of sequences with intermediate properties. Groups 1 and 2 are the exception to this, with a continuous transition of intermediates at their interface, in line with their similar microbial functions (though via varied mechanisms) (Parisi *et al.*, 2018). The number of covarying residues indicate a high degree of cooperation in determining defensin function.

The identified groups correlate well with currently known functions within the superfamily. There are clear separations between plant and animal defensins, as well as antimicrobial, signalling, and toxic functions. The PC axes that separate these groups also describe the key biophysical properties that characterise the different functions. This information will be useful in identifying likely biological roles of sequences, as well as which sequence elements determine that function, and whether it can be further engineered by mutagenesis. Newly identified sequences of interest can be added to the sequence space to identify their relative position, identify cluster centrality, and find nearby neighbours with known functions, mechanisms or structures.

Plant, animal, and fungal sequences are largely segregated, and mostly form clearly identifiable clusters. Even the taxonomically diverse C6 class is organised within group 3, with the long n-loop C6 defensins further clustered in group 6. Key sequence properties of the C6 subclusters have therefore been conserved since the plant-fungal-animal split, despite extensive sequence change.

The sequence space organisation lays the groundwork for improving our understanding of how divergent evolution gave rise to the superfamily's functional diversity. How have the different antimicrobial mechanisms evolved? Is the ability of antimicrobial plant defensins to traverse cell walls and enter the nucleus related to their role in development, possibly by acting as transcription factors or inducing apoptosis? What is the function of the well-defined cluster of uncharacterised histidine-rich defensins? What activities are exhibited by those defensins that lie in the sparsely populated inter-cluster regions?

Benefits and limitations

The extreme sequence diversity of the cis-defensin superfamily affords both opportunities and difficulties. It provides excellent library of functional peptides, with a scaffold compatible with a range of different activities. Conversely, the same diversity limits the application of standard sequence analysis methods and protein engineering. Although any displayed sequence space is necessarily a simplification of the full mega-dimensional sequence space, the simplified projections presented here retain a great deal of useful information, and are relatively intuitive to understand.

This quantitative sequence space map of the cis-defensin superfamily provides a foundation for classifying defensins and understanding their sequence-function relationships. Biologically relevant clusters are clearly identifiable within the sequence

space, as well as sequences with intermediate properties. Sequence space axes report on key properties that separate proteins with a layout is deterministic rather than stochastic.

Since an MSA is used, proteins analysed by this method must be related to one another, such that columns of the MSA contain homologous residues (unlike e.g. similarity networks where only pairwise alignments are needed). Proteins not derived from a common ancestor do not have homologous residues aligned and so cannot be compared. However, it may be cautiously extendable to sequences for which analogous residues can be confidently compared.

Confidently resolving the "twilight zone" evolutionary relationships of highly divergent sequences is currently beyond the scope of either phylogenetic, or non-phylogenetic sequence analysis methods (Rost, 1999; Atkinson *et al.*, 2009; Inkpen and Doolittle, 2016; Pearson and Sierk, 2005). This method therefore forgoes an explicit evolutionary model, and so does not distinguish whether changes in sequences' biophysical properties are conserved from an ancestral state or converged upon from different states (Jackson *et al.*, 2018). Convergence of multiple biophysical properties would generate similar homoplasy to that which can occur in phylogenies. Particularly in the case of small CRPs, like the defensins, the strong selective pressures and high evolvability due to the scaffold's structural robustness make it likely that some covarying property sets have been arrived upon in separate lineages.

https://doi.org/10.1093/bioinformatics/bty697

Conclusions

Quantitatively mapping the sequence space of the cis-defensins provides a useful and informative classification system. It identifies key biophysical properties that separate biological functions, and hierarchically clusters groups of sequences by their biophysical properties. Biological kingdoms and annotated functions are well organised within their own regions of the sequence space. The cluster-centrality of a sequence is clearly identifiable, as are intermediates between the main clusters.

This sequence analysis technique may also be broadly applicable to other cysteine-rich protein superfamilies, whose sequences are similarly too diverse for standard classification techniques. When phylogenetics is applicable, sequence space maps provide complementary information on how the biophysical properties of a protein have evolved.

The accompanying webtool facilitates easy identification of sequences of interest within the sequence space map and provides additionally information on superfamily and known motifs.

Availability

Project home page: TS404.shinyapps.io/DefSpace

Repository: GitHub.com/TS404/DefSpace

Operating system(s): Platform independent

Programming language: R

Other requirements: R 3.1 or higher

License: Academic Free License 3.0

https://doi.org/10.1093/bioinformatics/bty697

Methods

Sequence gathering

Homologous cis-defensin sequences were gathered as per reference (Shafee, Lay, *et al.*, 2016). Briefly, cis-defensin structures were gathered using DALI (Holm and Rosenström, 2010), with the plant defensin NaD1 (PDB:1MR4) as the initial query. Unique proteins whose structures had Z-scores >2 were collected and used as queries in turn until no new structures were identified. These sequences were used as query sequences for BLASTp searches against the non-redundant protein database and Sol genomics network database (E-value cutoff <0.005) (Mueller *et al.*, 2005). Additional rounds of sequence gathering were performed by using the least redundant sequences as queries for subsequent rounds of BLASTp until no new sequences were identified. This yielded 2019 total sequences from 352 species.

Sequence alignment

Because defensin sequences have high sequence length variation, sequence alignments are more accurate when structurally homologous cysteines are constrained, allowing the homologous inter-cysteine loops to align (Shafee, Robinson, *et al.*, 2016). Briefly, multiple sequence alignments were generated using the CysBar webserver (Shafee, Robinson, *et al.*, 2016) and ClustalΩ (Sievers *et al.*, 2011). Homologous cysteines were barcoded to constrain the alignment and allow the correct inter-cysteine loops to align. Overall protein structure was not used to otherwise inform the sequence alignment. Alternative alignments were compared with AlignStat (Shafee and Cooke, 2016). The final MSA (supplementary data file 1) was used for subsequent analyses (Fig S13). The MSA showed good alignment of homologous secondary structure elements, with indels mainly occurring in solvent-exposed loop regions, and with agreement with structure alignments. To test the dependence of the analyses on the MSA, a set of 100 alternative MSAs of the same sequences were also generated using Guidance2 as described later in the section.

Structure similarity network

Structure similarity was calculated as per reference (Shafee, Lay, *et al.*, 2016). Briefly, for cis-defensins with known structure, pairwise structure alignment using combinatorial extension was performed with the proCKSI webserver to generate an RMSD distance matrix. This matrix was hierarchically clustered by neighbour joining of the most similar structures to generate a bifurcating dendrogram which was then mapped onto the sequence space coordinates in [R]. Disulphide classes were annotated based on cysteine motif and connectivity (Shafee *et al.*, 2017).

Homology model structure

Published protein structures are used with the exception for NbD2 (no structures have yet been solved for any his-rich group 5 protein). For illustrations, a simple homology model was made of NbD2 using SWISS-model, with PDB:3PSM selected as the best template structure (Arnold *et al.*, 2006).

https://doi.org/10.1093/bioinformatics/bty697

Phylogeny

A maximum likelihood phylogeny with 1000 bootstraps was generated based on the MSA using RaxML (Stamatakis, 2014). The optimal substitution model was identified using ProtTest as Whelan and Goldman model with a gamma distribution (Darriba *et al.*, 2011; Whelan and Goldman, 2001). A strictly trimmed alignment (using trimAl (Capella-Gutiérrez *et al.*, 2009)) did not give notably different results from the full alignment. Phylogenies were annotated using iTOL (Letunic and Bork, 2016).

Numericisation

To quantitatively place sequences in a multidimensional sequence space, each residue of each sequence in the MSA was described by its biophysical properties. The variables used were R-group molecular weight (Daltons), net charge (Coulombs), hydrophobicity (Doolittle index) (Kyte and Doolittle, 1982), disorder propensity (TOP-IDP) (Campen *et al.*, 2008), disulphide potential (binary descriptor), and occupancy (binary descriptor). These properties encompass the main differences between the naturally occurring amino acids. Disulphide potential is included in this case since disulphides are particularly important to defensin structures. MSA column occupancy accounts for different sequence lengths. See supplementary table S3 for the values used. Note that this is also compatible with non-natural amino acids, or additional biophysical properties of interest.

The MSA with 2019 sequences and 196 columns was therefore converted into a numerical matrix with 2019 rows and 1176 columns (196 × n where n is the number of biophysical properties used, in this case, n=6 leading to 196 × 6 = 1176 columns). The resulting matrix of numerical values was used to represent the raw protein sequence space before multidimensional scaling.

Values were normalised within each property. Gaps were given the average value of their column for each property (other than occupancy) such that they had no effect on subsequent multidimensional scaling.

Multidimensional scaling

The highly multidimensional, numericised protein sequence space was analysed using Principal Components Analysis (PCA) to summarise the main covarying sets of properties (using [R] prcomp) (Dev. Core Team, 2011). The resulting principal components describe covarying sets of residue properties.

Bayesian clustering

Bayesian clustering was performed using the Gaussian finite mixture method (using [R] mclust) (Fraley and Raftery, 2012). The first 40 PCs were used for clustering since they summarised the most important 30% of the information contained in the 1176 dimensions.

Briefly, this algorithm calculates the models the distribution of data points as a set of spheroid clusters with varied sizes, elongation and orientation. The optimal number of clusters is chosen based on goodness of fit (Bayesian Information Criterion). Adding

clusters to the model improves the models fit to the data until an optimal number of clusters is reached, after which additional clusters fail to improve the model's fit.

Repeatability metrics

Bootstraps of the sequence space were generated by iteratively ignoring 10% of columns in the same MSA and generating the sequence space each time. Variation in the coordinates in the sequence space were reported as well as the optimal number of Bayesian clusters to summarise the data. Similarly jack-knife replicates were performed by randomly ignoring 10% of sequences in the MSA.

To further assess the dependence of the methods on the input MSA, 100 alternative MSAs were generated using Guidance2, which generates plausible variant alignments of indels (Sela *et al.*, 2015). These MSAs were used to generate maximum likelihood phylogenies and sequence space maps as described below.

To assess phylogenetic dependence on the input MSA, each of the 100 alternative MSAs was used to generate a maximum likelihood phylogeny using the same model parameters as for the original tree. For each of the 100 trees, the 7 largest clades were identified by ward clustering in [R] and compared for repeatability.

To assess SeqSpace dependence on the input MSA, the 100 were each numericised, scaled and clustered as described above. The Bayesian clusters found in each repeat were then compared for repeatability in [R].

Sequence similarity networks do not depend on an MSA (rather they typically use all-vs-all pairwise alignments) so a comparable measure was not possible in this regard.

Visualisation

Data was visualised with custom [R] scripts based on rgl, ggplots, igraph, and phytools (Revell, 2012; Csárdi and Nepusz, 2006; Adler *et al.*, 2012; Wickham, 2009). Structures were visualised with Pymol.

https://doi.org/10.1093/bioinformatics/bty697

References

Adler,D. *et al.* (2012) rgl: 3D visualization device system (OpenGL).

Arnold,K. *et al.* (2006) The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.

Atchley,W.R. *et al.* (2005) Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci.*, **102**, 6395–6400.

Atkinson,H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*, **4**, e4345.

Bleackley,M.R. *et al.* (2016) Nicotiana alata Defensin Chimeras Reveal Differences in the Mechanism of Fungal and Tumor Cell Killing and an Enhanced Antifungal Variant. *Antimicrob. Agents Chemother.*, **60**, 6302–6312.

Cabezas-cruz,A. *et al.* (2016) Antiplasmodial activity is an ancient and conserved feature of tick defensins. *Front. Microbiol.*, **7**, 1–12.

Campen,A. *et al.* (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.*, **15**, 956–63.

Capella-Gutiérrez,S. *et al.* (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–8.

Cheng,S. *et al.* (2014) Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. *Front. Ecol. Evol.*, **2**, 1–13.

Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal Complex Syst.*, **1695**, 1695.

Darriba,D. *et al.* (2011) ProtTest-HPC: Fast selection of best-fit models of protein evolution. In, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*., pp. 177–184.

Dev. Core Team,R. (2011) R: A Language and Environment for Statistical Computing.

Du,Q.-S. *et al.* (2006) Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction. *J. Biomol. Struct. Dyn.*, **23**, 635–640.

Fraley,C. and Raftery,A.E. (2012) MCLUST Version 4 for R : Normal Mixture Modeling for Model-Based Clustering , Classification , and Density Estimation. *Tech. Rep. - Dep. Stat. Univ. Washingt.*, 1–57.

Gao,B. and Zhu,S. (2012) Alteration of the mode of antibacterial action of a defensin by the amino-terminal loop substitution. *Biochem. Biophys. Res. Commun.*, **426**, 630–5.

Gopalakrishnakone,P. *et al.* (2015) Molecular Description of Scorpion Toxin Interaction with Voltage-Gated Sodium Channels. *Scorpion Venoms*, 1–575.

Harms,M.J. and Thornton,J.W. (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, **14**, 559–71.

Higgins,D.G. (1992) Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Comput. Appl. Biosci.*, **8**, 15–22.

Holm,L. and Rosenström,P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.*, **38**, W545-9.

Inkpen,S.A. and Doolittle,W.F. (2016) Molecular Phylogenetics and the Perennial Problem of Homology. *J. Mol. Evol.*, 1–9.

Jackson,M. *et al.* (2018) Molecular basis for the production of cyclic peptides by the plant asparaginyl endopeptidases. *Nat. Commun.*, **(in press)**.

Keefe,A.D. and Szostak,J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–8.

Koehbach,J. (2017) Structure-Activity Relationships of Insect Defensins. *Front. Chem.*, **5**.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–32.

Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.

Ma,Y. *et al.* (2012) Extreme diversity of scorpion venom peptides and proteins revealed by transcriptomic analysis: Implication for proteome evolution of scorpion venom

arsenal. *J. Proteomics*, **75**, 1563–1576.

Mirouze,M. *et al.* (2006) A putative novel role for plant defensins: A defensin from the zinc hyper-accumulating plant, Arabidopsis halleri, confers zinc tolerance. *Plant J.*, **47**, 329–342.

Mueller,L.A. *et al.* (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol.*, **138**, 1310–1317.

Oren,D.A. *et al.* (1998) An excitatory scorpion toxin with a distinctive feature: an additional alpha helix at the C terminus and its implications for interaction with insect sodium channels. *Struct. with Fold. Des.*, **6**, 1095–1103.

Orengo,C.A. and Thornton,J.M. (2005) Protein Families and Their Evolution - a Structural Perspective. *Annu. Rev. Biochem.*, **74**, 867–900.

Parisi,K. *et al.* (2018) The evolution, function and mechanisms of action for plant defensins. *Semin. cell Dev. Biol.*, **(in press)**.

Payne,J.A.E. *et al.* (2016) The plant defensin NaD1 introduces membrane disorder through a specific interaction with the lipid, phosphatidylinositol 4,5 bisphosphate. *Biochim. Biophys. Acta - Biomembr.*, **1858**, 1099–1109.

Pearson,W.R. and Sierk,M.L. (2005) The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.*, **15**, 254–260.

Poon,I.K. *et al.* (2014) Phosphoinositide-mediated oligomerization of a defensin induces cell lysis. *eLife*, **3**, e01808–e01808.

Rackovsky,S. (2009) Sequence physical properties encode the global organization of protein structure space. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14345–14348.

Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.

Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–76.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sela,I. *et al.* (2015) GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.*, **43**, W7–W14.

Shafee,T. *et al.* (2017) Convergent evolution of defensin sequence, structure and function. *Cell. Mol. Life Sci.*, **74**, 663–682.

Shafee,T., Robinson,A.J., *et al.* (2016) Structural homology guided alignment of cysteine rich proteins. *Springerplus*, **5**, 27.

Shafee,T., Lay,F.T., *et al.* (2016) The defensins consist of two independent, convergent protein superfamilies. *Mol. Biol. Evol.*, **33**, 1–23.

Shafee,T. *et al.* (2018) The origin of defense: centipede toxin structure reveals a pre-eukaryotic origin of the CSαβ defensin superfamily. *Structure*, **(in review**.

Shafee,T. and Cooke,I. (2016) AlignStat: a web-tool and R package for statistical comparison of alternative multiple sequence alignments. *BMC Bioinformatics*, **17**, 434.

Sievers,F. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

Silverstein,K. *et al.* (2005) Genome organization of more than 300 defensin-like genes in Arabidopsis. *Plant Physiol.*, **138**, 600–610.

Silverstein,K. *et al.* (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *Plant J.*, **51**, 262–80.

Smith,J.M. (1970) Natural selection and the concept of a protein space. *Nature*, **225**, 563–4.

Stamatakis,A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stemmer,W.P.C. (1995) Searching Sequence Space. *Nat Biotech*, **13**, 549–553.

Takeuchi,H. and Higashiyama,T. (2012) A Species-Specific Cluster of Defensin-Like Genes Encodes Diffusible Pollen Tube Attractants in Arabidopsis. *PLoS Biol.*, **10**.

Undheim,E.A.B. *et al.* (2016) Toxin structures as evolutionary tools: Using conserved 3D folds to study the evolution of rapidly evolving peptides. *BioEssays*, **38**, 539–548.

Vriens,K. *et al.* (2014) Antifungal plant defensins: Mechanisms of action and production. *Molecules*, **19**, 12280–12303.

Wake,D.B. (1991) Homoplasy: The Result of Natural Selection, or Evidence of Design Limitations. *Am. Nat.*, **138**, 543–567.

Wallace,I.M. and Higgins,D.G. (2007) Supervised

multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinformatics*, **8**, 135.

Wang,B. and Kennedy,M. (2014) Principal components analysis of protein sequence clusters. *J. Struct. Funct. Genomics*, **15**, 1–11.

Van der Weerden,N.L. and Anderson,M.A. (2013) Plant defensins: Common fold, multiple functions. *Fungal Biol. Rev.*, **26**, 121–131.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Wickham,H. (2009) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag.

Zhu,S. (2008) Discovery of six families of fungal defensin-like peptides provides insights into origin and evolution of the CSalphabeta defensins. *Mol. Immunol.*, **45**, 828–38.

Zhu,S. *et al.* (2014) Experimental conversion of a defensin into a neurotoxin: implications for origin of toxic function. *Mol. Biol. Evol.*, **31**, 546–59.

Zhu,S. *et al.* (2005) Phylogenetic distribution, functional epitopes and evolution of the CSalphabeta superfamily. *Cell. Mol. Life Sci.*, **62**, 2257–69.

https://doi.org/10.1093/bioinformatics/bty697

Supplementary Data

https://doi.org/10.1093/bioinformatics/bty697

Supplementary figures

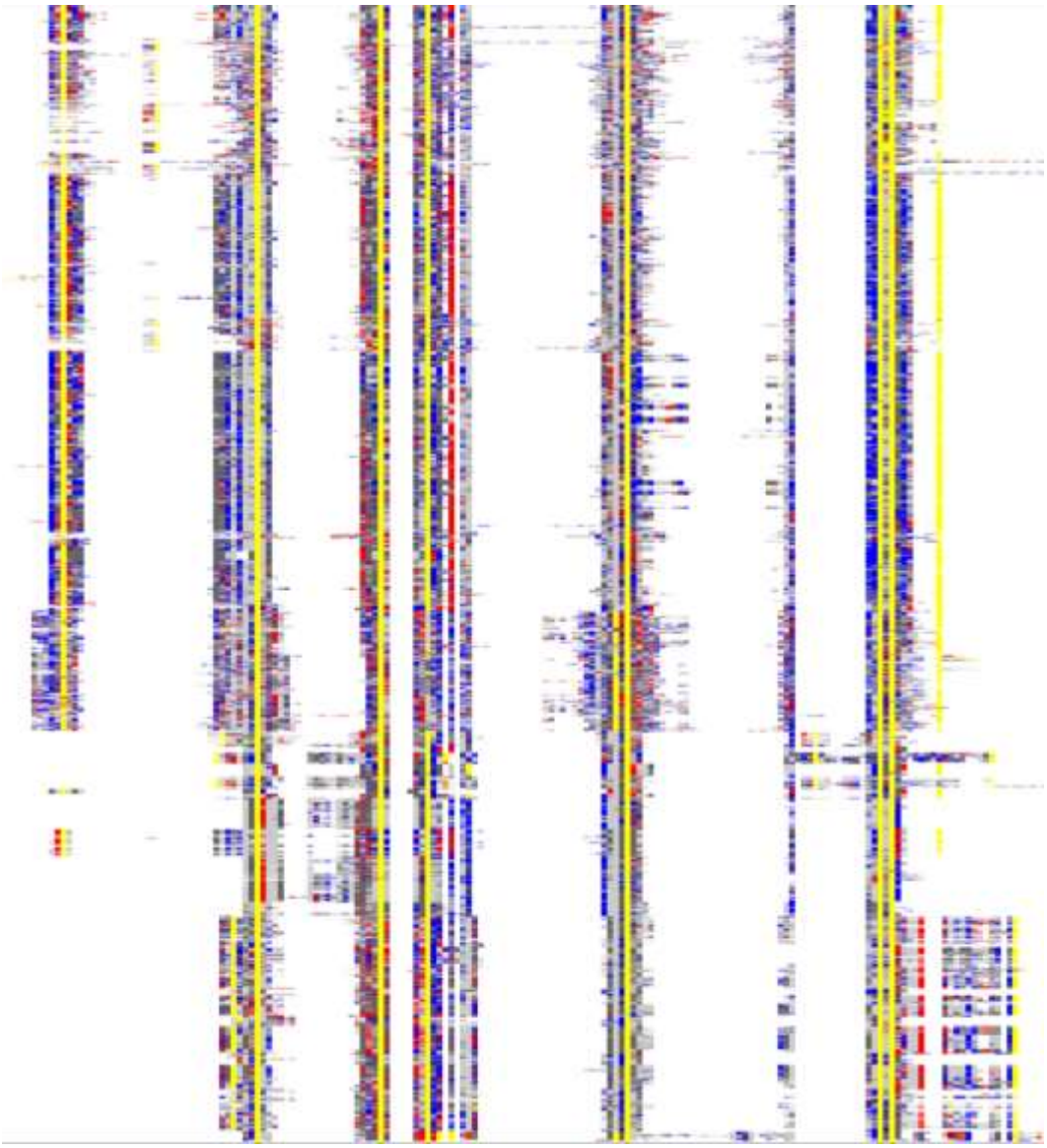
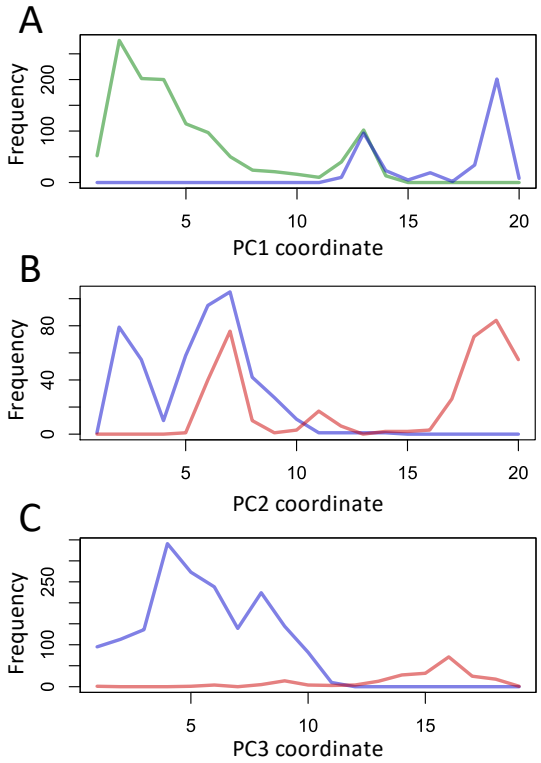


Figure S1 | Multiple sequence alignment property overview.
MSA of 2019 sequences coloured by residue type. Yellow = Cysteine; Blue = Cationic (K,R); Red =Anionic (D,E); Dark grey = polar (T,S,Q,N,H,Y); Light grey = hydrophobic (A,G,P,F,I,L,M,V,W); White = gap. Fasta file of alignment in supplementary data file 1.

https://doi.org/10.1093/bioinformatics/bty697



https://doi.org/10.1093/bioinformatics/bty697

Figure S2 | Functions segregate along the axes of the projected sequence space.
(A) Principal component axis 1. Plant in green, animal in blue. (B) Principal component axis 2. Toxin in red, non-toxin in blue. (C) Principal component axis 3. Signalling in red, non-signalling in blue.

https://doi.org/10.1093/bioinformatics/bty697

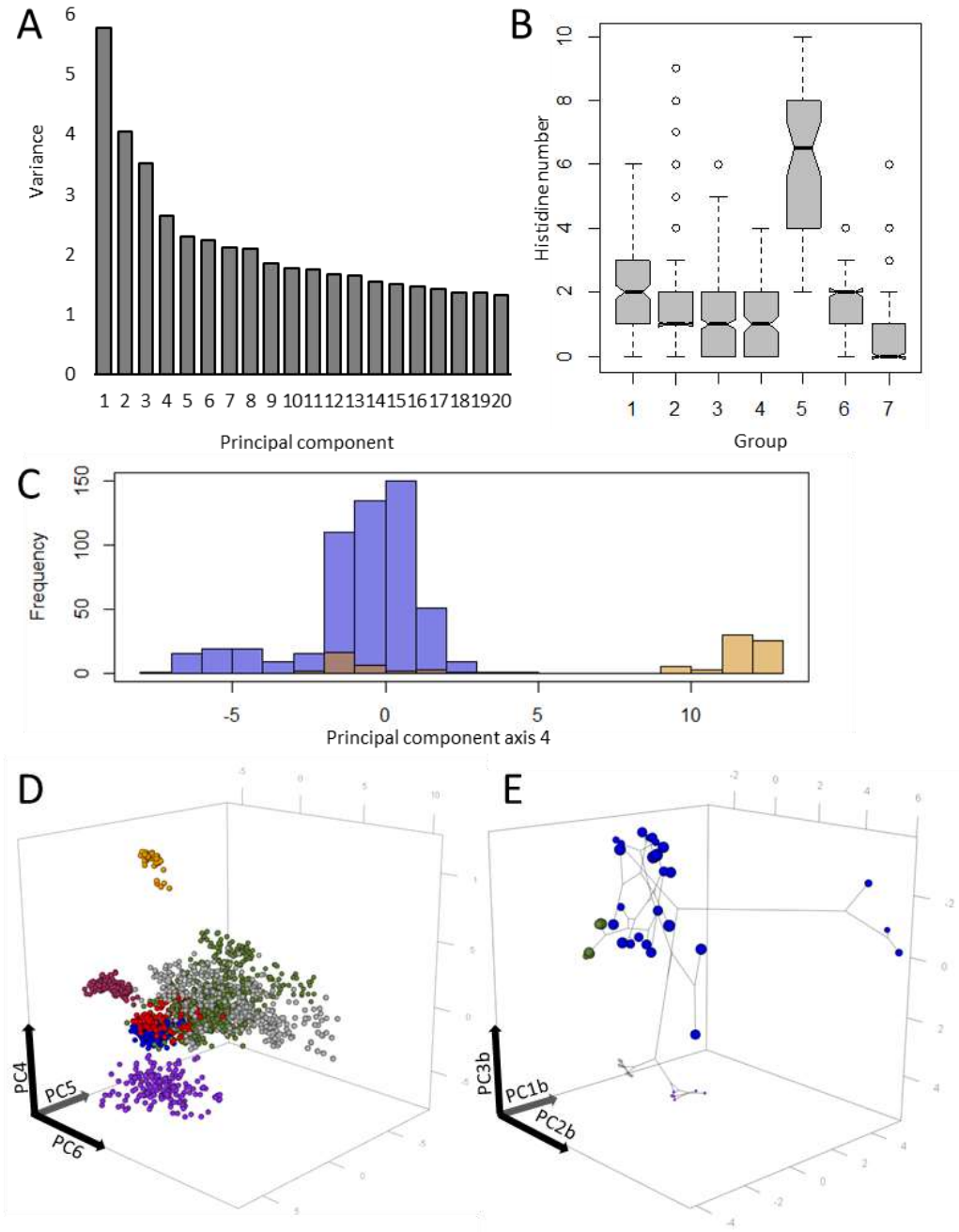


Figure S3 | Loadings variance and histidine-rich group 5.
(A) Scree plot of the variance that is encompassed by first 20 principal components. (B) The number of histidines per sequence in each group. (C). Principal component axis 4. Histidine-rich (>5), histidine-poor (≤4) in blue. (D) Sequence space plot using PCs 4-6 showing separation of the His-rich group (coloured by group as in Fig 1). (E) Subclusters within group 5 overlayed with an unrooted UPGMA tree of their sequences. Point size indicates the number of histidines (min = 2, max = 10). Colours indicate the four subclusters.

https://doi.org/10.1093/bioinformatics/bty697

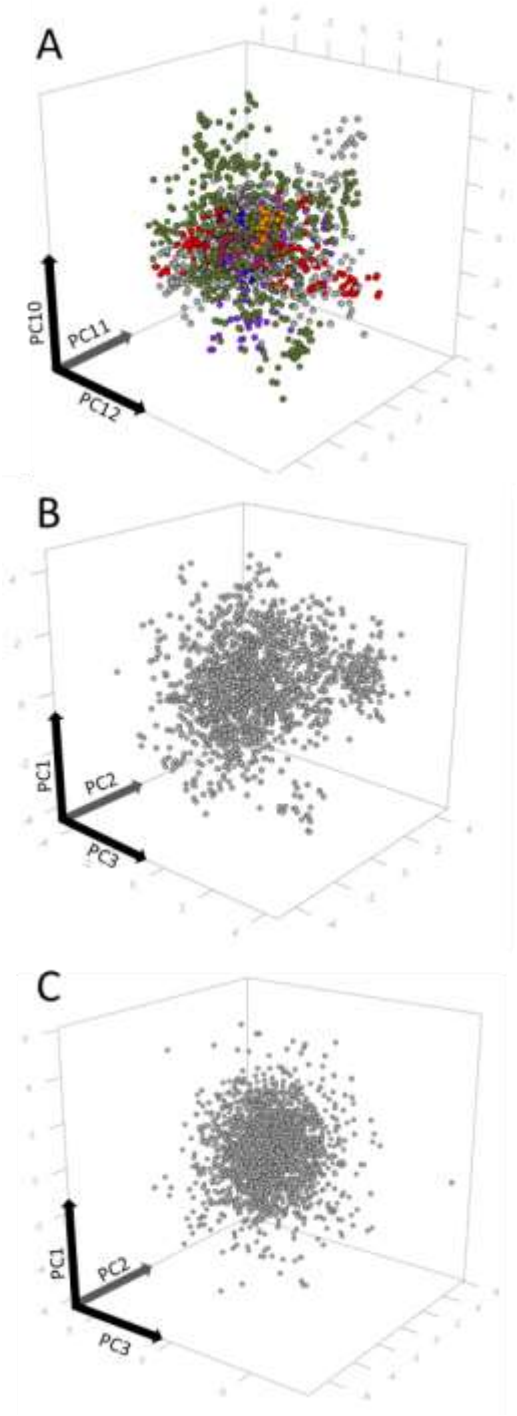


Figure S4 | Additional representations of defensin sequence space.
(A) Sequence space plot using PCs 10:12 showing random spread without well-separated groups (coloured by group as in Fig 1). (B) Sequence space of cis-defensin MSA using Cys and NGAP only (ignoring charge, hydrophobicity,

https://doi.org/10.1093/bioinformatics/bty697

disorder propensity and molecular weight). Separate groups are not discernible. (C) 2,000 random sequences of length 50aa.

https://doi.org/10.1093/bioinformatics/bty697

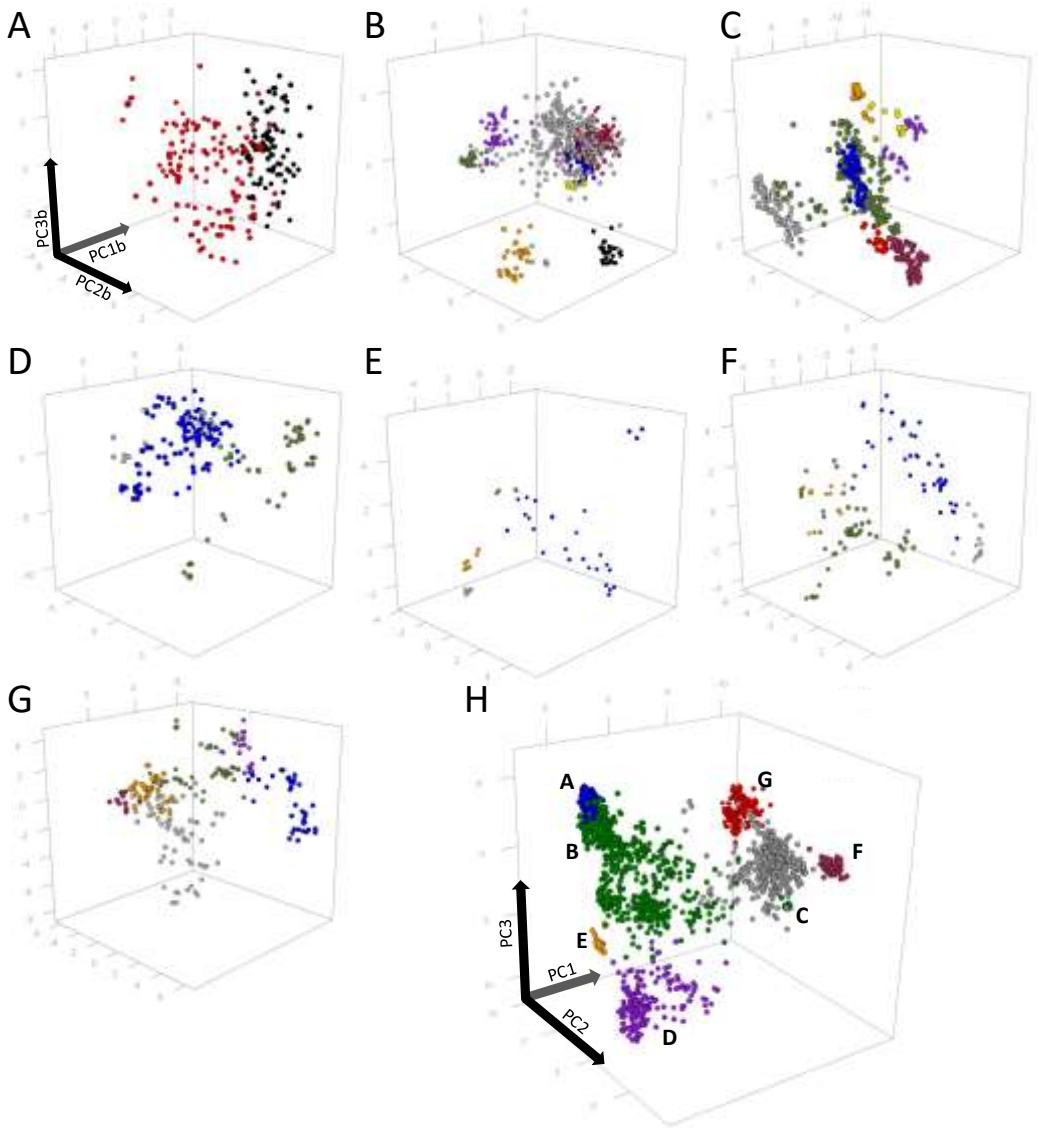


Figure S5 | Subclusters identified upon reanalysis of each main group.
Subclusters are clearly identifiable when the sequence space is calculated separately for each of the seven main groups. (A) Group 1 plant AMPs, (B) Group 2 plant AMPs, (C) Group 3 intermediate, (D) Group 4 plant signalling (E) Group 5 plant his-rich. Arranged by principal components of the group. Coloured by subclusters. (H) a summary the seven main clusters for reference.

https://doi.org/10.1093/bioinformatics/bty697

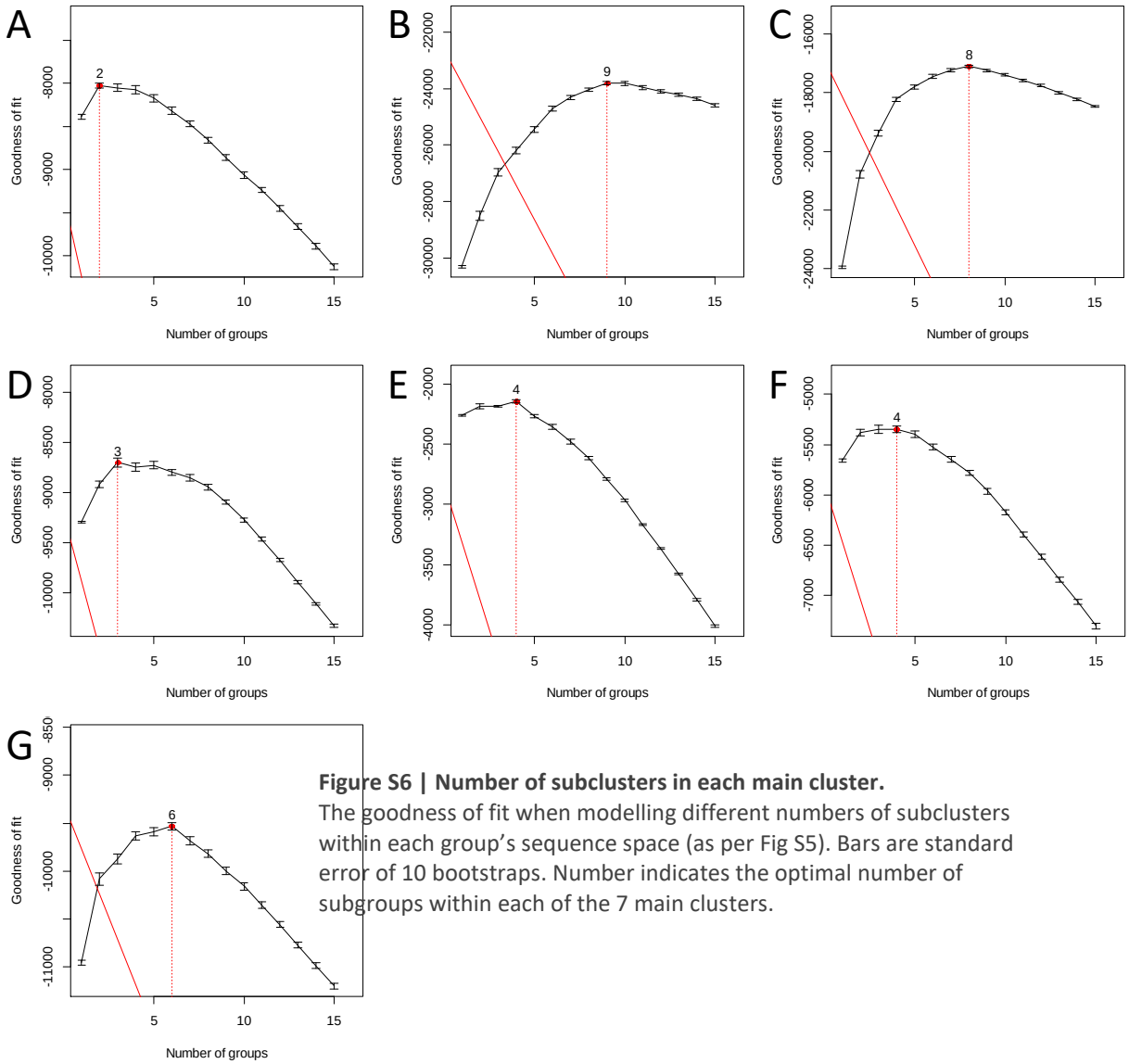


Figure S6 | Number of subclusters in each main cluster.
The goodness of fit when modelling different numbers of subclusters within each group's sequence space (as per Fig S5). Bars are standard error of 10 bootstraps. Number indicates the optimal number of subgroups within each of the 7 main clusters.

https://doi.org/10.1093/bioinformatics/bty697

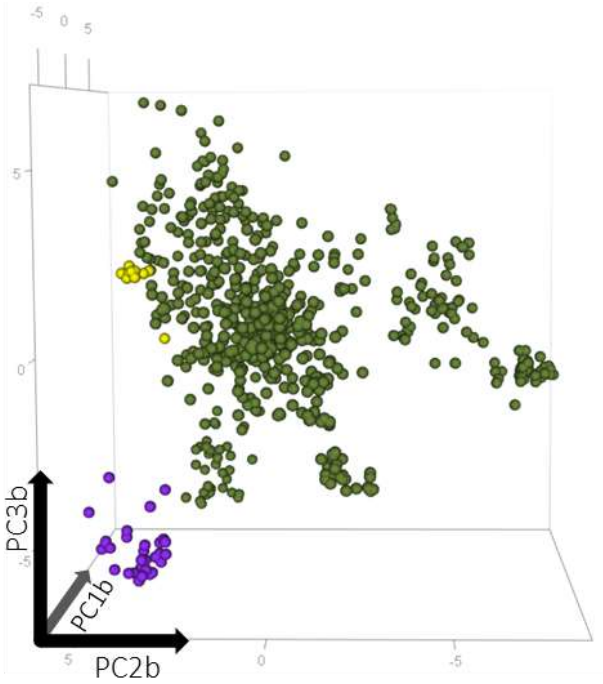
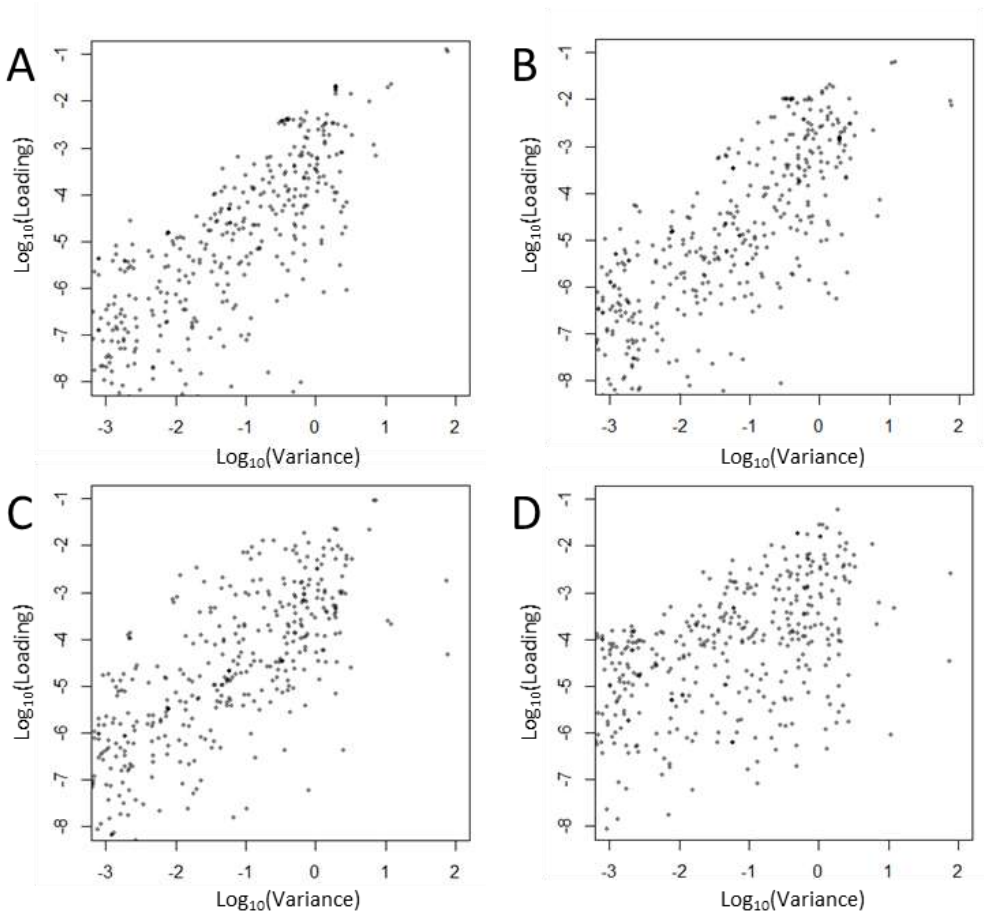


Figure S7 | Subclusters of group 2 separate C8 defensins from different taxa.
Subclusters of group 2 when the group's sequence space is calculated separately, with plant C8 defensins coloured green, mollusc C8 defensins coloured yellow, and insect C8 defensins coloured purple.

https://doi.org/10.1093/bioinformatics/bty697



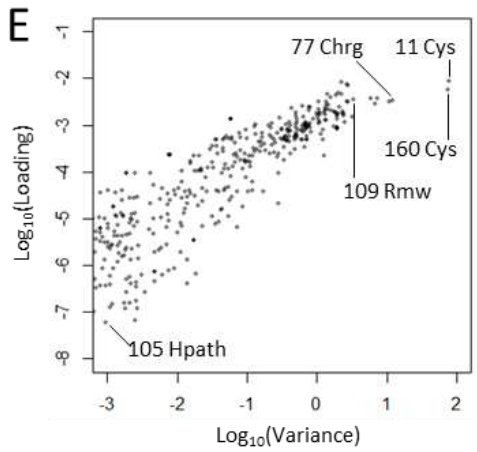


Figure S8 | The effect of sequence diversity upon residue property loading.
The dependency of loading weight up on property variance for (A) PC1, (B) PC2, (C) PC3, and (D) PC4. In general, more variable properties are more highly loaded. This can be seem more clearly for (E) the average loadings for the first 10 PCs. Notable residue properties highlighted: 11 and 160 are the terminal cysteines present in C8 defensins. Properties of 77 and 109 vary by kingdom and function. 105 is a conserved glycine that is nearly invariant except for the S-locus proteins where it is a cysteine.

https://doi.org/10.1093/bioinformatics/bty697

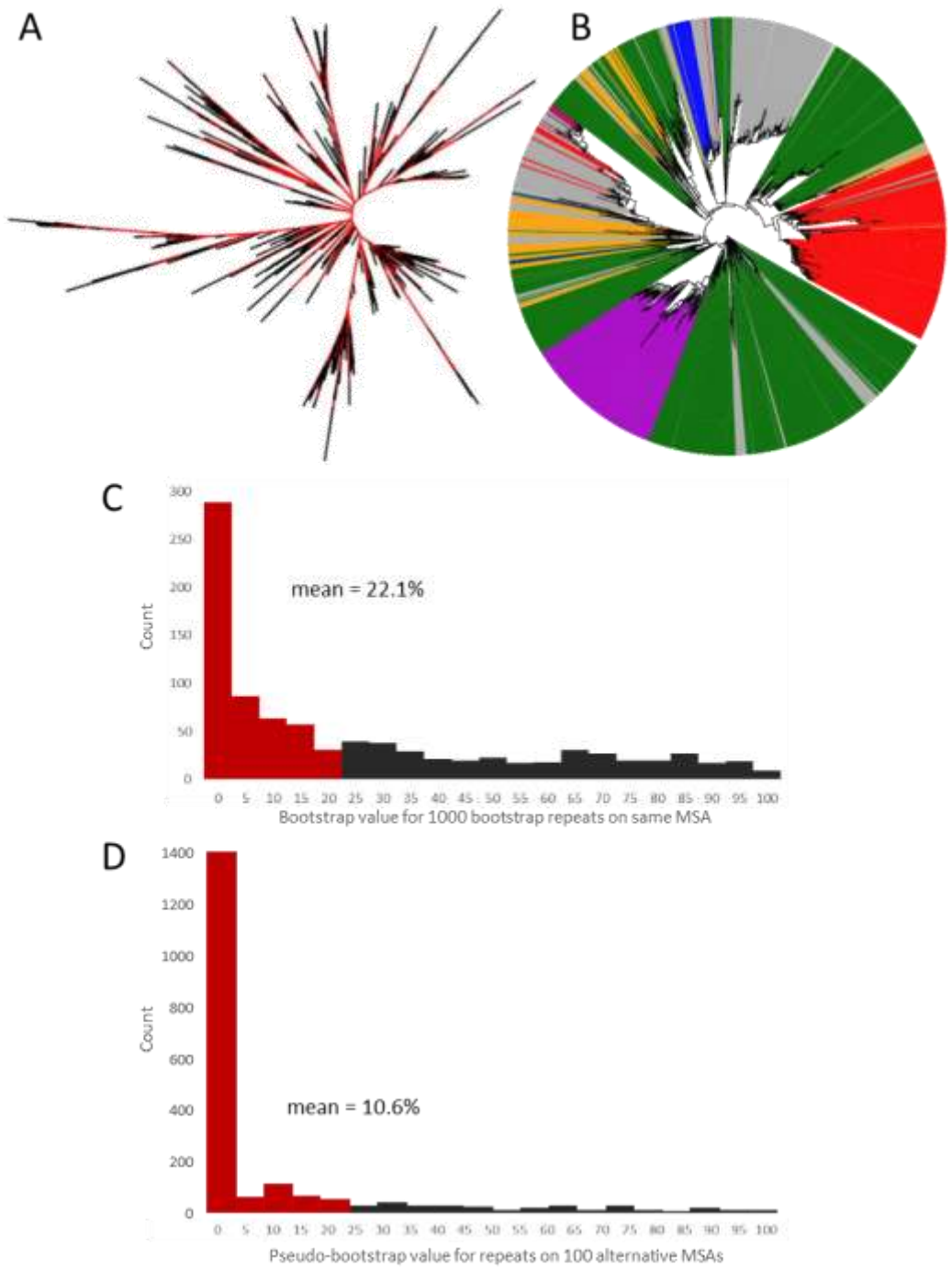


Figure S9 | Bootstrap support of maximum-likelihood phylogeny.
(A) branches with <20% bootstrap support in red, >20% in black. (B) main groups coloured onto phylogeny (note: low bootstraps as indicated in part A should be taken into account for the centre-most nodes). (C) Distribution of bootstrap values of all nodes for 1000 bootstrap repeats using same MSA. (D) Distribution of 'pseudo-bootstrap' values of all nodes for repeats using 100 alternative MSAs of same sequences.

https://doi.org/10.1093/bioinformatics/bty697

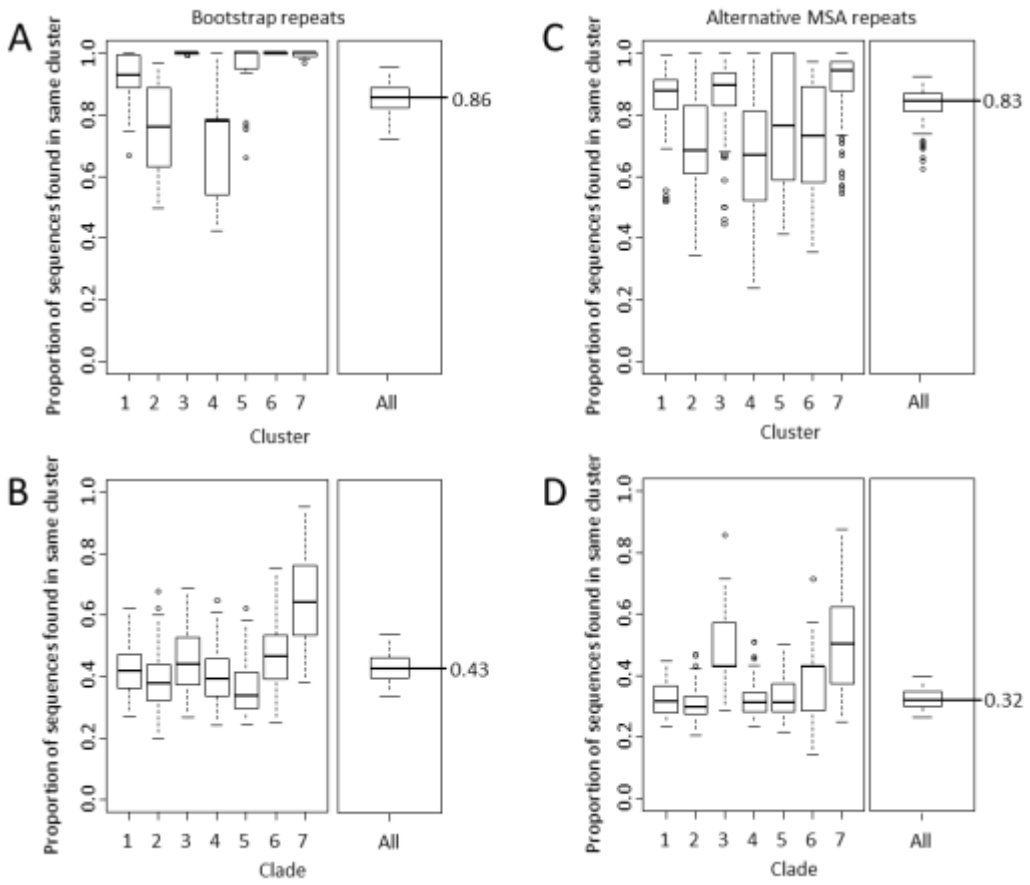


Figure S10 | Cluster stability upon bootstrap repeats or alternative MSAs.
The proportion of sequences are found in the same cluster in 100 bootstrap repeats of the same MSA using for (A) the sequence space method examining the 7 Bayesian clusters, and for (B) the maximum likelihood phylogeny, examining the seven largest clades. (C,D) The same analysis performed for 100 analyses run on alternative MSAs of the same sequences

https://doi.org/10.1093/bioinformatics/bty697

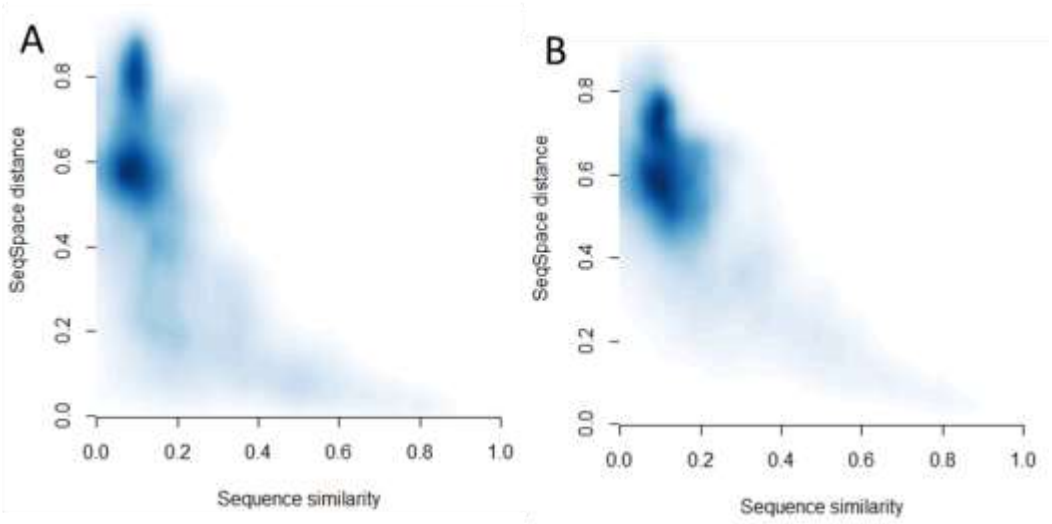


Figure S11 | Sequence space representation separates sequences with low similarity.
Most defensin sequences have very low sequence similarity, however the sequence space method arranges sequences over a wide dynamic range for (A) the first 3 PCs and (B) the first 10 PCs

https://doi.org/10.1093/bioinformatics/bty697

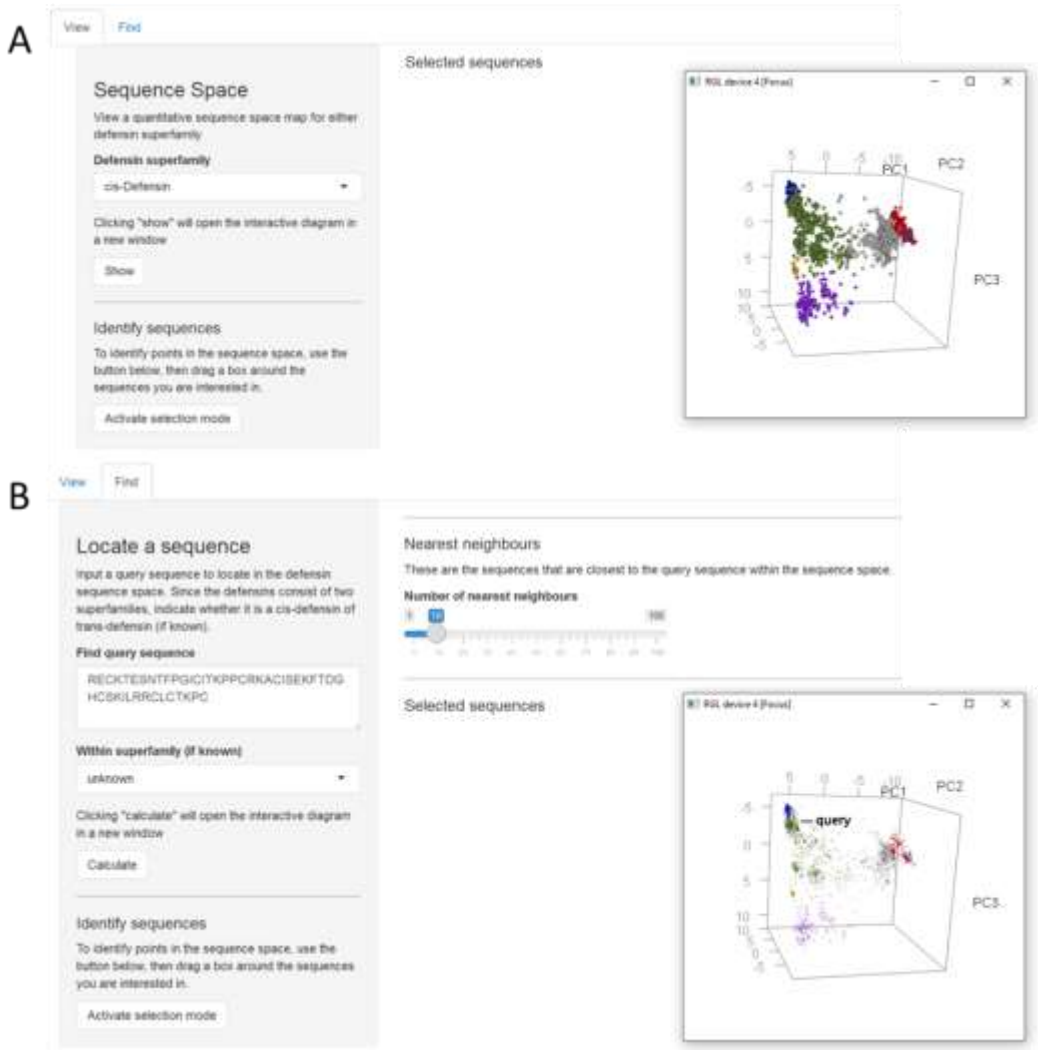
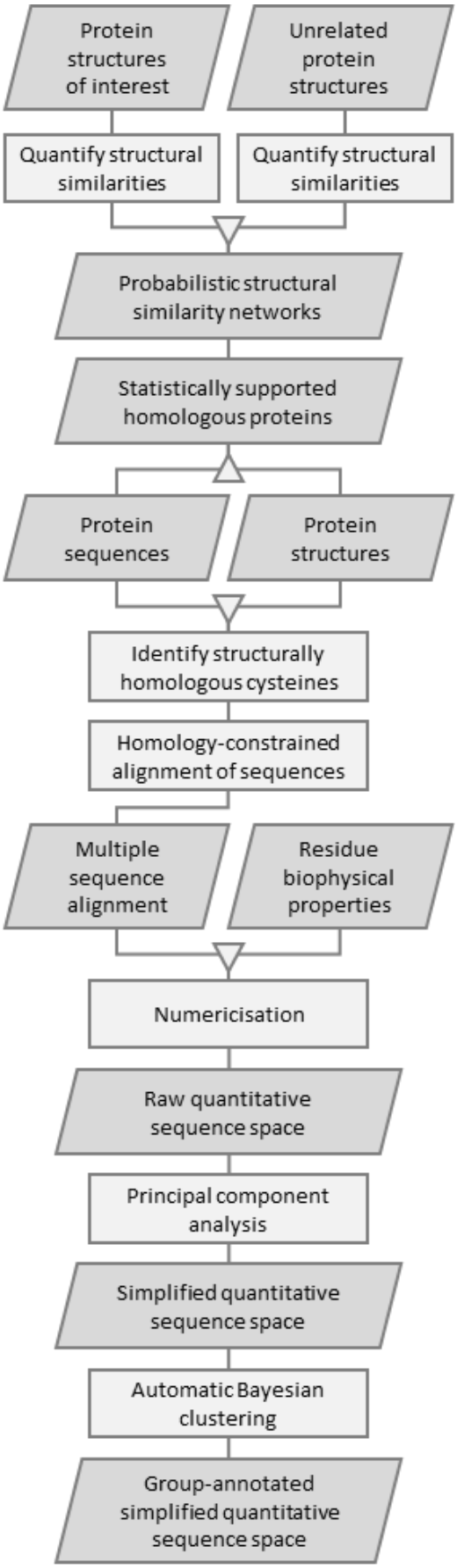


Figure S12 | Overview of webtool interface.
In addition to some basic background information (not shown) two tabs exist. (A) The view tab opens an interactive diagram of the sequence space. (B) The find tab indicates the location of a query sequence within the sequence space. If the user is unsure of the whether the sequence is from the cis- or trans- defensin superfamily, the program will attempt to assign it based on cysteine motifs and sequence similarities. In both cases, activating 'selection mode' allows the user to select one or more sequences to view them in a multiple sequence alignment.

https://doi.org/10.1093/bioinformatics/bty697



https://doi.org/10.1093/bioinformatics/bty697

Figure S13 | Flow chart of sequence space analysis workflow.
See methods section for details of individual processes. Rectangles indicate processes, parallelograms indicate inputs/outputs. Upward triangles indicate data splitting, downward triangles indicate data merging.

https://doi.org/10.1093/bioinformatics/bty697

https://doi.org/10.1093/bioinformatics/bty697

Supplementary tables

Table S1 | Sequence space residue properties for the first 4 principal components.
Most highly loaded residue properties for the first four PC axes. MSA column refers to the MSA in supplementary data Fig S1. RMW = side chain molecular weight, HPATH = hydropathy, CHRG = charge, DISORD = disorder propensity, CYS = cysteine, NGAP = occupancy.

consensus	resn	property	PC1_load	consensus	resn	property	PC2_load	consensus	resn	property	PC3_load	consensus	resn	property	PC4_load
C	11	CYS	0.40	-	173	CYS	0.28	G	105	CYS	0.31	G	105	CYS	-0.26
C	160	CYS	0.38	F	40	CYS	0.28	C	73	CYS	-0.31	C	73	CYS	0.25
-	173	CYS	-0.19	-	38	NGAP	0.17	-	111	NGAP	0.20	-	76	NGAP	0.23
F	40	CYS	-0.18	-	155	NGAP	0.17	-	110	NGAP	0.20	V	72	HPATH	-0.16
C	11	NGAP	0.18	-	39	NGAP	0.16	-	102	NGAP	0.16	-	13	CHRG	0.15
-	10	NGAP	0.18	F	151	DISORD	-0.16	E	77	CHRG	0.16	Y	106	HPATH	0.15
-	12	NGAP	0.18	F	151	RMW	0.15	F	80	DISORD	0.15	F	80	RMW	-0.15
-	9	NGAP	0.17	F	40	NGAP	0.15	-	134	DISORD	0.14	G	109	RMW	0.14
-	13	NGAP	0.17	-	154	NGAP	0.13	-	8	NGAP	0.13	T	153	DISORD	-0.14
C	160	NGAP	0.17	-	165	NGAP	0.13	-	101	NGAP	0.13	-	48	NGAP	-0.13
-	37	NGAP	0.16	-	173	NGAP	0.13	-	112	NGAP	0.13	G	104	RMW	0.12
G	42	RMW	-0.16	C	11	CYS	0.13	R	149	HPATH	0.12	R	148	CHRG	-0.12
-	14	NGAP	0.15	-	172	NGAP	0.13	-	154	DISORD	-0.12	-	110	NGAP	0.12
E	77	CHRG	-0.13	-	170	NGAP	0.13	-	134	HPATH	-0.12	G	109	CHRG	0.12
-	81	NGAP	-0.10	-	164	NGAP	0.13	F	80	RMW	-0.12	A	66	HPATH	-0.12
-	61	NGAP	-0.10	-	166	NGAP	0.13	G	104	RMW	0.12	T	153	RMW	0.12
R	149	RMW	0.09	-	163	NGAP	0.13	G	109	RMW	0.12	-	111	NGAP	0.12
-	163	NGAP	-0.09	-	161	NGAP	0.13	-	41	CHRG	-0.12	-	12	DISORD	-0.11
-	169	NGAP	-0.09	-	169	NGAP	0.13	R	149	CHRG	-0.12	F	40	HPATH	-0.11
-	157	NGAP	-0.09	C	160	CYS	0.12	-	7	NGAP	0.12	-	134	DISORD	0.11
-	164	NGAP	-0.09	-	41	NGAP	0.12	V	72	HPATH	-0.11	-	155	DISORD	0.11
-	103	NGAP	0.09	G	45	HPATH	0.12	-	154	HPATH	0.11	-	154	CHRG	-0.11
-	165	NGAP	-0.09	-	157	NGAP	0.12	-	133	NGAP	0.11	-	8	NGAP	-0.11
-	161	NGAP	-0.09	-	103	NGAP	-0.12	-	47	NGAP	0.11	-	134	RMW	-0.11
-	170	NGAP	-0.09	-	156	NGAP	0.12	-	154	CHRG	-0.11	-	133	DISORD	-0.10
-	173	NGAP	-0.09	F	80	RMW	-0.11	D	62	NGAP	-0.11	K	74	CHRG	0.10
-	156	NGAP	-0.09	-	162	NGAP	0.11	-	48	NGAP	0.11	-	13	RMW	0.10
-	172	NGAP	-0.09	-	81	NGAP	0.10	-	6	NGAP	0.10	R	149	RMW	-0.10
R	149	CHRG	0.08	-	134	NGAP	0.10	-	100	NGAP	0.10	-	154	DISORD	-0.10
E	77	DISORD	0.08	G	45	DISORD	-0.10	C	11	CYS	0.10	-	7	NGAP	-0.10
-	166	NGAP	-0.08	F	80	HPATH	0.10	G	105	RMW	0.10	Y	106	DISORD	-0.09
R	149	HPATH	-0.08	E	77	CHRG	-0.10	G	45	RMW	-0.09	V	72	RMW	0.09
-	162	NGAP	-0.08	-	61	NGAP	-0.09	E	77	DISORD	-0.09	E	77	DISORD	-0.09
F	40	DISORD	-0.08	V	72	HPATH	0.09	F	151	DISORD	0.09	C	160	CYS	0.09
F	40	RMW	0.08	G	109	DISORD	-0.09	G	45	DISORD	0.09	-	103	CHRG	0.09
R	148	RMW	0.08	-	47	NGAP	-0.09	A	66	RMW	0.09	S	63	DISORD	-0.09
F	80	RMW	0.08	-	60	NGAP	-0.09	G	79	RMW	0.08	-	101	NGAP	-0.09
R	148	CHRG	0.08	R	148	RMW	0.09	-	49	NGAP	0.08	G	45	DISORD	0.09
G	42	DISORD	0.08	-	58	NGAP	-0.09	F	151	HPATH	-0.08	G	45	HPATH	-0.09
-	155	DISORD	0.08	K	74	HPATH	-0.09	C	73	DISORD	-0.08	-	103	HPATH	-0.09
F	80	DISORD	-0.08	-	59	NGAP	-0.09	-	9	CHRG	-0.08	-	102	NGAP	-0.09
-	38	NGAP	0.08	-	48	NGAP	-0.08	G	45	HPATH	-0.08	-	6	NGAP	-0.09
-	155	HPATH	-0.08	T	153	CHRG	-0.08	C	73	RMW	0.07	-	13	HPATH	-0.08
-	39	NGAP	0.07	-	135	RMW	-0.08	F	80	HPATH	-0.07	-	47	NGAP	-0.08
-	154	CHRG	0.07	-	155	HPATH	0.08	R	149	DISORD	-0.07	-	14	NGAP	0.08
-	154	RMW	0.07	G	109	RMW	0.08	-	134	NGAP	0.07	R	148	HPATH	0.08
F	40	NGAP	0.07	-	53	NGAP	-0.08	-	14	NGAP	-0.07	F	40	RMW	0.08
G	109	RMW	-0.07	N	64	RMW	0.08	R	149	RMW	-0.07	-	49	NGAP	-0.08
-	135	RMW	0.06	-	14	NGAP	0.07	-	113	NGAP	0.07	-	154	HPATH	0.08

https://doi.org/10.1093/bioinformatics/bty697

Y	106	DISORD	0.06	N	64	DISORD	-0.07	-	156	NGAP	0.06	G	105	RMW	-0.08

Table S2 | Cysteine motifs found in the two defensin superfamilies.
Motifs encompass variation from the 5th percentile to the 95th percentile (full ranges are skewed by outliers). Motifs written as regular expressions for text searches. Accessions are PDB IDs where available, Genbank ID otherwise.

Description	Motif	Example	Accession
C8 defensin	C.{10}C.{5}C.{3}C.{9,10}C.{6,8}C.C.{3}C	NaD1	1MR4
C6 defensin	C.{5,12}C.{3}C.{9,10}C.{4,5}C.C	NvD1	2KOZ
Petunia defensin	C.{3}C.{6}C.{5}C.{2}CC.{9,11}C.{6}C.C.{3}C	PhD1	1N4N
Mollusc defensin	C.{6}C.{3}C.{4,5}C.{4}C.{8}C.C.{2}C	MgD1	1FJN
Hydra defensin	C.{6}C.{14}C.{3}C.{9}C.{6}C.{8}C.C	Hydramacin1	2K35
Annelid defensin	C.{6}C.{14}C.{3}C.{1,2}C.{7}C.{6,7}C.{5,9}C.C.{11,16}C	Theromacin	2LN8
alpha-toxin	C.{3}C.{5,6}C.{3}C.{9}C.{6,9}C.C.{14,15}C	BmK M1	1SN1
Maurotoxin	C.{5}C.{3}C.{5}C.{4}C.{4}C.C.{2}C	Maurotoxin	1TXM
Excitotoxin	C.{10}C.{3}C.{10}CC.{3}C.C.{19}C	Bj-xtrIT	1BCG
Small-toxin	C.{2}C.{10}C.{2}CC.{5,6}C.{4}C.C	Insectotoxin 15A	1SIS
S-locus 11	C.{3,9}C.{6,7}C.{3,15}C.{1,9}C.{8,9}C.C.{3,14}C	S8-SP11	1UGL
alpha-defensin	.C.C.{4}C.{9}C.{9}C.	HD5	2LXZ
beta-defensin	.{4}C.{6}C.{3,4}C.{9}C.{5,6}C.	HBD1	1IJV
theta-defensin	.C.C.{4}C	Retrocyclin2	2ATG
Big defensin	.{45,51}C.{6}C.{3}C.{13}C.{4}C.	TtBigDef	2RNG
Plant fusion defensin	C.{3,5}C.{4,8}C.{3}C.{9,11}C.{5,9}CCC	MtD36	357449491
Fungal N-ter defensin	C.{5}C.{7}C.{3}C.{10}C.{5}C.{5}C.C	Cglosin 1N	88178907
Fungal C-ter defensin	CC.{9}CC.{3}C.{9,10}C.{5}C.C	Cglosin 1C	88178907
S-locus 11 b	C.{9}C.{7}C.{14}C.C.C.{8}C.C.C.C	BoS14	283131299
S-locus 11 c	C.{9}C.{9}C.{16}C.C.{9}C.C.{3}C.{2}CC	PtS2	550331862
S-locus 11 d	C.C.{8}C.{7}C.{15}C.C.{8}C.C.{4}C.{6}C	BrS14	90819164
S-locus 11 e	C.{9}C.{6}C.{7}C.{6}C.C.{8,11}C.C	BoS7	283131295
S-locus 11 f	C.{5,9}C.{7}C.{12,17}C.C.{1,2}C.{6,10}C.C	EsS2	557114862
S-locus 11 g	C.{9,10}C.{7,8}C.{13,17}C.C.{9,12}C.C.{3,4}C	AtS32	254763280
+CCC	C.{10}C.{5}C.{3}C.{9,10}C.{6,8}CCC	AtPDF1.3	15225238
+CxCC	C.{10}C.{5}C.{3}C.{9,10}C.{6,8}C.CC	DLP96	332659178
+CxCxC	C.{10}C.{5}C.{3}C.{9,10}C.{6,8}C.C.C	Fabatin-2	3913646

https://doi.org/10.1093/bioinformatics/bty697

Table S3 | Residue properties used to characterise defensin sequences.
Net charge in Coulombs, disorder propensity as in TOP-IDP (Campen *et al.*, 2008), hydrophobicity as in the Doolittle index (Kyte and Doolittle, 1982), molecular weight of [R] group in Daltons, and disulphide potential and occupancy as binary descriptors.

	Net charge	***Disorder propensity***	***Hydro-phobicity***	***Molecular weight (R)***	***Disulphide potential***	***Occupancy***
A	0	0.06	1.8	15.09	0	1
C	0	0.02	2.5	47.16	1	1
D	-1	0.19	-3.5	59.10	0	1
E	-1	0.74	-3.5	73.13	0	1
F	0	-0.70	2.8	91.19	0	1
G	0	0.17	-0.4	1.07	0	1
H	0	0.30	-3.2	81.16	0	1
I	0	-0.49	4.5	57.18	0	1
K	1	0.59	-3.9	72.19	0	1
L	0	-0.33	3.8	57.18	0	1
M	0	-0.40	1.9	75.21	0	1
N	0	0.01	-3.5	58.12	0	1
P	0	0.99	-1.6	41.13	0	1
Q	0	0.32	-3.5	72.15	0	1
R	1	0.18	-4.5	100.2	0	1
S	0	0.34	-0.8	31.09	0	1
T	0	0.06	-0.7	45.12	0	1
V	0	-0.12	4.2	43.15	0	1
W	0	-0.88	-0.9	130.23	0	1
Y	0	-0.51	-1.3	107.19	0	1
X	0	0.03	-0.5	62.90	0	1
-	NA	NA	NA	NA	NA	0