Extreme Convergent Evolution in Defensin Proteins & Quantitative Maps of Their Sequence Space

Thomas Shafee | Fung Lay | Mark Hulett | Marilyn Anderson
La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia
Thomas.Shafee@Gmail.com

Background : Defensins are small, charged, disulphide-rich eukaryotic proteins with diverse sequences, structures, and functions. Their antimicrobial activities are of particular interest for protecting crops and humans from pathogens.

They have been traditionally treated as a single superfamily. However, we present evidence that there exist two independent evolutionary origins of defensins, based on their secondary structure element order, disulphide topology, and tertiary structures. These two superfamilies, the *cis*-defensins and *trans*-defensins, exhibit some of the most extensive convergent evolution of protein sequence, structure and function.

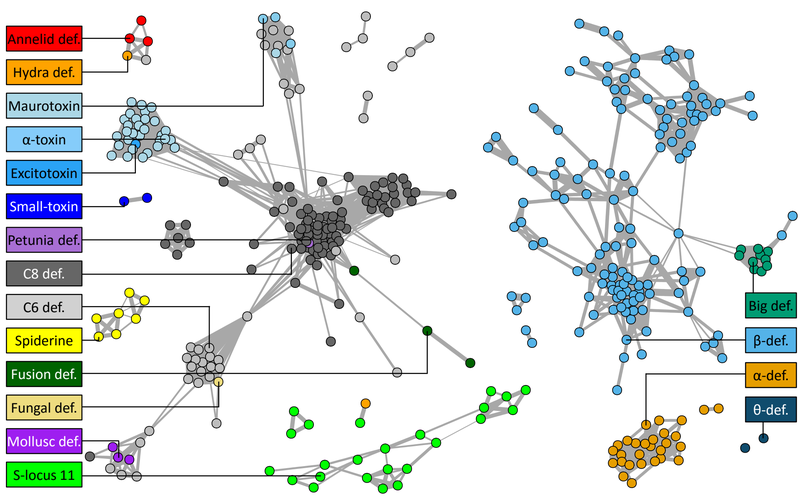
We have developed new methods of sequence alignment and analysis to overcome the difficulties of investigating such short and divergent sequences. Multivariate analysis of protein sequence space allows grouping of defensins into naturally occurring clusters which describe the residue properties that separate phyla and functions. It can be further used to design synthetic, cluster-central, archetypal defensin sequences.

1 | Evidence of Independent Evolutionary Origins

Sequence similarity is insufficient to establish relatedness in such divergent proteins. Defensins motifs occur in 4% of random cys-rich sequences.

Example of an erroneous published alignment between a cis-defensin and trans-defensin

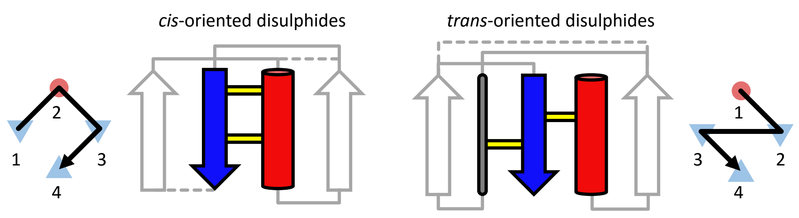
```
C-VSSGG-----DCLYSACFIFTKIQGTCYRGKAICCC  HBD1|H.Sapiens|P60022
CDLLSKWNWNHTACAGH-CIAKGFKGGVCNDKAV-CVCR  DmD1|D.melanogaster|29367272
```



A sequence-similarity network identifies several clusters of similar sequences. Coloured by structural class (or cysteine motif for sequences of unknown structure)

Primary Structure

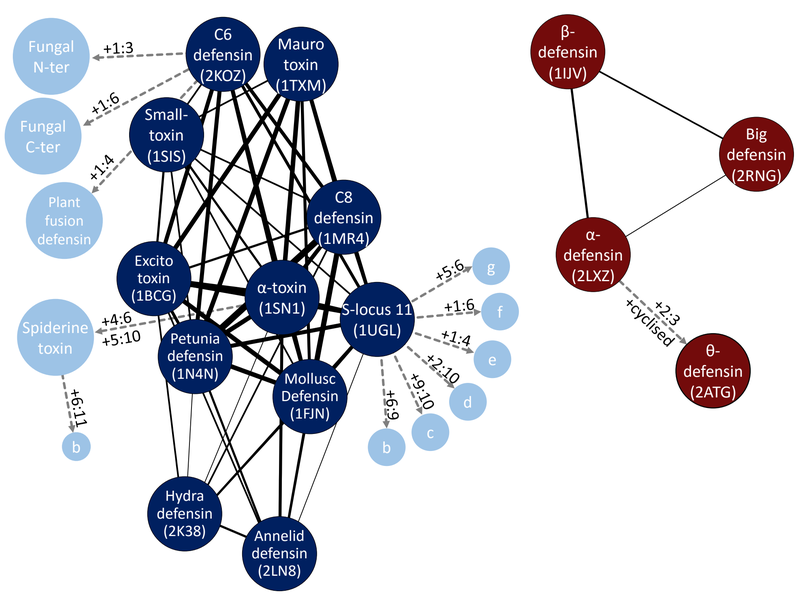
The order and orientation of the secondary structure elements precludes conversion from one fold to the other by simple rearrangements, such as circular permutation. The most conserved feature is the orientation and connectivity of the disulphide bonds from the C-terminal β-strand, which is oriented parallel to the α-helix.



The order of secondary structure elements in the two superfamilies. In one superfamily, a CxC motif in the final β-strand orients a pair of disulphide towards the same direction, in the other superfamily, a CC motif forces disulphide to bind two different secondary structure elements

Secondary Structure

Structural similarity separates the defensins into two independent superfamilies. Secondary structure and disulphide orientation and is conserved within each group, but differ between them.

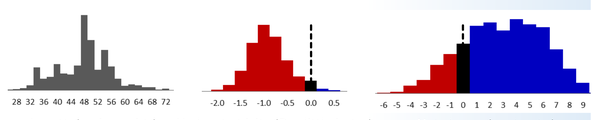


A structural similarity network separates the defensins into two groups. Dark circles indicate disulphide classes with solved structures, light circles represent disulphide classes for which the structure is not yet known. Black lines indicate probable structural homology (widths proportional to relatedness Z-score). Dashed arrows indicate evidence of common origin from cysteine motif or gene organisation. Numbers adjacent to arrows indicate additional disulphide bonds of that class.

Tertiary Structure

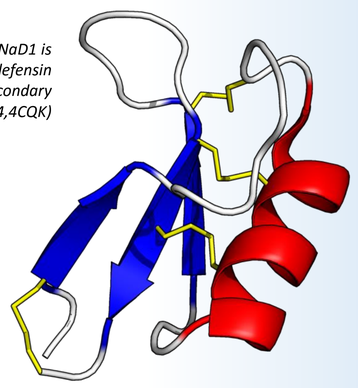
2 | Extent of Convergent Evolution

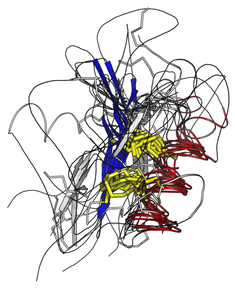
cis-Defensins



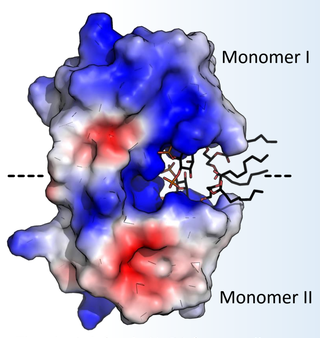
Sequence properties of 1820 cis-defensins

The plant defensin NaD1 is a prototypical cis-defensin Coloured by secondary structure (PDB:1MR4,4CQK)





Overlay of 12 cis-defensins with the conserved secondary structure coloured



The plant defensin NaD1 forms a dimer and binds the polar head groups of lipids to disrupt membranes Coloured by surface charge (PDB:4CQK)

Convergence

Sequence

- Short, cationic, cysteine-rich
- Residue bias due to relaxed hydrophobic core constraints
- GxC motif due to packing constraints
- Extreme sequence divergence

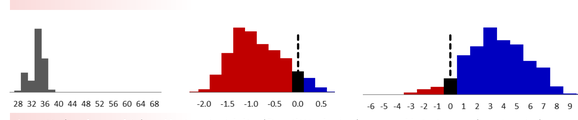
Structure

- Stable disulphide core displaying flexible surface loops
- Scaffold elaborated by addition and reduction of loops and disulphides
- Some dimerise via extended β-sheet
- Some additionally form higher-order oligomers

Function

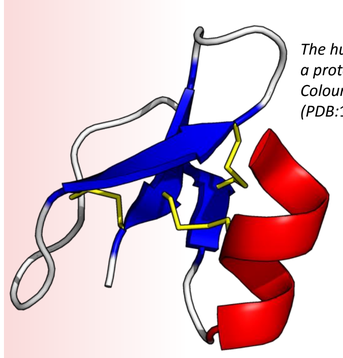
- Antimicrobial activity by lipid binding and cell membrane disruption
- Neurotoxic activity by ion channel binding and perturbation
- Signalling via cell-surface receptors

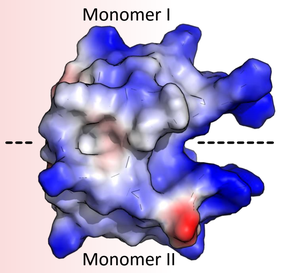
trans-Defensins



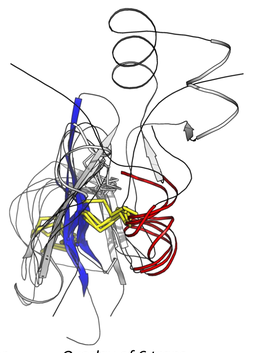
Sequence properties of 839 trans-defensins

The human defensin HBD1 is a prototypical trans-defensin Coloured by secondary structure (PDB:1IJV)

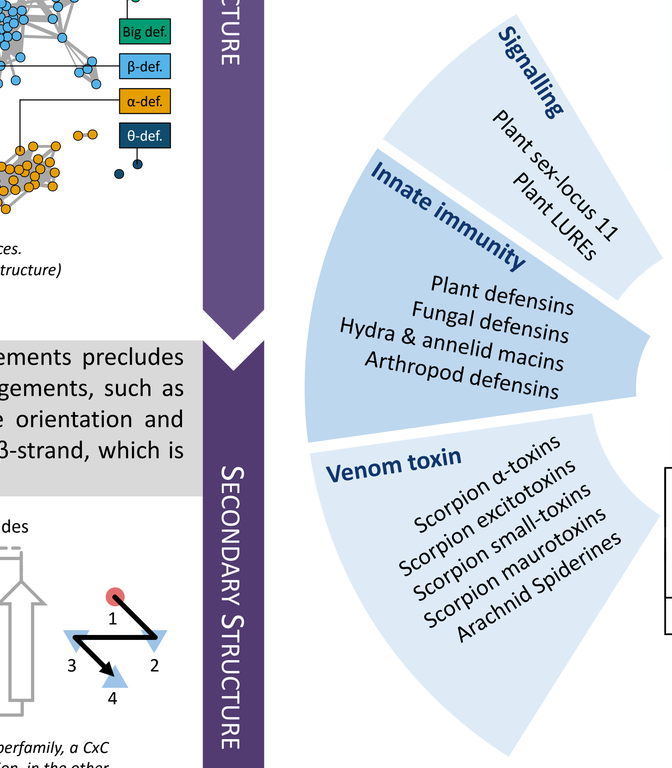




The human defensin HBD2 forms a dimer and displays an analogous cationic loop and is also known to bind lipids. Coloured by surface charge (PDB:1FD4)

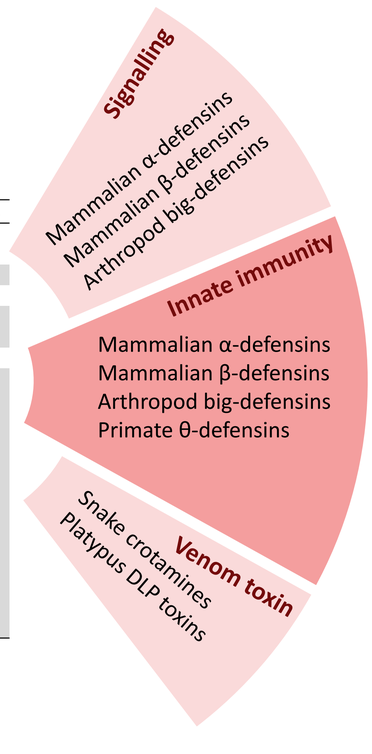


Overlay of 6 trans-defensins with the conserved secondary structure coloured



Phylum	cis-defensins	trans-defensins
Vertebrata		α | β | θ
Cephalochordata	C6	Big
Other deutrostomia		
Mollusca	C6 | C8 | Mollusc | Annelid | Hydra	Big
Annelida	Annelid | Hydra	
Other lophotrochozoa		
Arthropoda	C6 | C8 | α-Tx | Small-Tx | MauroTx | ExcitoTx | Mollusc | *Spiderine* (*a*, *b*)	Big | β
Nematoda	C6 | Annelid	
Other ecdysozoa		
Cnidaria	C6 | Hydra	
Porifera		
Ascomycota	C6 | *Fungal N* | *Fungal C*	
Basidomycota	C6	
Glomeromycota	C6	
Zygomycota	C6	
Other fungi		
Angiospermae	C6 | C8 | Petunia | *Fusion* | S-locus (a, *b*, *c*, *d*, *e*, *f*, *g*)	
Gymnospermae	C6 | C8	
Other viridiplantae		
Other eukarya		

A simplified phylogeny of Eukaryotic phyla, annotated with the occurrence of different structural classes (or cysteine motif in italics for sequences of unknown structure). Phyla with no known defensins from each of the superfamilies are filled in grey.



3 | Mapping Protein Sequence Space

Given the inherent limitations of phylogenetic-based sequence grouping for divergent proteins, sequence-space-based clustering methods were developed for the defensins using Principal Component Analysis (PCA)

Align
- Align low-similarity, low-complexity, high-indel sequences
  - Replace structurally homologous cysteines with barcode sequences
  - Re-align free loops
  - Restore cysteines

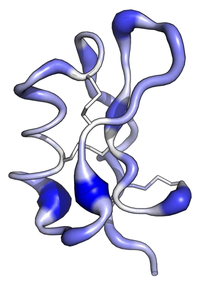
Numericise
- Convert sequences to matrix of quantitative residue properties
  - Size
  - Charge
  - Hydrophobicity
  - Disorder propensity
  - Cysteine

Transform
- Simplification of multidimensional space with PCA
  - PCs summarise main covariation
  - Loadings report key residue properties

Cluster
- Bayesian clustering of sequences with similar properties
  - 8 clusters optimally describe broad groupings

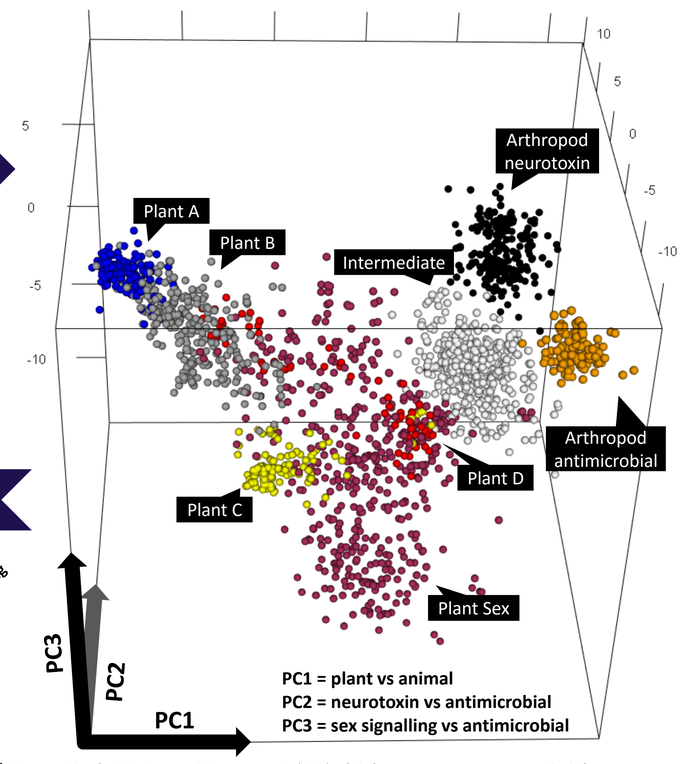
The main PCA resolves separates phyla and activities where the phylogeny failed to resolve deep roots. The PC loadings explain the relevant sequence features of these separations in an easily-interpretable manner. Subsequent sub-cluster PCAs should allow for archetype sequence design.

Interpret



	Consensus	Residue #	Property	Loading
PC3	S	46	HPATH	0.18
	F	134	HPATH	0.18
	C	73	CYS	0.17
	G	105	CYS	-0.17
	F	110	MW	-0.16
	etc.			

Loadings for each principal component indicate the main contributing residue properties. Loading magnitude be mapped to the structure (left, PDB:1UGL) or the multiple sequence alignment (right).



The first 3 Principal Components (PCs) of defensin sequence space with defensins indicated as points. The PCs separate sequences by kingdom and function. Coloured by groups identified by Bayesian clustering.

Conclusions : Understanding defensin evolution requires specialised techniques. Their secondary and tertiary structure indicates that they consist of two independent superfamilies. The *cis*-defensins and *trans*-defensins have undergone extreme convergent evolution of their sequences, structures and functions. Due to their extreme divergence, analysis of defensin sequence space surpasses traditional phylogenetic methods. It identifies naturally occurring clusters that successfully predict activities and kingdoms based on their sequence properties.

References :

1. T Shafee, A Robinson, N van der Weerden, M Anderson (2015) *SpringerPlus* Structural homology guided alignment of cysteine rich proteins
2. T Shafee*, F Lay, M Hulett, M Anderson (2016) *Molecular Biology and Evolution* The defensins consist of two independent, convergent protein superfamilies
3. T Shafee, F Lay, M Hulett, M Anderson (2016) *Cell and Molecular Life Sciences* Convergent Evolution of Defensin Sequence, Structure and Function
4. J Payne, M Bleackley, T Lee, T Shafee, I Poon, M Hulett, M Aguilar, N van der Weerden, M Anderson (2015) *Biochimica et Biophysica Acta biomembranes* The plant defensin NaD1 introduces membrane disorder through a specific interaction with the lipid, phosphatidylinositol 4,5 bisphosphate
5. *T Shafee*, M Anderson (*2016 in prep*) A quantitative map of protein sequence space for the cis-defensin superfamily