Video Article

Creating and Applying a Reference to Facilitate the Discussion and Classification of Proteins in a Diverse Group

D. Ellen K. Tarr[1]

[1]Department of Microbiology and Immunology, Arizona College of Osteopathic Medicine, Midwestern University

Correspondence to: D. Ellen K. Tarr at ellentarr@gmail.com

URL: https://www.jove.com/video/56107
DOI: doi:10.3791/56107

Keywords: Behavior, Issue 126, Protein superfamily, protein nomenclature, protein classification, sequence alignment, phylogeny, invertebrate defensins, CS-αβ superfamily

Date Published: 8/16/2017

Citation: Tarr, D.E. Creating and Applying a Reference to Facilitate the Discussion and Classification of Proteins in a Diverse Group. *J. Vis. Exp.* (126), e56107, doi:10.3791/56107 (2017).

Abstract

Related proteins that have been studied in different labs using varying organisms may lack a uniform system of nomenclature and classification, making it difficult to discuss the group as a whole and to place new sequences into the appropriate context. Developing a reference that prioritizes important sequence features related to structure and/or activity can be used in addition to established names to add some coherency to a diverse group of proteins. This paper uses the cysteine-stabilized alpha-helix (CS-αβ) superfamily as an example to show how a reference generated in spreadsheet software can clarify relationships between existing proteins in the superfamily, as well as facilitate the addition of new sequences. It also shows how the reference can help to refine sequence alignments generated in commonly used software, which impacts the validity of phylogenetic analyses. The use of a reference will likely be most helpful for protein groups that include highly divergent sequences from a broad spectrum of taxa, with features that are not adequately captured by molecular analyses.

Video Link

The video component of this article can be found at https://www.jove.com/video/56107/

Introduction

A protein's name should reflect is characteristics and relationship to other proteins. Unfortunately, names are generally assigned at the time of discovery and, as research continues, the understanding of the larger context may change. This can lead to multiple names if a protein was independently identified by more than one lab, to changes in nomenclature or in the characteristics thought to be definitive when assigning the name, and to the name no longer sufficiently differentiating the protein from others.

Invertebrate defensins provide a good example of degeneration in nomenclature and classification. The first invertebrate defensins were reported from insects, and the name "insect defensin" was proposed based on the perceived homology to mammalian defensins[1,2]. The term defensin is still used, even though it is now clear that invertebrate and mammalian defensins do not share a common ancestor[3,4]. Depending on the species, an invertebrate "defensin" may have six or eight cysteines (that form three or four disulfide bonds) and a variety of antimicrobial activities. To complicate the situation, proteins with the same characteristics as defensins are not always called "defensins," such as the recently identified cremycins from *Caenorhabditis remanei*[5]. In addition, invertebrate big defensins are more likely to be evolutionarily related to vertebrate β-defensins than to other invertebrate defensins[6]. Despite this, researchers sometimes rely on the name "defensin" when determining which sequences should be included in analyses.

Structural studies revealed the similarity between insect defensins and scorpion toxins[7], and the CS-αβ fold was subsequently established as the defining structural characteristic of insect defensins[8]. This fold defines the scorpion toxin-like (CS-αβ) superfamily in the Structural Classification of Proteins (SCOP) database[9], which currently includes five families: insect defensins, short-chain scorpion toxins, long-chain scorpion toxins, MGD-1 (from a mollusk), and plant defensins. This superfamily is synonymous with the recently described cis-defensins[4] and Superfamily 3.30.30.10 in the CATH/Gene 3D database[10,11]. Studies from a variety of invertebrate taxa, plants, and fungi show that the names of proteins that contain this fold are not clearly related to cysteine number or bonding pattern, antimicrobial activity, or evolutionary history[12].

The lack of consistency and clear criteria make it challenging to name and classify newly-identified sequences in this superfamily. A major obstacle to comparing proteins in this superfamily is that cysteines are numbered with respect to each individual sequence (the first cysteine in each sequence is C1), with no way to account for the structural role. This means that only sequences with the same number of cysteines can be compared. There is little sequence conservation other than the cysteines forming the CS-αβ fold, which makes alignments and phylogenetic analyses difficult. By developing a numbering system that prioritizes structural features, superfamily sequences can be more easily compared and aligned. Conserved features, as well as those defining subgroups, can be visualized quickly, and new sequences can be more easily placed into the appropriate context.

This paper uses a spreadsheet software (*e.g.,* Excel) to generate a reference numbering system for the CS-αβ superfamily. It shows how this clarifies comparisons between sequences and applies it to new CS-αβ sequences identified from tardigrades. Using the CS-αβ superfamily as an example, the protocol was written to provide guidance when using sequences of interest; however, it is not intended to be specific to this superfamily or to cysteine-rich sequences. This method will likely be most useful for groups of proteins that have been researched independently in divergent taxa and/or have little overall sequence homology, with discrete characteristics that may not be easily recognized by molecular analysis software. This method requires some *a priori* decisions regarding important features, so it will be of limited utility if no important features have been identified. The primary goal is to show how a simple visualization of the sequence relationships can be achieved. This can then be used to inform sequence alignment and analysis, but if alignment and analysis are the primary goals, a barcode method would be a suitable alternative that has more capacity for automation[13]. The current method displays the features of each peptide in a linear form, so it will not be helpful for the direct visualization of 3D structure.

Protocol

1. Determine the Defining Features of the Protein Group of Interest

1. Consult previous publications to determine if there is a consensus regarding the features that are necessary to be considered part of the group. Take note of any inconsistencies or differences in opinion between research groups, and include characteristics that may serve to differentiate one subgroup from another.
2. If previous literature does not address defining characteristics, use sequences that are considered representative of the group as a starting point to identify conserved features.

2. Collect Relevant Sequences

1. If reviews have been written that include analyses of sequences that are representing the group, include these sequences in the raw dataset. Retrieve sequences using accession numbers referenced in the literature and save in a standard sequence editing program (*e.g.,* EditSeq in the Lasergene suite or one of many available for free online).
2. If the group in question has been defined in one of the structural databases, include the sequences the database lists as being part of the group. Retrieve sequences using accession numbers provided in the database and save in a standard sequence editing program, as above.
   NOTE: For example, the sequences categorized in the CS-αβ (scorpion toxin-like) superfamily in the SCOP database can be found here: http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.h.c.h.html.
3. **Perform Basic Local Alignment Search Tool (BLAST)[14] searches of public, online databases available through the National Center for Biotechnology Information (NCBI) to find sequences that may have not been included in the literature or structural databases. For the most complete results, use both the protein BLAST (blastp) and translated blast with protein query (tblastn) programs; these are both available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi.**
   1. Use sequences known to be part of the group of interest as query sequences. Copy and paste the sequence into the search box at the top, or provide a GenBank accession number or gi identifier, if available.
   2. Choose the database from the dropdown menu. Choose non-redundant protein sequences (nr) for blastp and expressed sequence tags for tblastn.
   3. Search for results in specific taxa in the organism setting by typing the organism or taxon name and choosing from the list that appears while typing. To add additional organisms or taxa to exclude, click the "+" button and another field will appear. Exclude any unwanted taxa in the organism box by typing the organism or taxon name, choosing from the list that appears while typing, and checking the "Exclude" box on the right.
   4. Access additional parameters by clicking on "Algorithm parameters" near the bottom of the page. Leave at default unless there is a rationale for changing a parameter.
   5. Click the "BLAST" button to run the analysis; it may take some time for the results to appear. In general, retrieve hits with an expect value (or e-value) of "-05" or better and save in a standard sequence editing program.
      1. If all hits are above this threshold, rerun the search with an increased number of target sequences (in the algorithm parameters section) to obtain all relevant sequences.
4. If necessary, trim the sequences to exclude irrelevant information (*e.g.,* the CS-αβ fold only applies to the mature peptide). Identify signal peptides and pro-peptides for removal using ProP[15] (available online), or SignalP for more sophisticated signal peptide prediction[16] (available online).

3. Generate a Reference in a Spreadsheet Based on the Important Features That Were Identified

1. **Identify the defining characteristics of the group of interest. For example, use the CS-αβ fold definitively established by the solution structure of insect defensin A from *Phormia terraenovae* (Figure 1)[8].**
   1. This fold includes a smaller motif called the cysteine-stabilized helix (CSH)[17]; identify this motif by a CXXXC (where X is any amino acid) upstream of a CXC that form two disulfide bonds (**Figure 1**, solid pink lines).
      NOTE: To complete the CS-αβ motif, a third disulfide bond is formed from additional cysteines placed before each half of the CSH motif (**Figure 1**, dotted pink lines).
2. **Enter these defining features into a spreadsheet. See Figure 2.**
   1. Use columns for the conserved features and to represent the spaces between these features. Keep the columns wide enough to fit numbers and ensure that they have a consistent width. Set the width using the "Format| Column Width" function (**Figure 2**, pink arrow).
   2. Use the rows for the sequence names.

   3. When a sequence has the feature, fill in the box using the fill function (**Figure 2**, pink square). For spacing between features, enter the number of amino acids in the box between and leave it unfilled. For example, using the insect defensin sequence gives a reference that includes six cysteines, with defined spacings between C2 and C3 and between C5 and C6.

3. **Add representative sequences that have been previously established as members of the group based on the structural databases and literature.**
   NOTE: For example, previous literature and the SCOP database identify several groups for inclusion: insect defensins, short-chain scorpion toxins, long-chain scorpion toxins, MGD-1, plant defensins, nematode ABFs, drosomycins from *Drosophila,* and macins. The literature also identifies a bacterial sequence with only four cysteines that might represent the ancestor of this superfamily[18]. Adding these sequences increases the number of cysteines in the reference from six to ten but maintains the alignment of the important structural features (**Figure 3**).
   1. To add a feature that is likely to define a subgroup of sequences (for example, an extra cysteine), use the "Insert" function (**Figure 3**, pink arrow).
   2. If there are features missing from a given sequence, leave the box unfilled and combine it with boxes representing intervening amino acids. If necessary, merge the cells using the merge and center feature (**Figure 3**, pink box).

4. **Continue adding sequences to the groups to gain a better picture of the variation in each group of the larger superfamily. Summarize the group characteristics to facilitate comparisons (Figure 4).**
   1. When the number of amino acids between major features varies, use a hyphen to indicate a range, such as 6 - 12 (6 to 12 amino acids), and a slash to indicate either/or, such as 7/10 (7 or 10 amino acids).
   2. Choose a way to annotate features of sequences that may be relevant but do not occur often enough to include in the reference. For example, since cysteines are important in this superfamily, label additional cysteines (**Figure 4**, pink boxes).

5. Add newly-identified sequences to the spreadsheet using the established sequences as a guide. For example, adding sequences from tardigrades (yellow) shows that the tardigrade sequences fall into several different groups of the superfamily (**Figure 5** shows summaries instead of a row per sequence for space purposes).
6. Show variability within a taxonomic group by rearranging the rows (**Figure 6**).

4. Use the Reference to Refine Amino Acid Alignments

NOTE: There are many programs that can be used for multiple sequence alignments, but this demonstration will use Molecular Evolutionary Genetics Analysis (MEGA6)[19] because it is available to download for free.

1. Download and install the software.
2. Begin a new alignment in MEGA by selecting "Edit/Build Alignment" under the Align tab. Select "Create a new alignment" in the box that appears and click "OK." Then select "Protein."
3. Select "Insert Sequence from File" in the "Edit" menu to import the sequences.
   NOTE: Sequences will need to be in FASTA format for import into MEGA. Background colors that reflect different amino acid types are used by default, but this option can be turned off under the "Display" menu.
4. **Once all sequences are entered, click the flexing arm icon and then "Align Protein" to align the sequences using the MUSCLE algorithm[20].**
   NOTE: ClustalW is also available.
   1. If a message saying that nothing has been selected pops up and asks to select all, click "OK."
   2. NOTE: This opens a window that allows one to change some parameters, but they should only be changed there is reason to do so. This analysis uses a subset of the sequences analyzed in a previous paper[12].

5. Check the alignment based on the important features; note that the top bar above the sequences will show any columns where the amino acid is completely conserved (*). See **Figure 7**. See that the initial alignment shows only three of the four conserved cysteines (**Figure 7**, pink boxes); looking down the column, the AlCRP sequence is clearly misaligned (**Figure 7**, pink arrow).
6. To get rid of the large gap between the I and the conserved C, highlight the dashes and press the "delete" key. Do not highlight any amino acids, or they will be deleted as well.
7. **To move amino acids to the right, highlight and press the space bar.**
   1. Note that the AlCRP now has the structural cysteines aligned and that the last C of the CXXXC motif is conserved throughout the alignment (**Figure 8**). Adjust the alignment as necessary to prioritize the most important features of the sequences.

5. Compare the Groups Identified Using the Reference with Results from Phylogenetic Analyses

1. **From preliminary alignments, determine which sequences should be included in a phylogenetic analysis; for a small number of sequences, this step may be unnecessary.**
   1. Keep an alignment file that includes all sequences, but for a phylogenetic analysis, remove redundant sequences (**Figure 9**, pink boxes show pairs of redundant sequences).
   2. If the data set includes a large number of sequences, run a preliminary analysis and select representatives from groups that always form a clade.

2. **Determine the best amino acid substitution model.**
   1. Export the alignment in MEGA format (under the data tab).
   2. Go to the Models menu and select "Find Best DNA/Protein Model." Choose the file just saved and open it; this will open a window that has some parameters that can be changed.

3. Use the default parameters unless there is a reason to change them. Click "Compute" to begin the analysis.

3. **Run a maximum likelihood (ML) analysis in MEGA.**
    1. Choose "Construct/Test Maximum Likelihood Tree" from the Phylogeny menu.
    2. Choose the model determined to be the best fit for the data from step 5.2 (the output will give the substitution model as well as the best "rates among sites" parameter).
    3. Choose 1,000 bootstrap replicates to obtain the measures of support for the tree.
    4. Click "Compute" to run the analysis; MEGA has a "Tree Explorer" to visualize the tree.

4. **Run a Bayesian analysis in MrBayes open-source software[21].**
    NOTE: A MrBayes manual is also available from this site. This is intended to provide basic steps and is not a comprehensive guide to conducting Bayesian phylogenetic analysis.
    1. Export the MEGA alignment in PAUP (Nexus) format to the same folder as the MrBayes program.
    2. Open MrBayes and type "exe *Filename*" (*e.g.,* "exe Alignment.nex").
    3. Specify the model and analysis parameters. Choose either the model specified in step 5.2 or choose the "mixed" setting that will try various models and report the frequency of the model in the trees with the best posterior probabilities (prset aamodelpr=mixed). Type "showmodel" to report the current model settings and "help mcmc" to show current parameter settings, with a brief explanation of each.
    4. Set the number of generations using the "mcmcp ngen=" command (1 million is typical).
    5. Type "mcmc" to begin the analysis.
    6. When the number of generations has completed, the program will ask to add more generations. If the average standard deviation of split frequencies is less than 0.1, type no. If it is above 0.1, the analysis should be allowed to continue, or some parameters should be changed (see the manual).
    7. Use the "sumt" command to generate the tree files.
    8. After the analysis is complete and a consensus tree is generated, the tree can be viewed in FigTree (available online).

5. Compare the trees to see if the methods generate consistent results.
    NOTE: Some sequences do not provide a lot of information: the trees may not be well resolved and the branches may have minimal support (**Figure 10**).
6. Compare trees to the groups identified using the reference to see if the phylogenetic analyses support these groups.

Representative Results

Groups of sequences in the CS-αβ superfamily reported in the literature are shown in **Figure 4**. The cysteine pairings based on the numbering for each sequence suggest five basic groups (**Table 1**, middle column). Group 1 has six cysteines that from three disulfide bonds and includes sequences from insects, arachnids, mollusks, nematodes, and fungi. Groups 2, 3, and 4 have 8 cysteines that form four disulfide bonds. Group 2 includes insect, arachnid, and plant sequences; group 3 includes arachnid, mollusk, and nematode sequences; and group 4 includes sequences from cnidarians, annelids, mollusks, and fungi. Group 5 includes the 10 cysteine macins. Some sequences did not quite fit these patterns but were generally closer to one group than the others.

Groups 1 and 2 seem to share two bonds: C2-C5 and C3-C6; however, beginning the numbering of each sequence with its first cysteine does not acknowledge the structural context of the bonds. C2-C5 in Group 1 sequences forms one of the two bonds in the CSH motif, while C2-C5 in Group 2 sequences forms the final bond needed to stabilize the CS-αβ fold. The homologous bond to the Group 1 C2-C5 is Group2 C3-C6, which is not obvious from the numbering. It is also not obvious that in Group 3, the C2-C6 bond plays the same structural role.

Using sequences from the literature generated a reference with a total of ten cysteines. The CSH motif is formed from bonds C3-C8 and C4-C9, with C2-C6 completing the CS-αβ fold. Renumbering the cysteine pairs based on the reference numbers clarifies the bonds present in each sequence (**Table 1**, right column). It is now obvious that all the sequences have C2-C6, C3-C8, and C4-C9, reflecting the structural fold that defines the superfamily. The use of a reference allows for easy comparison between sequences that have inconsistent nomenclature and ambiguous classification criteria. It can also help to identify features that define a subgroup of sequences. For example, the C1-C7 bond may differentiate macins from other superfamily members, making it appropriate to classify sequences with this bond as "macins" rather than "defensins" (**Table 1** and **Figure 4**).

Searches of public online databases revealed sixteen sequences from tardigrades that clearly have the CS-αβ fold, eight each from *Hypsibius dujardini* and *Milnesium tardigradum.* Four of the new sequences have six cysteines, nine have eight, one has nine, and two have ten. This gives very little information, but by aligning the sequences to the reference, it becomes clear that tardigrade sequences with the same number of cysteines do not always have the structurally-important cysteines at the same place within the sequence (**Figure 5** and **Figure 6**). The alignment with the reference also allows for the inference of bonding patterns (**Table 2**, inferred bonding patterns shown in parentheses). Some of the tardigrade sequences clearly fit patterns 1 - 4. Others are most similar to the proposed bacterial ancestor, scorpion Cl- toxin, or a family of fungal defensin-like peptides. Pattern 2 may have two subgroups, one represented by scorpion Na+ toxins, drosomycin, and plant defensins, and the other by scorpion Cl- toxins. Further work investigating the function of the tardigrade proteins is needed to determine if some should be considered toxins rather than defensins.

Phylogenetic analyses are often used to study how a group of proteins may have evolved. The sequences in the CS-αβ superfamily are generally short and highly divergent; resulting trees are often poorly resolved and offer little insight. Both the ML and Bayesian trees for the subset of sequences analyzed here were poorly resolved, with low support for many clades (**Figure 10**, **Supplementary Files 1 - 4**). It is common practice to only show bootstrap levels over 70 (or posterior probabilities over 0.7), but **Figure 10** retains all numbers to demonstrate the overall low levels of support. Five groups were supported above 70/0.7 in at least one of the two trees: (a) a 6C and an 8C scorpion toxin; (b) macins; (c) tick and scorpion defensins; (d) plant defensins; and (e) 6C defensins from insects, arachnids, and mollusks. In the ML tree, clade e also includes an 8C toxin and an 8C tardigrade defensin, but support was very low (**Figure 10A**). In general, these reflect the categories identified using the reference cysteine numbering but also show that sequences with different cysteine numbers within a large taxonomical group may be more closely related than sequences with the same pattern from different groups. While only a small number of sequences were used in this study, a larger analysis of 250 sequences did not eliminate the lack of resolution (**Supplementary Files 5 - 8**)[12]. The spreadsheet reference alignment may offer easier visualization of similarities with structural or functional relevance compared to phylogenetic trees.

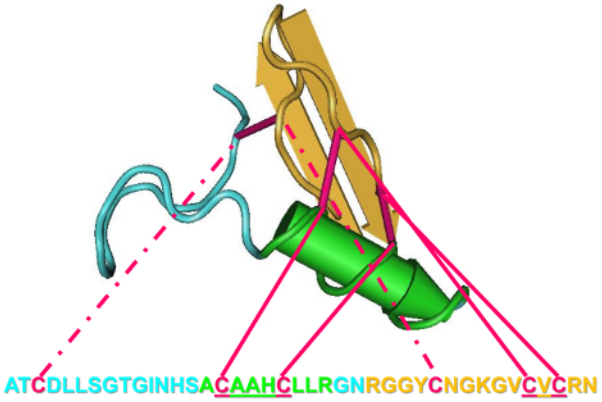
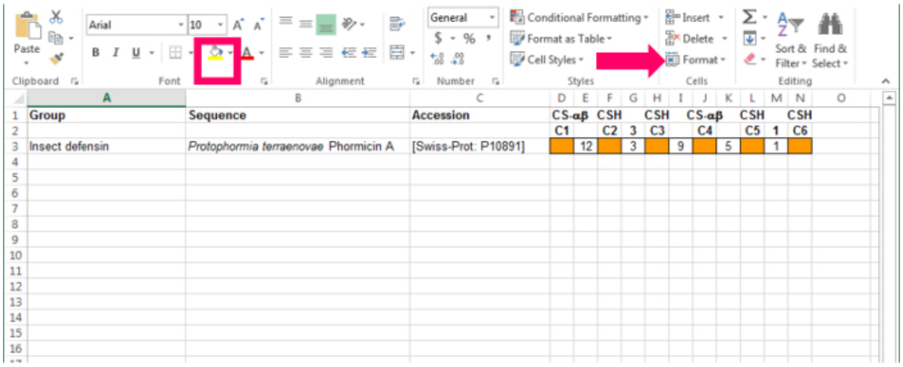
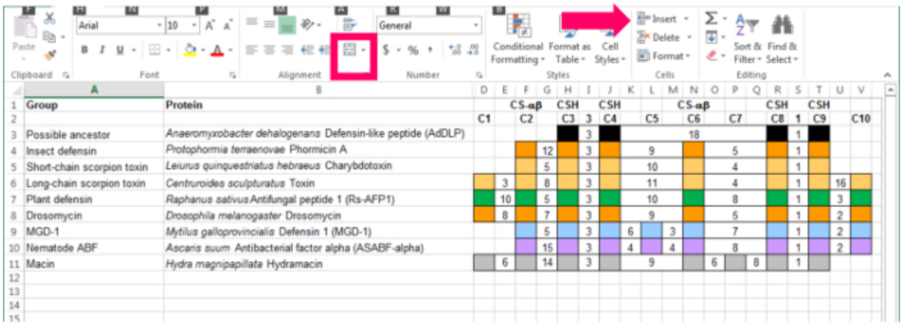


Figure 1: Defining Sequence and Structural Features of the CS-αβ Superfamily. Amino acids and 3D structure are color coded: loop (blue), alpha-helix (green), beta-sheets (gold), and disulfide bonds (pink). Please click here to view a larger version of this figure.



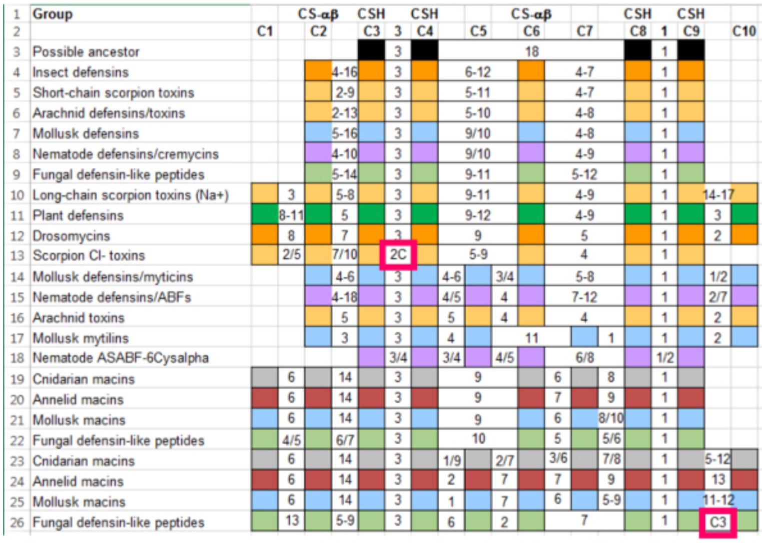
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Group	Sequence	Accession	CS-αβ		CSH		CSH		CS-αβ		CSH		CSH	
2				C1		C2	3	C3		C4		C5	1	C6	
3	Insect defensin	*Protophormia terraenovae* Phormicin A	[Swiss-Prot: P10891]	🟧	12	🟧	3	🟧	9	🟧	5	🟧	1	🟧	
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															

Figure 2: Preliminary Six-cysteine Reference Based on the Sequence of Insect Defensin. Columns indicate the conserved cysteines (C1-C6) and, for the CSH motif, the number of conserved amino acids between the cysteines. The filled boxes indicate that the sequence has the given cysteine and the numbers indicate amino acids between the cysteines. Please click here to view a larger version of this figure.



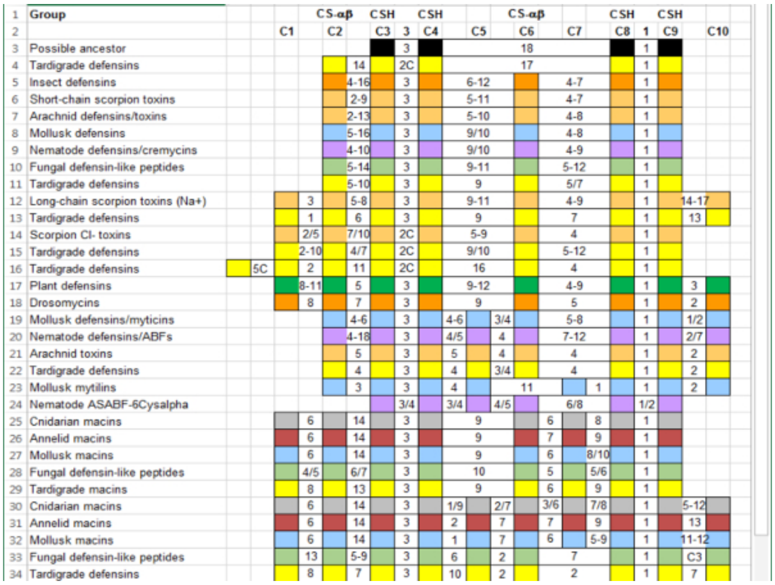
	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	**Group**	**Protein**			CS-αβ		CSH		CSH				CS-αβ				CSH		CSH		
2			C1		C2		C3	3	C4		C5		C6		C7		C8	1	C9		C10
3	Possible ancestor	*Anaeromyxobacter dehalogenans* Defensin-like peptide (AdDLP)					⬛	3	⬛				18				⬛	1	⬛		
4	Insect defensin	*Protophormia terraenovae* Phormicin A			🟧	12	🟧	3	🟧		9		🟧		5		🟧	1	🟧		
5	Short-chain scorpion toxin	*Leiurus quinquestriatus hebraeus* Charybdotoxin			🟨	5	🟨	3	🟨		10		🟨		4		🟨	1	🟨		
6	Long-chain scorpion toxin	*Centruroides sculpturatus* Toxin	🟨	3	🟨	8	🟨	3	🟨		11		🟨		4		🟨	1	🟨	16	🟨
7	Plant defensin	*Raphanus sativus* Antifungal peptide 1 (Rs-AFP1)	🟩	10	🟩	5	🟩	3	🟩		10		🟩		8		🟩	1	🟩	3	🟩
8	Drosomycin	*Drosophila melanogaster* Drosomycin	🟧	8	🟧	7	🟧	3	🟧		9		🟧		5		🟧	1	🟧	2	🟧
9	MGD-1	*Mytilus galloprovincialis* Defensin 1 (MGD-1)			🟦	5	🟦	3	🟦	6	🟦	3	🟦		7		🟦	1	🟦	2	🟦
10	Nematode ABF	*Ascaris suum* Antibacterial factor alpha (ASABF-alpha)			🟪	15	🟪	3	🟪	4	🟪	4	🟪		8		🟪	1	🟪	2	🟪
11	Macin	*Hydra magnipapillata* Hydramacin	⬜	6	⬜	14	⬜	3	⬜		9		⬜	6	⬜	8	⬜	1	⬜		
12																					
13																					
14																					
15																					

Figure 3: Refined Ten-cysteine Reference Based on Representative Sequences from Groups of the CS-αβ Superfamily. The columns indicate conserved cysteines and the amino acids between them. Cysteines contributing to the CSH motif (C3, C4, C8, and C9) and to the CS-αβ fold (C2 and C6) are labeled. Sequences are color-coded by taxonomic group: Arachnida (light orange), Bacteria (black), Cnidaria (grey), Hexapoda (orange), Mollusca (blue), Nematoda (purple), and Plantae (green). Please click here to view a larger version of this figure.



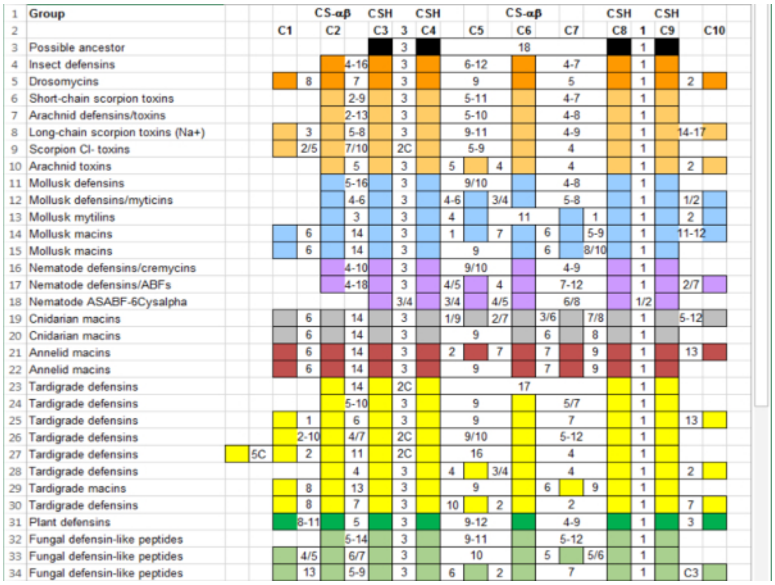
1	**Group**			CS-αβ		CSH		CSH				CS-αβ				CSH		CSH		
2		C1		C2		C3	3	C4		C5		C6		C7		C8	1	C9		C10
3	Possible ancestor					⬛	3	⬛				18				⬛	1	⬛		
4	Insect defensins			🟧	4-16	🟧	3	🟧		6-12		🟧		4-7		🟧	1	🟧		
5	Short-chain scorpion toxins			🟨	2-9	🟨	3	🟨		5-11		🟨		4-7		🟨	1	🟨		
6	Arachnid defensins/toxins			🟨	2-13	🟨	3	🟨		5-10		🟨		4-8		🟨	1	🟨		
7	Mollusk defensins			🟦	5-16	🟦	3	🟦		9/10		🟦		4-8		🟦	1	🟦		
8	Nematode defensins/cremycins			🟪	4-10	🟪	3	🟪		9/10		🟪		4-9		🟪	1	🟪		
9	Fungal defensin-like peptides			🟩	5-14	🟩	3	🟩		9-11		🟩		5-12		🟩	1	🟩		
10	Long-chain scorpion toxins (Na+)	🟨	3	🟨	5-8	🟨	3	🟨		9-11		🟨		4-9		🟨	1	🟨	14-17	🟨
11	Plant defensins	🟩	8-11	🟩	5	🟩	3	🟩		9-12		🟩		4-9		🟩	1	🟩	3	🟩
12	Drosomycins	🟧	8	🟧	7	🟧	3	🟧		9		🟧		5		🟧	1	🟧	2	🟧
13	Scorpion Cl- toxins	🟨	2/5	🟨	7/10	🟨	2C	🟨		5-9		🟨		4		🟨	1	🟨		
14	Mollusk defensins/myticins			🟦	4-6	🟦	3	🟦	4-6	🟦	3/4	🟦		5-8		🟦	1	🟦	1/2	🟦
15	Nematode defensins/ABFs			🟪	4-18	🟪	3	🟪	4/5	🟪	4	🟪		7-12		🟪	1	🟪	2/7	🟪
16	Arachnid toxins			🟨	5	🟨	3	🟨	5	🟨	4	🟨		4		🟨	1	🟨	2	🟨
17	Mollusk mytilins			🟦	3	🟦	3	🟦	4	🟦		11		🟦	1	🟦	1	🟦	2	🟦
18	Nematode ASABF-6Cysalpha					🟪	3/4	🟪	3/4	🟪	4/5	🟪		6/8		🟪	1/2	🟪		
19	Cnidarian macins	⬜	6	⬜	14	⬜	3	⬜		9		⬜	6	⬜	8	⬜	1	⬜		
20	Annelid macins	🟥	6	🟥	14	🟥	3	🟥		9		🟥	7	🟥	9	🟥	1	🟥		
21	Mollusk macins	🟦	6	🟦	14	🟦	3	🟦		9		🟦	6	🟦	8/10	🟦	1	🟦		
22	Fungal defensin-like peptides	🟩	4/5	🟩	6/7	🟩	3	🟩		10		🟩	5	🟩	5/6	🟩	1	🟩		
23	Cnidarian macins	⬜	6	⬜	14	⬜	3	⬜	1/9	⬜	2/7	⬜	3/6	⬜	7/8	⬜	1	⬜	5-12	⬜
24	Annelid macins	🟥	6	🟥	14	🟥	3	🟥	2	🟥	7	🟥	7	🟥	9	🟥	1	🟥	13	🟥
25	Mollusk macins	🟦	6	🟦	14	🟦	3	🟦	1	🟦	7	🟦	6	🟦	5-9	🟦	1	🟦	11-12	🟦
26	Fungal defensin-like peptides	🟩	13	🟩	5-9	🟩	3	🟩	6	🟩	2	🟩		7		🟩	1	🟩	C3	🟩

Figure 4: Summary of CS-αβ Superfamily Sequences Aligned with Reference by Group Characteristics. The columns indicate conserved cysteines and the amino acids between them. Cysteines contributing to the CSH motif (C3, C4, C8, and C9) and to the CS-αβ fold (C2 and C6) are labeled. Sequences are color-coded by taxonomic group: Annelida (dark red), Arachnida (light orange), Bacteria (black), Cnidaria (grey), Fungi (light green), Hexapoda (orange), Mollusca (blue), Nematoda (purple), and Plantae (green). Numbers separated by a dash indicate a range of intervening amino acids; numbers separated by a slash represent either/or. A "C" indicates an additional cysteine that does not occur with enough frequency to warrant addition to the reference. Please click here to view a larger version of this figure.



1	Group					CS-αβ		CSH		CSH				CS-αβ				CSH		CSH		
2				C1		C2		C3	3	C4		C5		C6		C7		C8	1	C9		C10
3	Possible ancestor							⬛	3	⬛				18				⬛	1	⬛		
4	Tardigrade defensins					🟨	14	🟨	2C	🟨				17				🟨	1	🟨		
5	Insect defensins					🟧	4-16	🟧	3	🟧		6-12		🟧		4-7		🟧	1	🟧		
6	Short-chain scorpion toxins					🟧	2-9	🟧	3	🟧		5-11		🟧		4-7		🟧	1	🟧		
7	Arachnid defensins/toxins					🟧	2-13	🟧	3	🟧		5-10		🟧		4-8		🟧	1	🟧		
8	Mollusk defensins					🟦	5-16	🟦	3	🟦		9/10		🟦		4-8		🟦	1	🟦		
9	Nematode defensins/cremycins					🟪	4-10	🟪	3	🟪		9/10		🟪		4-9		🟪	1	🟪		
10	Fungal defensin-like peptides					🟩	5-14	🟩	3	🟩		9-11		🟩		5-12		🟩	1	🟩		
11	Tardigrade defensins					🟨	5-10	🟨	3	🟨		9		🟨		5/7		🟨	1	🟨		
12	Long-chain scorpion toxins (Na+)			🟧	3	🟧	5-8	🟧	3	🟧		9-11		🟧		4-9		🟧	1	🟧	14-17	🟧
13	Tardigrade defensins			🟨	1	🟨	6	🟨	3	🟨		9		🟨		7		🟨	1	🟨	13	🟨
14	Scorpion Cl- toxins			🟧	2/5	🟧	7/10	🟧	2C	🟧		5-9		🟧		4		🟧	1	🟧		
15	Tardigrade defensins			🟨	2-10	🟨	4/7	🟨	2C	🟨		9/10		🟨		5-12		🟨	1	🟨		
16	Tardigrade defensins	🟨	5C	🟨	2	🟨	11	🟨	2C	🟨		16		🟨		4		🟨	1	🟨		
17	Plant defensins			🟩	8-11	🟩	5	🟩	3	🟩		9-12		🟩		4-9		🟩	1	🟩	3	🟩
18	Drosomycins			🟧	8	🟧	7	🟧	3	🟧		9		🟧		5		🟧	1	🟧	2	🟧
19	Mollusk defensins/myticins					🟦	4-6	🟦	3	🟦	4-6	🟦	3/4	🟦		5-8		🟦	1	🟦	1/2	🟦
20	Nematode defensins/ABFs					🟪	4-18	🟪	3	🟪	4/5	🟪	4	🟪		7-12		🟪	1	🟪	2/7	🟪
21	Arachnid toxins					🟧	5	🟧	3	🟧	5	🟧	4	🟧		4		🟧	1	🟧	2	🟧
22	Tardigrade defensins					🟨	4	🟨	3	🟨	4	🟨	3/4	🟨		4		🟨	1	🟨	2	🟨
23	Mollusk mytilins					🟦	3	🟦	3	🟦	4	🟦		11		🟦	1	🟦	1	🟦	2	🟦
24	Nematode ASABF-6Cysalpha							🟪	3/4	🟪	3/4	🟪	4/5	🟪		6/8		🟪	1/2	🟪		
25	Cnidarian macins			⬜	6	⬜	14	⬜	3	⬜		9		⬜	6	⬜	8	⬜	1	⬜		
26	Annelid macins			🟥	6	🟥	14	🟥	3	🟥		9		🟥	7	🟥	9	🟥	1	🟥		
27	Mollusk macins			🟦	6	🟦	14	🟦	3	🟦		9		🟦	6	🟦	8/10	🟦	1	🟦		
28	Fungal defensin-like peptides			🟩	4/5	🟩	6/7	🟩	3	🟩		10		🟩	5	🟩	5/6	🟩	1	🟩		
29	Tardigrade macins			🟨	8	🟨	13	🟨	3	🟨		9		🟨	6	🟨	9	🟨	1	🟨		
30	Cnidarian macins			⬜	6	⬜	14	⬜	3	⬜	1/9	⬜	2/7	⬜	3/6	⬜	7/8	⬜	1	⬜	5-12	⬜
31	Annelid macins			🟥	6	🟥	14	🟥	3	🟥	2	🟥	7	🟥	7	🟥	9	🟥	1	🟥	13	🟥
32	Mollusk macins			🟦	6	🟦	14	🟦	3	🟦	1	🟦	7	🟦	6	🟦	5-9	🟦	1	🟦	11-12	🟦
33	Fungal defensin-like peptides			🟩	13	🟩	5-9	🟩	3	🟩	6	🟩	2	🟩		7		🟩	1	🟩	C3	🟩
34	Tardigrade defensins			🟨	8	🟨	7	🟨	3	🟨	10	🟨	2	🟨		2		🟨	1	🟨	7	🟨

Figure 5: Addition of Tardigrade CS-αβ Sequences to Superfamily Alignment with Reference by Group Characteristics. The columns indicate conserved cysteines and the amino acids between them. Cysteines contributing to the CSH motif (C3, C4, C8, and C9) and to the CS-αβ fold (C2 and C6) are labeled. Sequences are color-coded by taxonomic group: Annelida (dark red), Arachnida (light orange), Bacteria (black), Cnidaria (grey), Fungi (light green), Hexapoda (orange), Mollusca (blue), Nematoda (purple), Plantae (green), and Tardigrada (yellow). Numbers separated by a dash indicate a range of intervening amino acids; numbers separated by a slash represent either/or. A "C" indicates an additional cysteine that does not occur with enough frequency to warrant addition to the reference. Please click here to view a larger version of this figure.



1	Group					CS-αβ		CSH		CSH				CS-αβ				CSH		CSH		
2				C1		C2		C3	3	C4		C5		C6		C7		C8	1	C9		C10
3	Possible ancestor							⬛	3	⬛				18				⬛	1	⬛		
4	Insect defensins					🟧	4-16	🟧	3	🟧		6-12		🟧		4-7		🟧	1	🟧		
5	Drosomycins			🟧	8	🟧	7	🟧	3	🟧		9		🟧		5		🟧	1	🟧	2	🟧
6	Short-chain scorpion toxins					🟧	2-9	🟧	3	🟧		5-11		🟧		4-7		🟧	1	🟧		
7	Arachnid defensins/toxins					🟧	2-13	🟧	3	🟧		5-10		🟧		4-8		🟧	1	🟧		
8	Long-chain scorpion toxins (Na+)			🟧	3	🟧	5-8	🟧	3	🟧		9-11		🟧		4-9		🟧	1	🟧	14-17	🟧
9	Scorpion Cl- toxins			🟧	2/5	🟧	7/10	🟧	2C	🟧		5-9		🟧		4		🟧	1	🟧		
10	Arachnid toxins					🟧	5	🟧	3	🟧	5	🟧	4	🟧		4		🟧	1	🟧	2	🟧
11	Mollusk defensins					🟦	5-16	🟦	3	🟦		9/10		🟦		4-8		🟦	1	🟦		
12	Mollusk defensins/myticins					🟦	4-6	🟦	3	🟦	4-6	🟦	3/4	🟦		5-8		🟦	1	🟦	1/2	🟦
13	Mollusk mytilins					🟦	3	🟦	3	🟦	4	🟦		11		🟦	1	🟦	1	🟦	2	🟦
14	Mollusk macins			🟦	6	🟦	14	🟦	3	🟦	1	🟦	7	🟦	6	🟦	5-9	🟦	1	🟦	11-12	🟦
15	Mollusk macins			🟦	6	🟦	14	🟦	3	🟦		9		🟦	6	🟦	8/10	🟦	1	🟦		
16	Nematode defensins/cremycins					🟪	4-10	🟪	3	🟪		9/10		🟪		4-9		🟪	1	🟪		
17	Nematode defensins/ABFs					🟪	4-18	🟪	3	🟪	4/5	🟪	4	🟪		7-12		🟪	1	🟪	2/7	🟪
18	Nematode ASABF-6Cysalpha							🟪	3/4	🟪	3/4	🟪	4/5	🟪		6/8		🟪	1/2	🟪		
19	Cnidarian macins			⬜	6	⬜	14	⬜	3	⬜	1/9	⬜	2/7	⬜	3/6	⬜	7/8	⬜	1	⬜	5-12	⬜
20	Cnidarian macins			⬜	6	⬜	14	⬜	3	⬜		9		⬜	6	⬜	8	⬜	1	⬜		
21	Annelid macins			🟥	6	🟥	14	🟥	3	🟥	2	🟥	7	🟥	7	🟥	9	🟥	1	🟥	13	🟥
22	Annelid macins			🟥	6	🟥	14	🟥	3	🟥		9		🟥	7	🟥	9	🟥	1	🟥		
23	Tardigrade defensins					🟨	14	🟨	2C	🟨				17				🟨	1	🟨		
24	Tardigrade defensins					🟨	5-10	🟨	3	🟨		9		🟨		5/7		🟨	1	🟨		
25	Tardigrade defensins			🟨	1	🟨	6	🟨	3	🟨		9		🟨		7		🟨	1	🟨	13	🟨
26	Tardigrade defensins			🟨	2-10	🟨	4/7	🟨	2C	🟨		9/10		🟨		5-12		🟨	1	🟨		
27	Tardigrade defensins	🟨	5C	🟨	2	🟨	11	🟨	2C	🟨		16		🟨		4		🟨	1	🟨		
28	Tardigrade defensins					🟨	4	🟨	3	🟨	4	🟨	3/4	🟨		4		🟨	1	🟨	2	🟨
29	Tardigrade macins			🟨	8	🟨	13	🟨	3	🟨		9		🟨	6	🟨	9	🟨	1	🟨		
30	Tardigrade defensins			🟨	8	🟨	7	🟨	3	🟨	10	🟨	2	🟨		2		🟨	1	🟨	7	🟨
31	Plant defensins			🟩	8-11	🟩	5	🟩	3	🟩		9-12		🟩		4-9		🟩	1	🟩	3	🟩
32	Fungal defensin-like peptides					🟩	5-14	🟩	3	🟩		9-11		🟩		5-12		🟩	1	🟩		
33	Fungal defensin-like peptides			🟩	4/5	🟩	6/7	🟩	3	🟩		10		🟩	5	🟩	5/6	🟩	1	🟩		
34	Fungal defensin-like peptides			🟩	13	🟩	5-9	🟩	3	🟩	6	🟩	2	🟩		7		🟩	1	🟩	C3	🟩

Figure 6: Addition of Tardigrade CS-αβ Sequences to Superfamily Alignment with Reference by Taxonomic Group. The columns indicate conserved cysteines and the amino acids between them. Cysteines contributing to the CSH motif (C3, C4, C8, and C9) and to the CS-αβ fold (C2 and C6) are labeled. Sequences are color-coded by taxonomic group: Annelida (dark red), Arachnida (light orange), Bacteria (black), Cnidaria (grey), Fungi (light green), Hexapoda (orange), Mollusca (blue), Nematoda (purple), Plantae (green), and Tardigrada (yellow). Numbers separated by a dash indicate a range of intervening amino acids; numbers separated by a slash represent either/or. A "C" indicates an additional cysteine that does not occur with enough frequency to warrant addition to the reference. Please click here to view a larger version of this figure.

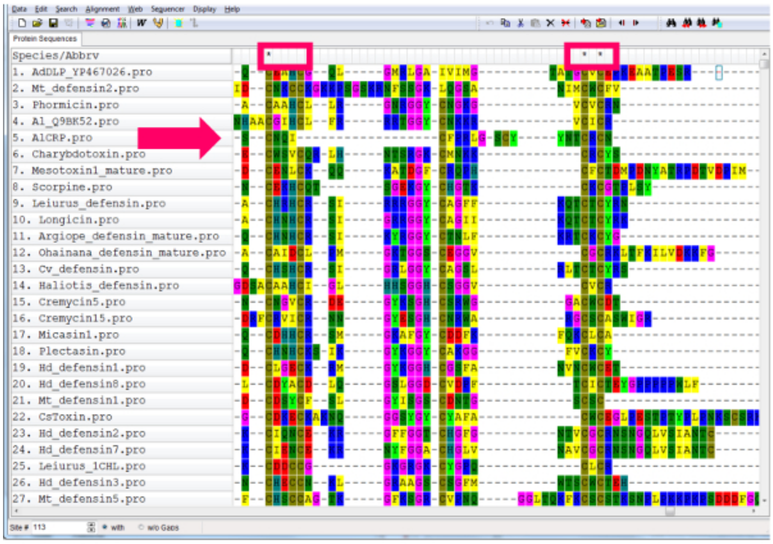


Figure 7: Misaligned Sequence Using Automated Alignment. Amino acids conserved in all sequences are indicated by * in the row above the first sequence (outlined in pink boxes). AlCRP is misaligned. The gap needs to be removed to correctly align the C (pink arrow). Please click here to view a larger version of this figure.

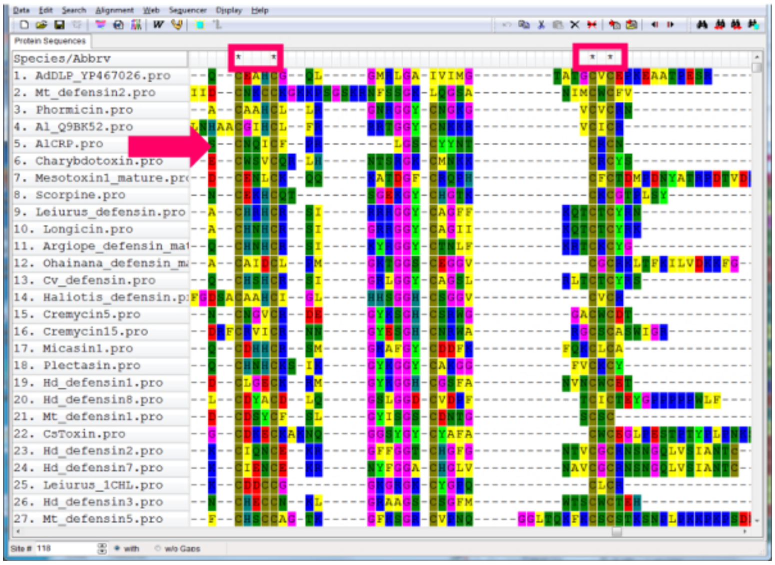
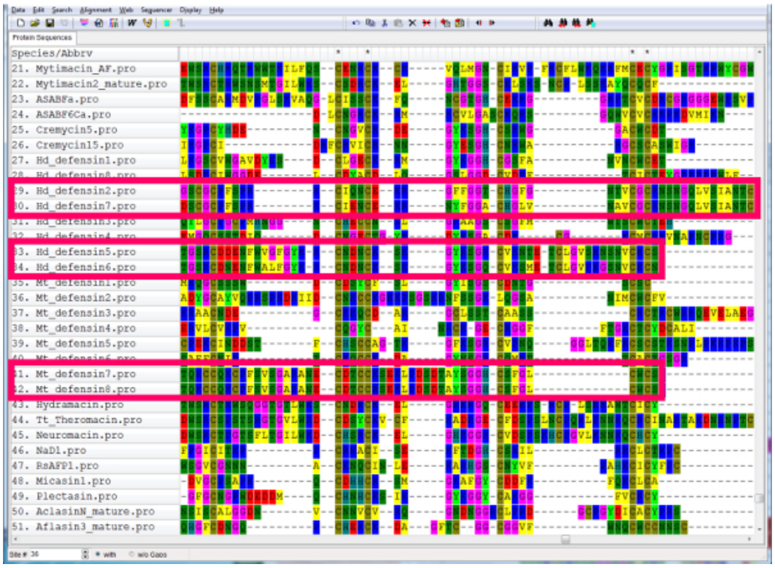
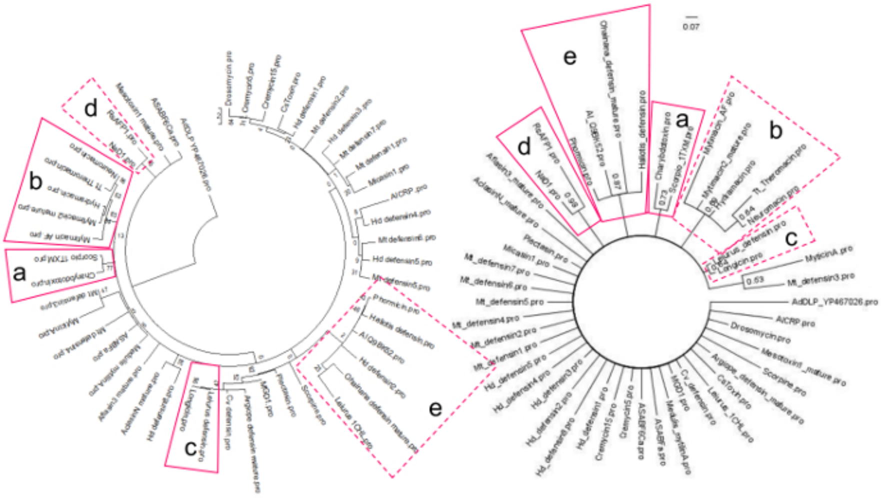


Figure 8: Manual Refinement of the Alignment Preserves the Structurally Important Features of the Sequences. AlCRP is now aligned correctly (pink arrow), and the CXXXC motif is fully conserved for the sequences (pink boxes). Please click here to view a larger version of this figure.



```
Species/Abbrv
21. Mytimacin_AF.pro
22. Mytimacin2_mature.pro
23. ASABFa.pro
24. ASABF6Ca.pro
25. Cremycin5.pro
26. Cremycin15.pro
27. Hd_defensin1.pro
28. Hd_defensin8.pro
29. Hd_defensin2.pro
30. Hd_defensin7.pro
31. Hd_defensin3.pro
32. Hd_defensin4.pro
33. Hd_defensin5.pro
34. Hd_defensin6.pro
35. Mt_defensin1.pro
36. Mt_defensin2.pro
37. Mt_defensin3.pro
38. Mt_defensin4.pro
39. Mt_defensin5.pro
40. Mt_defensin6.pro
41. Mt_defensin7.pro
42. Mt_defensin8.pro
43. Hydramacin.pro
44. Tt_Theromacin.pro
45. Neuromacin.pro
46. NaD1.pro
47. RsAFP1.pro
48. Micasin1.pro
49. Plectasin.pro
50. AclasinN_mature.pro
51. Aflasin3_mature.pro
```

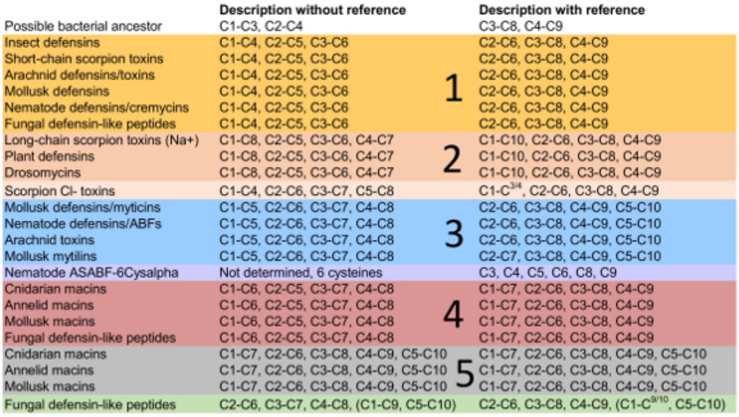
Figure 9: Redundant Sequences in an Alignment. If there are pairs of nearly identical sequences (pink boxes), one can be removed, since these will likely always cluster together in and contribute little to the overall topology of the tree. Please click here to view a larger version of this figure.



(A) Maximum likelihood analysis

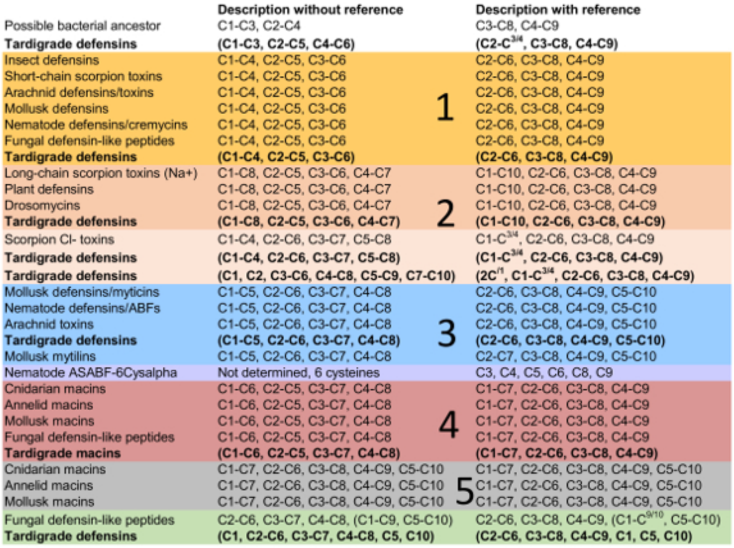
(B) Bayesian analysis

Figure 10: Comparison of Trees Generated from Phylogenetic Analyses. (**A**) Maximum likelihood analysis in MEGA, with 1,000 bootstrap replicates using the WAG+G+I model. (**B**) Bayesian analysis with 1,000,000 generations using the mixed-model setting. Clades supported at 70/0.7 are shown in solid pink lines; dashed pink lines show clades supported at 70/0.7 in the other tree. (a) A 6C and an 8C scorpion toxin; (b) macins; (c) tick and scorpion defensins; (d) plant defensins; and (e) 6C defensins from insects, arachnids, and mollusks. Please click here to view a larger version of this figure.



	Description without reference		Description with reference
Possible bacterial ancestor	C1-C3, C2-C4		C3-C8, C4-C9
Insect defensins	C1-C4, C2-C5, C3-C6	1	C2-C6, C3-C8, C4-C9
Short-chain scorpion toxins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Arachnid defensins/toxins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Mollusk defensins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Nematode defensins/cremycins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Fungal defensin-like peptides	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Long-chain scorpion toxins (Na+)	C1-C8, C2-C5, C3-C6, C4-C7	2	C1-C10, C2-C6, C3-C8, C4-C9
Plant defensins	C1-C8, C2-C5, C3-C6, C4-C7		C1-C10, C2-C6, C3-C8, C4-C9
Drosomycins	C1-C8, C2-C5, C3-C6, C4-C7		C1-C10, C2-C6, C3-C8, C4-C9
Scorpion Cl- toxins	C1-C4, C2-C6, C3-C7, C5-C8		C1-C$^{3/4}$, C2-C6, C3-C8, C4-C9
Mollusk defensins/myticins	C1-C5, C2-C6, C3-C7, C4-C8	3	C2-C6, C3-C8, C4-C9, C5-C10
Nematode defensins/ABFs	C1-C5, C2-C6, C3-C7, C4-C8		C2-C6, C3-C8, C4-C9, C5-C10
Arachnid toxins	C1-C5, C2-C6, C3-C7, C4-C8		C2-C6, C3-C8, C4-C9, C5-C10
Mollusk mytilins	C1-C5, C2-C6, C3-C7, C4-C8		C2-C7, C3-C8, C4-C9, C5-C10
Nematode ASABF-6Cysalpha	Not determined, 6 cysteines		C3, C4, C5, C6, C8, C9
Cnidarian macins	C1-C6, C2-C5, C3-C7, C4-C8	4	C1-C7, C2-C6, C3-C8, C4-C9
Annelid macins	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Mollusk macins	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Fungal defensin-like peptides	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Cnidarian macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10	5	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Annelid macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10		C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Mollusk macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10		C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Fungal defensin-like peptides	C2-C6, C3-C7, C4-C8, (C1-C9, C5-C10)		C2-C6, C3-C8, C4-C9, (C1-C$^{9/10}$, C5-C10)

Table 1: Groups within the CS-αβ Superfamily Based on Cysteine-pairing Patterns. Five basic patterns of bond formation are shown using internal numbers (middle column) or reference numbers (right column). Scorpion Cl- toxins, ASABF 6Cys-alpha, and a group of fungal peptides are placed with the pattern that most closely matches. A cysteine not included in the reference is indicated by a superscript of the cysteines before/after *(e.g.,* C$^{3/4}$ is between C3 and C4).



	Description without reference		Description with reference
Possible bacterial ancestor	C1-C3, C2-C4		C3-C8, C4-C9
Tardigrade defensins	**(C1-C3, C2-C5, C4-C6)**		**(C2-C$^{3/4}$, C3-C8, C4-C9)**
Insect defensins	C1-C4, C2-C5, C3-C6	1	C2-C6, C3-C8, C4-C9
Short-chain scorpion toxins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Arachnid defensins/toxins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Mollusk defensins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Nematode defensins/cremycins	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Fungal defensin-like peptides	C1-C4, C2-C5, C3-C6		C2-C6, C3-C8, C4-C9
Tardigrade defensins	**(C1-C4, C2-C5, C3-C6)**		**(C2-C6, C3-C8, C4-C9)**
Long-chain scorpion toxins (Na+)	C1-C8, C2-C5, C3-C6, C4-C7	2	C1-C10, C2-C6, C3-C8, C4-C9
Plant defensins	C1-C8, C2-C5, C3-C6, C4-C7		C1-C10, C2-C6, C3-C8, C4-C9
Drosomycins	C1-C8, C2-C5, C3-C6, C4-C7		C1-C10, C2-C6, C3-C8, C4-C9
Tardigrade defensins	**(C1-C8, C2-C5, C3-C6, C4-C7)**		**(C1-C10, C2-C6, C3-C8, C4-C9)**
Scorpion Cl- toxins	C1-C4, C2-C6, C3-C7, C5-C8		C1-C$^{3/4}$, C2-C6, C3-C8, C4-C9
Tardigrade defensins	**(C1-C4, C2-C6, C3-C7, C5-C8)**		**(C1-C$^{3/4}$, C2-C6, C3-C8, C4-C9)**
Tardigrade defensins	**(C1, C2, C3-C6, C4-C8, C5-C9, C7-C10)**		**(2C$^{/1}$, C1-C$^{3/4}$, C2-C6, C3-C8, C4-C9)**
Mollusk defensins/myticins	C1-C5, C2-C6, C3-C7, C4-C8	3	C2-C6, C3-C8, C4-C9, C5-C10
Nematode defensins/ABFs	C1-C5, C2-C6, C3-C7, C4-C8		C2-C6, C3-C8, C4-C9, C5-C10
Arachnid toxins	C1-C5, C2-C6, C3-C7, C4-C8		C2-C6, C3-C8, C4-C9, C5-C10
Tardigrade defensins	**(C1-C5, C2-C6, C3-C7, C4-C8)**		**(C2-C6, C3-C8, C4-C9, C5-C10)**
Mollusk mytilins	C1-C5, C2-C6, C3-C7, C4-C8		C2-C7, C3-C8, C4-C9, C5-C10
Nematode ASABF-6Cysalpha	Not determined, 6 cysteines		C3, C4, C5, C6, C8, C9
Cnidarian macins	C1-C6, C2-C5, C3-C7, C4-C8	4	C1-C7, C2-C6, C3-C8, C4-C9
Annelid macins	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Mollusk macins	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Fungal defensin-like peptides	C1-C6, C2-C5, C3-C7, C4-C8		C1-C7, C2-C6, C3-C8, C4-C9
Tardigrade macins	**(C1-C6, C2-C5, C3-C7, C4-C8)**		**(C1-C7, C2-C6, C3-C8, C4-C9)**
Cnidarian macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10	5	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Annelid macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10		C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Mollusk macins	C1-C7, C2-C6, C3-C8, C4-C9, C5-C10		C1-C7, C2-C6, C3-C8, C4-C9, C5-C10
Fungal defensin-like peptides	C2-C6, C3-C7, C4-C8, (C1-C9, C5-C10)		C2-C6, C3-C8, C4-C9, (C1-C$^{9/10}$, C5-C10)
Tardigrade defensins	**(C1, C2-C6, C3-C7, C4-C8, C5, C10)**		**(C2-C6, C3-C8, C4-C9, C1, C5, C10)**

Table 2: Addition of Tardigrade CS-αβ Sequences to Cysteine-pairing Pattern Groups. Tardigrade defensins and macins (bold) are put into the previously established groups where possible. Some tardigrade sequences may show a group-specific pattern. A cysteine not included in the reference is indicated by a superscript of the cysteines before/after *(e.g.,* C$^{3/4}$ is between C3 and C4). The notation "2C$^{/1}$" indicates there are two cysteines upstream of reference C1.

Supplementary File 1 (S1): Alignment of this Dataset in MEGA. Please click here to download this file.

Supplementary File 2 (S2): Maximum-likelihood Tree MEGA File for This Dataset. Please click here to download this file.

Supplementary File 3 (S3): Alignment of This Dataset in Nexus Format for MrBayes. Please click here to download this file.

Supplementary File 4 (S4): Consensus File from the MrBayes Analysis of This Dataset. Please click here to download this file.

Supplementary File 5 (S5): Alignment of 250 CS-αβ Sequences in MEGA. Please click here to download this file.

Supplementary File 6 (S6): Maximum Likelihood tree of 250 CS-αβ Sequences. Please click here to download this file.

Supplementary File 7 (S7): Alignment of 250 CS-αβ Sequences in Nexus Format for MrBayes. Please click here to download this file.

Supplementary File 8 (S8): Consensus File from the MrBayes Analysis of 250 CS-αβ Sequences. Please click here to download this file.

Discussion

The criteria for naming a protein within a group should be clear, but this is not always the case. Sequences that have the CS-αβ fold have been studied in many labs using a variety of organisms, resulting in different systems of nomenclature, as well as varying levels of characterization.

Attempting to impose a completely new nomenclature is not reasonable and would result in a great deal of confusion when consulting previous literature. A reference numbering system can be used in addition to the name of a protein to clarify its characteristics relative to the superfamily.

Groups of proteins with clear criteria for naming and classification will not likely benefit from generating a reference in a spreadsheet, although it may be useful for summarizing large numbers of sequences and visualizing important characteristics. Sequence alignments and logos are useful for investigating the level of conservation at each site, but do not actively prioritize sequence features important for structure or function. The CS-αβ example focused on the structure, but specific amino acids that form a binding site could also be incorporated as a defining feature. As sequence features that confer specific antimicrobial/toxic activities of CS-αβ peptides are identified, these could be added to the reference to clarify groups based on activity. Although only the predicted mature peptides were used in this example, if the presence of a signal peptide or pro-peptide is important, that information can be added for each sequence. Specific insertion or deletion events, as well as intron locations, can also be included if they are thought to be informative. An advantage of using MrBayes for the phylogenetic analysis is that it is not limited to molecular data-it can analyze data coding for other characteristics that may have evolutionary importance. These can be coded as present or absent, providing more information than the sequence alone.

Collecting the relevant sequences is a critical step of the protocol. Depending on the scope of the study and the distribution of the group members, this may span broad taxonomical groups. If the goal is to understand an entire group of proteins, consider that some sequences may be found outside the species that they are usually reported from. If a taxon is already well represented and additional sequences are unlikely or redundant, excluding them from the search may be appropriate. A basic rule-of-thumb for retrieving hits in a BLAST search is to use a cutoff of -05 for the e-value. The e-value is the number of hits expected by chance. While this is suitable for some situations, if there is a group of sequences that is highly divergent but shares specific characteristics, it can be less reliable-it may retrieve sequences that are similar but do not have the specific features wanted, and it may not return sequences that have the key characteristics but that are highly divergent. There are some potential ways of addressing this issue. The first is to look at the sequences identified in the search that are below the -05 cut-off to see if they meet the inclusion criteria. Second, if there is enough information, use Position-Specific Iterated BLAST (PSI-BLAST)[22] or Pattern-Hit Initiated BLAST (PHI-BLAST)[23]. PSI-BLAST uses the results from an initial search to generate a new model for the next round and can sometimes find divergent sequences that the initial search did not identify. PHI-BLAST requires a pattern to be submitted along with the query sequence. This restricts the retrieved sequences to those containing the pattern of interest. This tool is especially useful if a motif unique to the group can be clearly identified.

An accurate alignment is critical for phylogenetic analysis; interpretations of trees are only valid if they are generated using good alignment. Using the reference to inform the alignment can help to avoid errors that are only obvious when the structure or activity are considered. Sequence redundancy will need to be defined for the project. Two sequences that seem redundant may not be for phylogenetic purposes if they are from widely divergent taxa or are nearly identical in sequence but have different structural or functional properties. If there is ambiguity regarding which sequences should be included, multiple alignments can be generated and analyzed separately to see how alignment changes impact phylogenetic inferences. The method presented here does not eliminate the need for the manual adjustment of alignments, but it can help to clarify how the sequences should align and could possibly be used in conjunction with a more sophisticated barcoding technique than has been described previously[13].

For the reference to be useful, it is important to identify defining characteristics that are not currently obvious from the sequence alone. For example, consider the inability to compare cysteine bonding patterns between sequences with different numbers of cysteines when each sequence is numbered with respect to itself. The goal is to facilitate comparison and discussion, not to add another layer of confusion. This may involve several iterations of the reference and judgment calls in deciding which features to include. It is hoped that adopting a common method of discussing divergent sequences in a group will increase the understanding of the group as a whole.

Disclosures

The author has nothing to disclose.

Acknowledgements

Ongoing tardigrade antimicrobial peptide research is supported by intramural funding from the Midwestern University Office of Research and Sponsored Programs (ORSP). The ORSP had no role in study design, data collection, analysis, interpretation, or manuscript preparation.

References

1. Matsuyama, K., Natori, S. Purification of Three Antibacterial Proteins from the Culture Medium of NIH-Sape-4, an Embryonic Cell Line of *Sarcophaga peregrina. J Biol Chem.* **263** (32), 17112-17116 (1988).
2. Lambert, J. *et al.* 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. *PNAS.* **86** (262-266) (1989).
3. Dimarcq, J.-L., Bulet, P., Hetru, C., Hoffmann, J. Cysteine-rich antimicrobial peptides in invertebrates. *Biopolymers.* **47** 465-477 (1998).
4. Shafee, T. M. A., Lay, F. T., Hulett, M. D., Anderson, M. A. The Defensins Consist of Two Independent, Convergent Protein Superfamilies. *Mol Biol Evol.* **33** (9), 2345-2356 (2016).
5. Zhu, S., Gao, B. Nematode-derived drosomycin-type antifungal peptdies provide evidence for plant-to-ecdysozoan horizontal transfer of a disease resistance gene. *Nat Commun.* **5** (2014).
6. Zhu, S., Gao, B. Evolutionary origin of b-defensins. *Dev. Comp. Immunol.* **39**, 79-84 (2013).
7. Bonmatin, J.-M. *et al.* Two-dimensional ^{1}H NMR study of recombinant insect defensin A in water: Resonance assignments, secondary structure and global folding. *J Biomol NMR.* **2** (3), 235-256 (1992).
8. Cornet, B. *et al.* Refined three-dimensional solution structure of insect defensin A. *Structure.* **3** (5), 435-448 (1995).

9. Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. SCOP: a structural classification of proteins database for the investigations of sequences and structures. *J Mol Biol.* **247**, 536-540 (1995).
10. Sillitoe, I. *et al.* CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43** (Database issue), D376-D381 (2015).
11. Lam, S. D. *et al.* Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.* **44** (Database issue), D404-409 (2016).
12. Tarr, D. E. K. Establishing a reference array for the CS-ab superfamily of defensive peptides. *BMC Res Notes.* **9** 490 (2016).
13. Shafee, T. M. A., Robinson, A. J., van der Weerden, N., Anderson, M. A. Structural homology guided alignment of cysteine rich proteins. *SpringerPlus.* **5** (27) (2016).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic Local Alignment Search Tool. *J Mol Biol.* **215** (3), 403-410 (1990).
15. Duckert, P., Brunak, S., Blom, N. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel.* **17** (1), 107-112 (2004).
16. Petersen, T. N., Brunak, S., von Heijne, G., Nielsen, H. SignalP 4.0:discriminating signal peptides from transmembrane regions. *Nat Methods.* **8**, 785-786 (2011).
17. Kobayashi, Y. *et al.* The cysteine-stabilized a-helix: A common structural motif of ion-channel blocking neurotoxic peptides. *Biopolymers.* **31**, 1213-1220 (1991).
18. Gao, B., del Carmen Rodriguez, M., Lanz-Mendoza, H., Zhu, S. AdDLP, a bacterial defensin-like peptide, exhibits anti-*Plasmodium.* activity. *Biochem Biophys Res Commun.* **387**, 393-398 (2009).
19. Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis. *Mol Biol Evol.* **30** (12), 2725-2729 (2013).
20. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32** (5), 1792-1797 (2004).
21. Ronquist, F., Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* **19** (12), 1572-1574 (2003).
22. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** (17), 3389-3402 (1997).
23. Zhang, Z. *et al.* Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26** (17), 3986-3990 (1998).