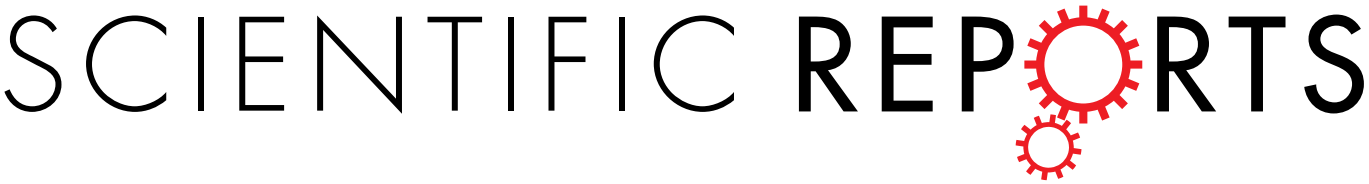


OPEN

Functional classification of protein toxins as a basis for bioinformatic screening

Received: 27 July 2017
Accepted: 2 October 2017
Published online: 24 October 2017

Surendra S. Negi[1], Catherine H. Schein[1,2], Gregory S. Ladics[3], Henry Mirsky[4], Peter Chang[4], Jean-Baptiste Rascle[5], John Kough[6], Lieven Sterck[7], Sabitha Papineni[8], Joseph M. Jez[9], Lucilia Pereira Mouriès[10] & Werner Braun[1]

Proteins are fundamental to life and exhibit a wide diversity of activities, some of which are toxic. Therefore, assessing whether a specific protein is safe for consumption in foods and feeds is critical. Simple BLAST searches may reveal homology to a known toxin, when in fact the protein may pose no real danger. Another challenge to answer this question is the lack of curated databases with a representative set of experimentally validated toxins. Here we have systematically analyzed over 10,000 manually curated toxin sequences using sequence clustering, network analysis, and protein domain classification. We also developed a functional sequence signature method to distinguish toxic from non-toxic proteins. The current database, combined with motif analysis, can be used by researchers and regulators in a hazard screening capacity to assess the potential of a protein to be toxic at early stages of development. Identifying key signatures of toxicity can also aid in redesigning proteins, so as to maintain their desirable functions while reducing the risk of potential health hazards.

Most genetically engineered (GE) food crops involve expressing an introduced protein, thus assessing the safety of the protein is required before commercialization[1–4]. GE crops are created by introducing gene(s) from one species into a crop plant species to improve the nutritional value, yield, drought resistance, herbicide tolerance or pest resistance. Biotechnology companies screen new constructs early in the product development process in order to remove potential hazards and ensure the safety of their product pipelines. National and international regulatory agencies have established guidelines for assessing both trait and GE crop safety through a weight of evidence approach[5–7]. The US-FDA, EPA, USDA, or EFSA and other international organizations require scientifically validated methods to ensure reliable results are generated that allow them to assess the safety of introduced proteins in GE crops.

In silico methods and webservers have been successfully developed to predict toxicity of small molecular weight compounds. These include the systems pharmacology approach[8] to predict drug toxicity and the EPA ToxCast program[9,10] to screen chemicals for potential toxicity to human and the environment. Similar broadly offered bioinformatics tools are not available to predict whether a protein poses the potential to have a toxic effect on mammals. One of the reasons is the absence of a comprehensive, publicly available database containing all proteins with experimentally verified toxic effects in humans or animal studies. Specific databases exist for animal toxins[11,12], spider venoms[13] and microbial pathogens[14]. Although the amino acid sequence determines the three-dimensional structure and the biochemical function of the protein, the specific determinants for the pathogenic effect are not known in many cases. Further, the amino acids that dictate toxic function may be

[1]Sealy Center for Structural Biology and Molecular Biophysics, Department of Biochemistry and Molecular Biology, University of Texas, Medical Branch, Galveston, TX, 77555-0304, USA. [2]Foundation for Applied Molecular Evolution, Inc., Alachua, FL, 32615-9495, USA. [3]DuPont Haskell Laboratory, 1090 Elkton Road, Newark, DE, 19711, USA. [4]Pioneer Hi-Bred, DuPont Agricultural Biotechnology, 200 Powder Mill Road, Wilmington, DE, 19880, USA. [5]Bayer SAS, 355 rue Dostoïevski, CS 90153, Valbonne, 06906, Sophia Antipolis, France. [6]Office of Pesticide Programs, Microbial Pesticides Branch, US Environmental Protection Agency, Washington, DC, USA. [7]Department of Plant Systems Biology, Department of Plant Biotechnology and Bioinformatics, Ghent University, B–9052, Ghent, Belgium. [8]Dow AgroSciences LLC, 9330 Zionsville Road, Indianapolis, IN, 46268, USA. [9]Department of Biology, Washington University in St. Louis, One Brookings Drive, CB 1137, St. Louis, MO, USA. [10]ILSI Health and Environmental Sciences Institute (HESI), 1156 Fifteenth St., NW, Washington, DC, 20005, USA. Correspondence and requests for materials should be addressed to W.B. (email: webraun@utmb.edu)

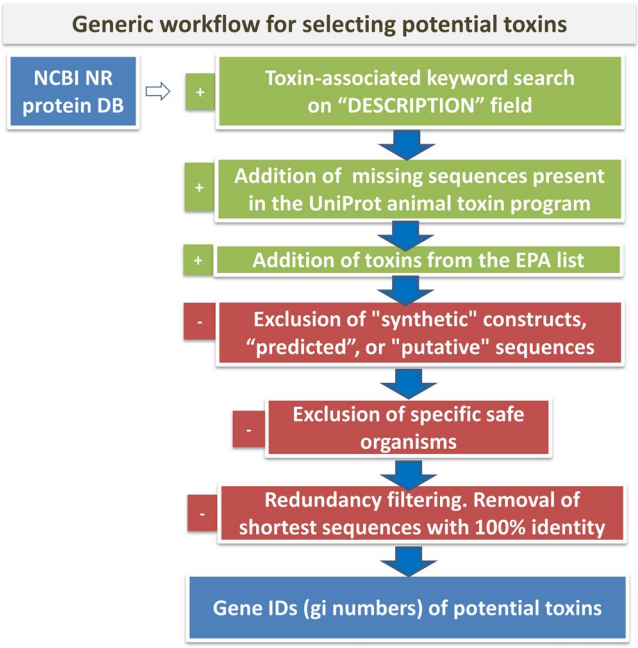


Figure 1. Workflow for selecting potential toxin sequences included in the database. Different selection of keywords were combined to provide a broad coverage of toxins.

quite distant from one another in the linear sequence of the protein, but may be close together within the folded three-dimensional structures of the proteins. These residues may be also distributed on several protein domains, or even on different monomers in multidomain protein toxins.

The extensive data sets of amino acid sequences, three-dimensional structures, biochemical and biological functions of gene products in publicly available databases can be the basis for bioinformatics approaches to determine the potential risk of toxicity. Proper cataloguing of this data, by discriminating the small proportion of proteins that are known toxins, is one part of an overall "weight-of-evidence" evaluation for the safety of GE products[4,5,15,16]. Here we document the first steps to establish a bioinformatics strategy for evaluating the toxic potential of a protein. Beginning with a manually curated list obtained through a keyword search, over 10,000 protein sequences were grouped based on their sequence identity, and then according to their similarity to protein families as classified in the PFAM database[17]. The clustering was automatically performed by a series of independent single linkage clustering with varying thresholds for sequence identities and the top 100 clusters manually inspected for common biochemical and physiological functions. In addition, for all toxin entries protein domains were assigned to PFAM classes. Both procedures indicated that there are only a limited number (<400) of potential mechanisms for protein toxicity. The current list is a starting point for a relational database of protein toxins for hazard screening. We show further that sequence alignments of the clustered toxins can establish structural and sequential motifs[18–20] for use in distinguishing toxins from their non-toxic homologues in the same PFAM class. Extending this classification and motif analysis to all known toxic proteins can aid in identifying possible mechanisms of toxicity during the first tier of hazard screening, and prevent potentially problematic proteins from entering the developmental pipeline.

Results

Selection of a representative set of toxin sequences. We began with a curated list of proteins whose signatures contained one of a selected series of key words that indicated protein toxicity (Fig. 1). The list was simplified by removing duplicates, putative or synthetic constructs, and by adding missing sequences catalogued in existing toxin databases, such as that maintained by the EPA (see materials and methods). The final list of the toxins contained 10,389 protein sequences. Sequence clustering and analysis is described in the Methods section.

About 400 clusters at the 35% sequence identity level contain most protein toxins. To examine the sequence variability of the extracted ~10,000 protein toxins, we generated clusters with cutoff levels at 5% intervals and manually analyzed the most populated individual clusters at the 95%, 65% and 35% sequence identity levels. Clusters at the 95% level were used to identify and remove highly redundant toxin sequences. Sequences with 35% identity can be generally considered to have similar 3D-structures[21–23]. This observation of structural similarity has been confirmed in many cases by the results of Protein Structure Initiative of the NIH[24,25].

However, functional similarity cannot be easily deduced from a simple sequence similarity cutoff. Even proteins that are very similar in both sequences and structures may have completely different functions, whereby only one group is toxic. We therefore used here an empirical approach with a varying sequence identity level to learn about the cluster properties of known protein toxins. The number of clusters decreased from a high of 6,295 at 95% to 3,562 at 65% sequence identity, and to 2,375 when grouped at 35% identity. The 335 most populated

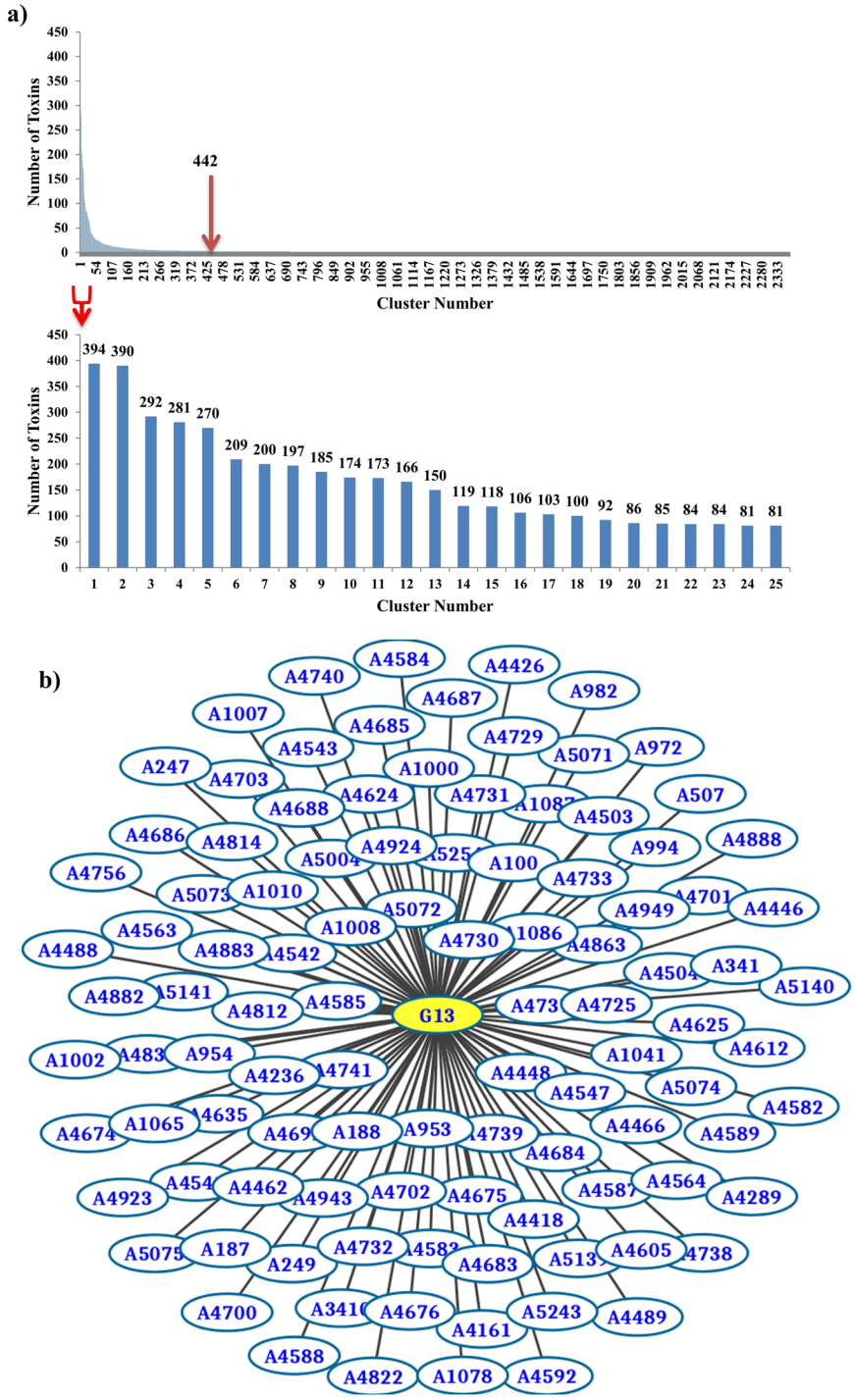


Figure 2. Cluster and network analysis of protein toxin sequences: (**a**) Toxin sequences in each of the 35% sequence identity clusters are shown. Most sequences were contained in about 442 clusters. (**b**) Relation of a cluster at the 95% sequence identity level (indicated by Axxxx) to larger clusters at 35% sequence identity level. The example shows a 35% cluster of conotoxin sequences (G13) composed of multiple 95% clusters.

clusters at the 35% identity level contain about 80% of the protein toxins (Fig. 2a). The rest of the toxins formed very small clusters or were unique sequences at this cut-off level.

Annotations may distinguish members within a cluster that are functionally related. Not surprisingly, the most closely related proteins in the most populated clusters had similar functions, although the annotations in the NCBI entry data files were sometimes quite different. For example, the highly similar proteins in the second highest populated cluster were called Shiga toxin 2 A or verocytotoxin 2 (Table S1), whereby both terms indicate similar activity. Analysis of the clustering at 65% and 35% identity indicated that although the

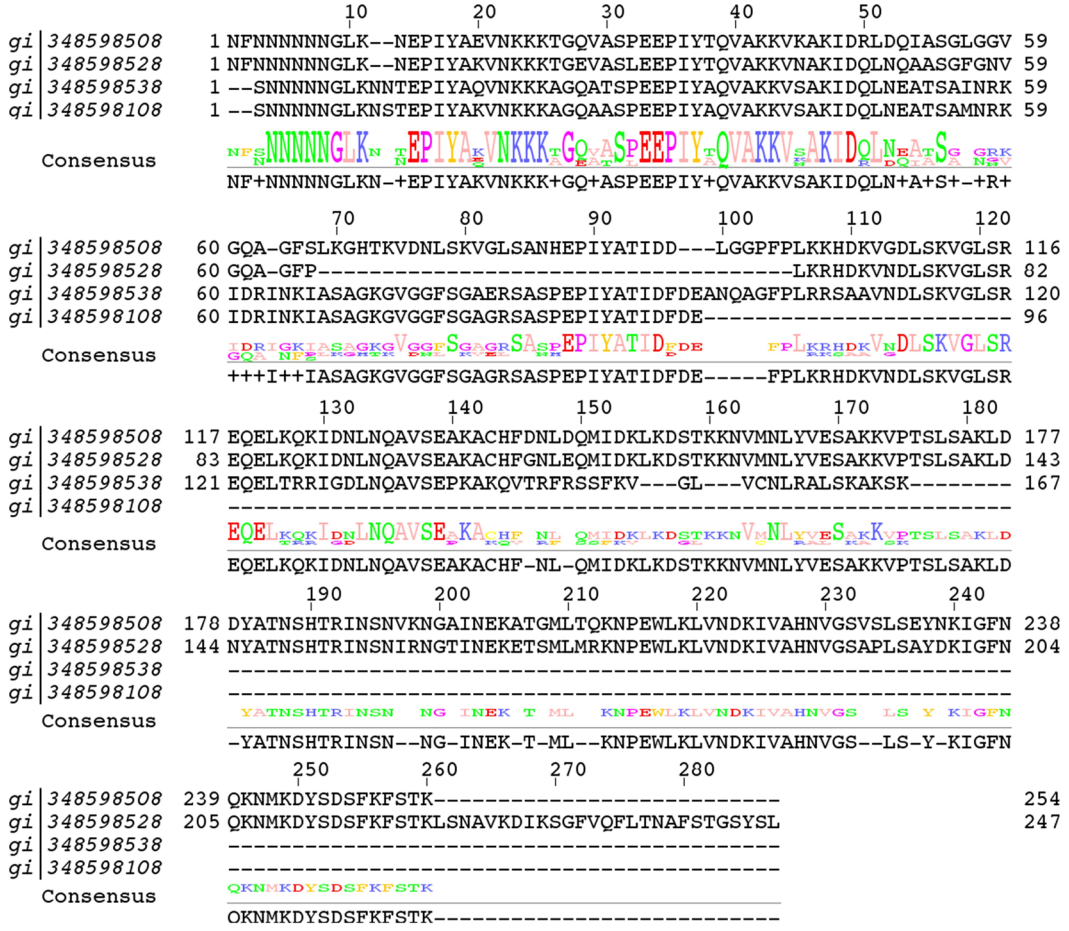


Figure 3. Most of the 1600 singlet sequences can be related to highly populated clusters using multiple sequence alignments. For example, the first sequence is a cytotoxin associated protein from *Heliobacter plyori* included in cluster 4, the other three are from different clusters with only one sequence. Those sequences are almost identical to the first sequence, but contain deletions from 69 to 105, or at the C-termini.

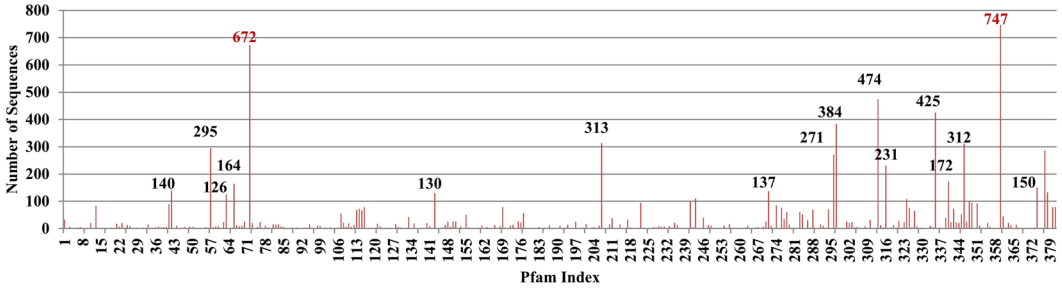
clusters were progressively larger, they still represented relatively homogeneous groups of proteins. For example all of the 394 entries in the top cluster at 35% identity are conotoxins (Table S2).

Functional relations of toxins within clusters. Comparison of the most populated clusters of bacterial toxins at the 35% sequence identity showed that our clustering was consistent with functional annotation. We manually analyzed the bacterial toxins in the top 100 clusters at 65% and 35% identity (Table S3). All toxins within the same 35% cluster were functionally related, despite diverse nomenclature used in the NCBI annotations. Proteins with similar annotations that clustered independently at 35% also often had biologically distinct functions. We thus undertook further analysis with PFAM, as discussed below.

Hierarchical relation between clusters at different sequence identity levels. To illustrate the relationship between the clusters at 95% and those at 65% or 35% sequence identity level, we used a network analysis. Each cluster at different sequence identity threshold was represented by a group ID, and the relationships between these groups were visualized using Cytoscape[26] (Fig. 2b). Comparison of these networks at different sequence identity levels showed that the larger clusters at 35% represent toxins within the same protein family or protein superfamily. We thus suggest that the 35% sequence identity level represents a good choice for a functional grouping of the toxins, and the 95% levels can be used to resolve nomenclature issues.

Unique sequences are short or partial sequences. Manual analysis of the clusters with only one sequence indicated that the overwhelming majority of those singlet sequences were fragments of whole toxins that were contained with high sequence identity in a larger cluster. For example, cluster number 4 at the 35% identity level contains 281 cytotoxin-associated proteins. Three singlet sequences could be aligned with the N-terminal of these sequences, but were too short to make the 35% cutoff for identity with the whole, much longer sequences within the cluster (Fig. 3). Thus, we can assign those entries manually to cluster 4.

The functional diversity of toxins is dramatically restricted. Most bacterial toxins are multidomain proteins, where only one domain may contain the enzymatic region responsible for their detrimental effects.



Pfam ID	No. of Sequences	Toxin type
PF03498	126	Cytolethal distending toxin A/C domain
PF00353	130	Haemolysin-type calcium-binding repeat
PF06414	133	Zeta toxin
PF00161	137	Ribosome inactivating protein
PF05791	140	Bacillus haemolytic enterotoxin (HBL)
PF15738	150	Bacterial toxin of type II toxin-antitoxin system, YafQ
PF16981	164	chi-Conotoxin or t superfamily
PF10530	172	Toxin with inhibitor cystine knot ICK or Knottin scaffold
PF07740	231	Ion channel inhibitory toxin
PF01123	271	Staphylococcal/Streptococcal toxin, OB-fold domain
PF06769	286	YoeB-like toxin of bacterial type II toxin-antitoxin system
PF03507	295	CagA exotoxin
PF07365	312	Alpha conotoxin precursor
PF05016	313	ParE toxin of type II toxin-antitoxin system, parDE
PF02876	384	Staphylococcal/Streptococcal toxin, beta-grasp domain
PF00537	425	Scorpion toxin-like domain
PF00087	474	Snake toxin
PF02950	672	Conotoxin
PF02691	747	Vacuolating cyotoxin

Figure 4. The toxin sequences grouped functionally to 381 PFAM domains. The number of sequences in each PFAM class varied widely (top). The most populated PFAM domains with the number of sequence entries are listed below.

Other toxins form oligomeric assemblies that form membrane pores or need specific protein receptors or lipids to initiate the oligomerization process[27]. Sequence similarity classification alone will thus be misleading for many proteins. Thus the database was further organized according to functional domains, using the PFAM classification[28,29]. Here, toxin identity to discrete areas in the protein known to be responsible for a given function is more important than absolute overall sequence similarity. In addition to expert curation of well-studied proteins, PFAM uses a Hidden Markov Model (HMM) approach, whereby patterns of amino acids in parts of the sequence can reveal similar function. Most of the toxins in the database could be automatically classified by the HMMER software[30] according to the PFAM classification of their domains.

The automated search with HMMER assigned 8570 toxins in the database to only 381 different protein domains out of 16,295 annotated domains in the PFAM database. For the remaining toxins we could manually assign about 500 entries within the same list of protein domains by association, based on their 35 or 65% identity to assigned sequences within the same clusters. This result is also consistent with our cluster analysis at the 35% level, where < 400 clusters accounted for most of the sequences. The functional building blocks of toxins thus come only from a limited subset of domains generally found in proteins. The functional domains found in the protein sequences correspond to less than 3% of the 16,295 annotated domains in the PFAM database. The number of sequences in the database in each PFAM domain is highly variable (Fig. 4). Also consistent with the cluster analysis, the most populated PFAM domains are from vacuolating cytotoxins, conotoxins, snake, scorpion and bacterial toxins. This list of PFAM domains and the associated HMM profiles could be a starting point to find more related toxins in public databases when used with selected keywords. A complete list of all 381 PFAM domains is given as supplementary material (Table S4).

Domain structures of toxins are critical for their function. The PFAM analysis clarifies the functional diversity of the large, multidomain proteins in the selected list of toxins. Many of the most studied bacterial toxins have several domains, which individually may serve regulatory or enzymatic functions not related to the pathogenic or cytotoxic effects of the whole protein[31]. For example, the hemolysin, HlyA of *Vibrio cholera* consists of four distinct domains, which all play a role in pore formation (Fig. 5A). The Pro region (PF12563) is cleaved to activate the toxin[32,33], while the beta trefoil (PF00652), the cytolysin (PF07968) and beta prism lectin jacalin (PF16458) domains form a heptameric pore[34]. The homologous hemolysin toxin in *Vibrio vulnificus* (VVH) does not contain the beta prism lectin domain (PF16458) (Fig. 5B)[35], while the chaperone-like Pro-domain is expressed as a separate gene product. The structural and sequence similarity between these toxins suggests that the VVH also forms a heptameric pore. Thus, even a high sequence similarity of a protein to the Pro or beta prism lectin region alone does not imply that the protein has a potential toxic effect, if the cytolysin and the beta trefoil lectins are absent. This example illustrates how a domain based approach can help in assessing functional similarity of a protein to known toxins.

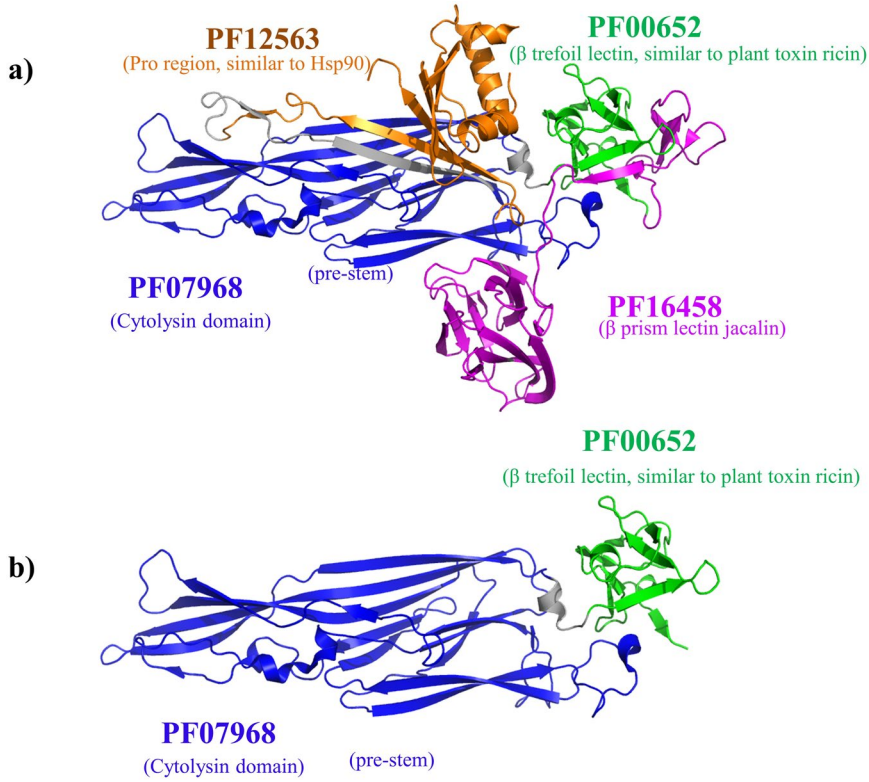


Figure 5. Domain structures of the hemolysins from *Vibrio cholera* (**a**) and *Vibrio vulnificus* (**b**). The membrane-active form of both is a heptameric, pore-forming structure.

Sequence signatures distinguish toxic from non-toxic proteins within a protein family. The diverse Kunitz-type protease inhibitor family includes (non-toxic) aprotinin or bovine pancreatic trypsin inhibitor (BPTI), domains of the Alzheimer's amyloid precursor protein (APP) and tissue factor pathway inhibitor (TFPI). The same PFAM also contains toxic proteins from snake venoms[36], including dendrotoxins from the venoms of mamba (*Dendroaspis*) snakes[37]. These dendrotoxins are highly homologous to BPTI, share the small (~6 kDa) prototype structure[38–40] (Fig. 6), but function by blocking subtypes of voltage-dependent potassium channels of the Kv1 subfamily in neurons[41]. A BLAST sequence search in the NCBI sequence database with BPTI identified several dendrotoxins, with highly significant E-values of 10^{-10} to 10^{-12}. However, dendrotoxins have little or no protease inhibitor activity and BPTI does not block potassium channels[37]. Therefore, E-values obtained from a BLAST search alone are not sufficient to distinguish toxic from non-toxic proteins in this family.

A functional motif analysis with PCPMer[18] successfully distinguished the two protein groups. An alignment of experimentally verified potassium channel blocking dendrotoxins yielded 3 PCP- motifs (1: 25 KYCKLP 30, 2: 41 PSFYYK 46, 3: 55 FDYSGCGGNANRF 67). The three motifs were then searched with PCPMer in eight trypsin inhibitor sequences, including that of BPTI. For motif 1, the average and standard deviations of the score values are 0.89 ± 0.3 for the toxic dendrotoxins *versus* 0.68 ± 0.8 for non-toxic members (P value $1.0*10^{-4}$), and for M2 0.87 ± 0.07 versus 0.56 ± 0.01 (P-value $3*10^{-7}$). The values for motif 3 are not significantly different (P value 0.03) (Table S5). Thus, motifs 1 and 2 are uniquely found in toxic members, whereas motif 3 is found in toxic and non-toxic members. As the mapping on the structure (Fig. 6) shows, motif 1 overlaps with the amino acid residues that have been shown by site-directed mutagenesis to be critical for inhibiting potassium channels[41]. Thus, the motif analysis coincides well with experimental results for areas responsible for the different activities.

Discussion

The objective of our research was to establish bioinformatics tools that can be used in the first tier of assessing the safety of proteins. The study, building on current industry practices to compare protein sequences to internal databases of known protein toxins, provides a validation of these approaches with quantitative data on the distribution of toxins in the protein landscape. We show here that known toxins belong to a restricted number of functional groups, as indicated by both a cluster analysis and specific annotation according to the PFAM classification. In addition, we demonstrate that motif recognition tools can distinguish the toxicity hazard of protein members within the same protein family. A detailed comparison will then allow the reviewer to determine the potential and possible mechanism for protein toxicity based on sequence or domain similarities with known toxic proteins. Such screening may eliminate unnecessary *in vivo* toxicity testing of a protein with valuable traits.

Our work demonstrates that the potential toxin sequences can be clustered into approximately 400 distinct groups, based on either sequence identity alone (Fig. 2a) or sequence features that link them to known functional protein families in PFAM (Fig. 4). As the list of potential toxins was independently established by four different

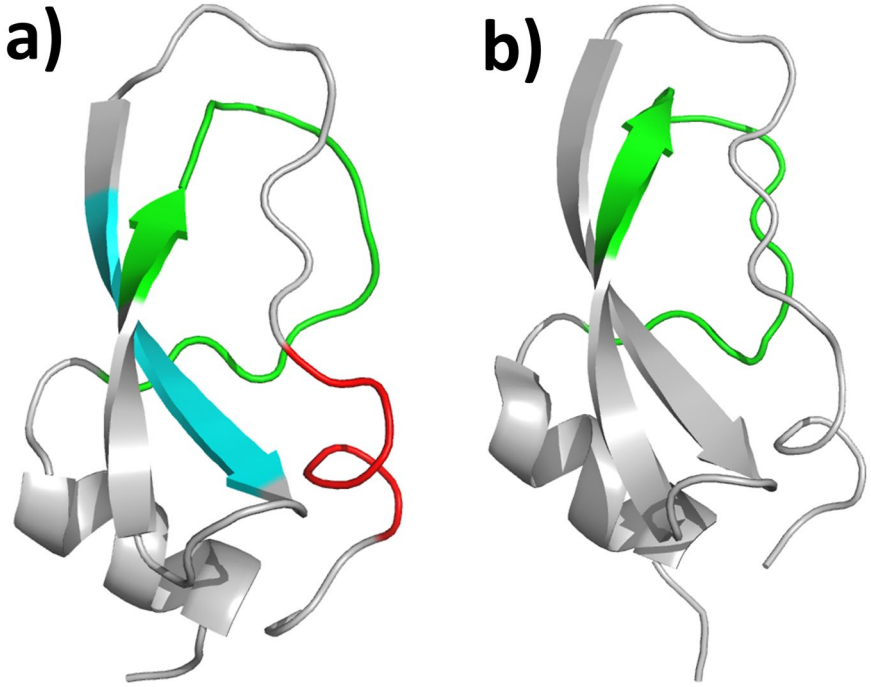


Figure 6. Dendrotoxin (**a**) and BPTI (**b**) group to the same Kunitz inhibitor PFAM domain and share the same 3D fold. Sequence motifs were generated in an alignment of 10 dendrotoxins (motif 1 in red, 2 in cyan and 3 in green). Only motif 3 had a significant score in trypsin inhibitors.

research groups with keyword searches and inclusion of specific known toxins, we believe that the selected list of toxins covers most currently identified toxins. Thus, our finding that the number of protein toxin domains represents only a small fraction of all those known for proteins opens an opportunity to focus the safety assessment on a well-defined list of protein domains and their related functional activities.

As we have shown, sequence identity and even structural similarity cannot be used as sole criteria for defining a protein as a toxin. Some toxins might require specific protein binding sites to guide the toxin to its target, e.g. an ion channel in the case of dendrotoxins. In other cases several proteins act together in a complex. In the anthrax toxin, the two enzymatic proteins, lethal factor (LF), a metalloprotease, and edema factor (EF), an adenylyl cyclase, require the pore forming protective antigen to enter cells and exert a toxic effect[42–45]. For those cases motif analysis tools can be helpful to identify the necessary catalytic activities and protein binding sites for toxicity.

Further computational analysis is required to establish a standardized database containing validated motifs of toxins, *i.e.* proteins that when administered to vertebrates have an adverse effect. One current bottleneck to assembling a database is the non-standard nomenclature used by biologists for proteins that are very similar to one another, as noted above for Shiga toxin 2 A/verocytotoxin. On the other hand, proteins with the same name and similar functions, such as hemolysins, can have quite different domain structures and/or sequences depending on the organism (Fig. 5). Although efforts are underway to standardize the nomenclature of toxins from specific organisms such as spiders[6], scorpions[46], centipedes[47] and snakes[48], a unifying standard nomenclature for all toxins reflecting their structural and functional similarities is currently not available. The sequence clustering we have achieved here will help to clarify such nomenclature issues by assigning most of the toxins to PFAM domains, and hierarchical clustering of the toxin sequences at the three identity levels of 95%, 65% and 35%.

The most highly studied toxins produced by bacteria, as Table S3 indicates, have many different pathogenic mechanisms/modes of action, including ribonucleases (YoeB), vacuolating cytotoxins, hemolysins, cytolysins[49], proteases, phospholipases[50], leukotrienes, neurotoxins or pore formation[51,52]. The snake venom toxins and the conotoxins also present many different activities that can inhibit the growth of plant, insect or mammalian cells or block neural cell receptors. Another example of intrinsic diversity within proteins with similar functions are the vacuolating cytotoxins of *Helicobacter*. These form at least 8 distinct clusters even at 35% identity, emphasizing that the annotated function covers several distinct families with similar annotations. This is also seen for the conotoxins, which despite their short sequences cluster into different functional families.

Our analysis is designed to be a first screening stage, on which to base more detailed computational and experimental investigations. The final use for the protein depends on the risk versus benefits analysis and is outside the scope of this article. For example, even the most virulent proteins can have potentially valuable traits, depending on the administered dose and the route of exposure[53]. Ricin and Botulinum toxins are highly toxic at very low doses[54], yet local injections of the later have proven useful for many therapeutic applications as well as the more widely publicized and profit-generating cosmetic ones. Ricin's effects vary greatly depending on whether it is injected or consumed orally. Inactivated pertussis toxin is both a vaccine and a potential adjuvant[55]. Similarly,

anthrax toxins and derivatives may have use as antitumor agents[56,57]. Rendering a protein toxic may also require posttranslational processing[58], specific cofactors for activity[59], or contact with specific receptors on target cells[60] to exert pathogenic effects.

In conclusion, our current data suggest that there are only a few hundred sequentially and functionally distinct toxin clusters. This implies that most likely, the majority of proteins selected will not share those biochemical functions and can be considered as safe. For those that do bear some similarity to known toxins, we have summarized the basic functions of the largest toxin clusters and present a complete list of all PFAM domains for those toxins. This, coupled with motif recognition tools, provide the first stages of a possible approach to address functional similarities for novel protein products.

Materials and Methods

Selection of toxins. The basis for our work is a collection of potentially toxic proteins that were assembled in internal databases of four biotechnology companies: DuPont Pioneer, Bayer Crop Science, Monsanto and BASF. The sequences were selected using keyword searches (e.g., 'toxic', 'toxin') in the GenBank database, and specific toxins as published in the toxin list 40 CFR 725.421 of the EPA or from the UniProt animal toxin database[11] were added. As keyword searches are not highly specific, proteins from safe organisms, short sequences with 100% sequence identity to longer entries, and those known to be non-toxic were removed (Fig. 1). The specific lists of keyword searches and the filtering processes were done independently by the research teams in the four companies. Finally, the combined database contained 10,389 sequences whose gi entries occurred in at least two of the databases. This selection criterion minimized the number of non-toxic entries in the database and at the same time gave comprehensive coverage of sequences with toxin annotations.

Cluster analysis and functional classification. Cluster analysis of the toxin sequences was done with BLASTCLUST, a standalone software package distributed from NCBI. BLASTCLUST automatically clusters protein sequences based on pairwise alignments generated by the BLAST algorithm using the sequence identity and coverage of the alignment as a criterion to determine if the two sequences are neighbors. Clusters are generated by the single-linkage method, which includes a sequence in a cluster if the sequence is a neighbor to at least one sequence in the cluster. Classification of protein domains occurring in the toxins was based on the PFAM classification (version 29.0, Dec 2015 release). PFAM[17] is a manually curated database of protein domains that contained 16295 entries. The identification of a domain for all toxin sequences in the database was determined by the HMMER[30] software.

Network analysis. Each sequence entry in the database received a unique identifier (cluster ID) for the membership in a cluster of a certain sequence identity level. For example, we denoted the clusters at the 95% level as Axxxx, where xxxx is the rank of the cluster among the 95% level clusters sorted according to the number of members. Thus A5 is the fifth largest cluster among the 95% level clusters. The IDs for the 65% and 35% level clusters were Dxxxx and Gxxxx respectively. A computer program in Perl was written to collect for each member in a cluster of the 35% level the memberships in the 65% and 95% clusters. The result was then represented and analyzed using Cytoscape[26].

Motif analysis. Homologous proteins with similar function usually share similar sequence regions, although the overall sequence identity can be as low as 20–30%. These critical regions, also known as motifs, important for the biological function and similar fold, are in most cases highly conserved[61–64]. In this study, we used PCPMer[18,20,65] to generate motifs of a toxin family and then used these motifs to search for similar sequence regions in other proteins. PCPMer identifies functionally important areas based on conservation of physical-chemical properties (PCPs) of amino acids in a multiple sequence alignment of proteins. The criteria for conservation are derived from the distributions of the PCP descriptors in each column of the alignment as compared to a background distributions, derived from a statistical study of non-redundant proteins from the Swiss-Prot database[66] as a random sample. If the distributions of the five PCP descriptors are significantly different from the background distribution as measured by the relative entropy (or Kullback-Leibler divergence)[67] for any of the five descriptors E1 to E5, that position is considered as conserved. The functional motifs are defined as continuous stretches of conserved residues with relative entropy values higher than an empirical or user specified threshold. The motifs are typically 5–15 amino acids in length, where the minimum length and inclusion of gaps can be specified by the user. The PCPMer approach has been successfully used to characterize functionally important sites in endonucleases, the cytochrome P450 protein, metal-binding proteins, the Ig domains of the muscle protein titin and several allergenic proteins[18,42,68–72].

Data Availability. The complete list of PFAM domains containing toxic proteins (Table S4) can be downloaded from our website http://curie.utmb.edu/SciRep/Negi_et_Table_S4.xlsx. All other data generated or analyzed during this study are included in this published article and its Supplementary Information files.

References

1. Hammond, B., Kough, J., Herouet-Guicheney, C. & Jez, J. M. Toxicological evaluation of proteins introduced into food crops. *Critical reviews in toxicology* **43**(Suppl 2), 25–42 (2013).
2. Joshi, S. S. *et al*. Assessment of potential adjuvanticity of Cry proteins. *Regulatory toxicology and pharmacology: RTP* **79**, 149–155 (2016).
3. Baktavachalam, G. B. *et al*. Transgenic maize event TC1507: Global status of food, feed, and environmental safety. *GM crops & food* **6**, 80–102 (2015).
4. Ladics, G. S. *et al*. Bioinformatics and the allergy assessment of agricultural biotechnology products: industry practices and recommendations. *Regulatory toxicology and pharmacology: RTP* **60**, 46–53 (2011).

5. Ladics, G. S. Current codex guidelines for assessment of potential protein allergenicity. *Food and chemical toxicology: an international journal published for the British Industrial Biological Research Association* **46**(Suppl 10), S20–23 (2008).
6. Delaney, B. *et al*. Evaluation of protein safety in the context of agricultural biotechnology. *Food and chemical toxicology: an international journal published for the British Industrial Biological Research Association* **46**(Suppl 2), S71–97 (2008).
7. Engel, K. H. *et al*. The role of the concept of "history of safe use" in the safety assessment of novel foods and novel food ingredients. Opinion of the Senate Commission on Food Safety (SKLM) of the German Research Foundation (DFG). *Mol Nutr Food Res* **55**, 957–963 (2011).
8. Bai, J. P. & Abernethy, D. R. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annual review of pharmacology and toxicology* **53**, 451–473 (2013).
9. McPartland, J., Dantzker, H. C. & Portier, C. J. Building a robust 21st century chemical testing program at the U.S. Environmental Protection Agency: recommendations for strengthening scientific engagement. *Environmental health perspectives* **123**, 1–5 (2015).
10. Kleinstreuer, N. C. *et al*. Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nature biotechnology* **32**, 583–591 (2014).
11. Jungo, F., Bougueleret, L., Xenarios, I. & Poux, S. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon: official journal of the International Society on Toxinology* **60**, 551–557 (2012).
12. He, Q. Y. *et al*. ATDB: a uni-database platform for animal toxins. *Nucleic acids research* **36**, D293–297 (2008).
13. Herzig, V. *et al*. ArachnoServer 2.0, an updated online resource for spider toxin sequences and structures. *Nucleic acids research* **39**, D653–657 (2011).
14. Zhou, C. E. *et al*. MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic acids research* **35**, D391–394 (2007).
15. Selgrade, M. K., Bowman, C. C., Ladics, G. S., Privalle, L. & Laessig, S. A. Safety assessment of biotechnology products for potential risk of food allergy: implications of new research. *Toxicol Sci* **110**, 31–39 (2009).
16. Koch, M. S., DeSesso, J. M., Williams, A. L., Michalek, S. & Hammond, B. Adaptation of the ToxRTool to Assess the Reliability of Toxicology Studies Conducted with Genetically Modified Crops and Implications for Future Safety Testing. *Critical reviews in food science and nutrition* **56**, 512–526 (2016).
17. Finn, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279–285 (2016).
18. Ivanciuc, O., Garcia, T., Torres, M., Schein, C. H. & Braun, W. Characteristic motifs for families of allergenic proteins. *Mol Immunol* **46**, 559–568 (2009).
19. Ivanciuc, O. *et al*. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem* **11**, 583–593 (2004).
20. Mathura, V. S., Schein, C. H. & Braun, W. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics* **19**, 1381–1390 (2003).
21. Abagyan, R. A. & Batalov, S. Do aligned sequences share the same fold? *Journal of molecular biology* **273**, 355–368 (1997).
22. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823–826 (1986).
23. Sillitoe, I. *et al*. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research* **43**, D376–381 (2015).
24. Serrano, P. *et al*. NMR in structural genomics to increase structural coverage of the protein universe: Delivered by Prof. Kurt Wuthrich on 7 July 2013 at the 38th FEBS Congress in St. Petersburg, Russia. *The FEBS journal* **283**, 3870–3881 (2016).
25. Nair, R. *et al*. Structural genomics is the largest contributor of novel structural leverage. *Journal of structural and functional genomics* **10**, 181–191 (2009).
26. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).
27. Dal Peraro, M. & van der Goot, F. G. Pore-forming toxins: ancient, but never really out of fashion. *Nature reviews. Microbiology* **14**, 77–92 (2016).
28. Merkeev, I. V. & Mironov, A. A. PHOG-BLAST–a new generation tool for fast similarity search of protein families. *BMC Evol Biol* **6**, 51 (2006).
29. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic acids research* **42**, D222–230 (2014).
30. Finn, R. D. *et al*. HMMER web server: 2015 update. *Nucleic acids research* **43**, W30–38 (2015).
31. Lee, M. S., Koo, S., Jeong, D. G. & Tesh, V. L. Shiga Toxins as Multi-Functional Proteins: Induction of Host Cellular Stress Responses, Role in Pathogenesis and Therapeutic Applications. *Toxins (Basel)* **8** (2016).
32. Huntley, J. S., Sathyamoorthy, V., Hall, R. H. & Hall, A. C. Membrane attack induced by HlyA, a pore-forming toxin of Vibrio cholerae. *Hum Exp Toxicol* **16**, 101–105 (1997).
33. Nagamune, K. *et al*. *In vitro* proteolytic processing and activation of the recombinant precursor of El Tor cytolysin/hemolysin (pro-HlyA) of Vibrio cholerae by soluble hemagglutinin/protease of V. cholerae, trypsin, and other proteases. *Infect Immun* **64**, 4655–4658 (1996).
34. Olson, R. & Gouaux, E. Crystal structure of the Vibrio cholerae cytolysin (VCC) pro-toxin and its assembly into a heptameric transmembrane pore. *Journal of molecular biology* **350**, 997–1016 (2005).
35. Kaus, K., Lary, J. W., Cole, J. L. & Olson, R. Glycan specificity of the Vibrio vulnificus hemolysin lectin outlines evolutionary history of membrane targeting by a toxin family. *Journal of molecular biology* **426**, 2800–2812 (2014).
36. Hollecker, M. & Creighton, T. E. Evolutionary conservation and variation of protein folding pathways. Two protease inhibitor homologues from black mamba venom. *Journal of molecular biology* **168**, 409–437 (1983).
37. Harvey, A. L. & Robertson, B. Dendrotoxins: structure-activity relationships and effects on potassium ion channels. *Curr Med Chem* **11**, 3065–3072 (2004).
38. Wlodawer, A., Deisenhofer, J. & Huber, R. Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. *Journal of molecular biology* **193**, 145–156 (1987).
39. Wagner, G. *et al*. Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *Journal of molecular biology* **196**, 611–639 (1987).
40. Berndt, K. D., Guntert, P. & Wuthrich, K. Nuclear magnetic resonance solution structure of dendrotoxin K from the venom of Dendroaspis polylepis polylepis. *Journal of molecular biology* **234**, 735–750 (1993).
41. Imredy, J. P. & MacKinnon, R. Energetic and structural interactions between delta-dendrotoxin and a voltage-gated potassium channel. *Journal of molecular biology* **296**, 1283–1294 (2000).
42. Chen, D. *et al*. Structure-based redesign of an edema toxin inhibitor. *Bioorg Med Chem* **20**, 368–376 (2012).
43. Abrami, L. & van der Goot, R. N. FG. Anthrax toxin: the long and winding road that leads to the kill. *Trends Microbiol. 2005 Feb;13(2):72-8* **13**, 72–78 (2005).
44. Krantz, B. *et al*. A phenylalanine clamp catalyzes protein translocation through the anthrax toxin pore. *Science* **309**, 777–781 (2005).
45. Liu, S., Moayeri, M. & Leppla, S. H. Anthrax lethal and edema toxins in anthrax pathogenesis. *Trends Microbiol* **22**, 317–325 (2014).
46. Tytgat, J. *et al*. A unified nomenclature for short-chain peptides isolated from scorpion venoms: alpha-KTx molecular subfamilies. *Trends in pharmacological sciences* **20**, 444–447 (1999).

47. Undheim, E. A. *et al*. Clawing through evolution: toxin diversification and convergence in the ancient lineage Chilopoda (centipedes). *Molecular biology and evolution* **31**, 2124–2148 (2014).
48. Hargreaves, A. D. & Mulley, J. F. A plea for standardized nomenclature of snake venom toxins. *Toxicon: official journal of the International Society on Toxinology* **90**, 351–353 (2014).
49. Prevost, G., Bouakham, T., Piemont, Y. & Monteil, H. Characterisation of a synergohymenotropic toxin produced by Staphylococcus intermedius. *FEBS Lett* **376**, 135–140 (1995).
50. Titball, R. W. *et al*. Molecular cloning and nucleotide sequence of the alpha-toxin (phospholipase C) of Clostridium perfringens. *Infect Immun* **57**, 367–376 (1989).
51. Pedelacq, J. D. *et al*. The structure of a Staphylococcus aureus leucocidin component (LukF-PV) reveals the fold of the water-soluble species of a family of transmembrane pore-forming toxins. *Structure* **7**, 277–287 (1999).
52. Yamashita, K. *et al*. Crystal structure of the octameric pore of staphylococcal gamma-hemolysin reveals the beta-barrel pore formation mechanism by two components. *Proc Natl Acad Sci U S A* **108**, 17314–17319 (2011).
53. Hong, J. *et al*. Anthrax edema factor potency depends on mode of cell entry. *Biochemical & Biophysical Research Communications* **335**, 850–857 (2005).
54. Hicks, R. P., Nichols, H. M., Bhattacharjee, D. A., van Hamont, A. K. & Skillman, J. E. DR. The medicinal chemistry of botulinum, ricin and anthrax toxins. *Curr Med Chem. 2005* **12**(6), 667–90 (2005).
55. Orr, B. *et al*. Adjuvant effects of adenylate cyclase toxin of Bordetella pertussis after intranasal immunisation of mice. *Vaccine* **25**, 64–71 (2007).
56. Peters, D. E. *et al*. Comparative toxicity and efficacy of engineered anthrax lethal toxin variants with broad anti-tumor activities. *Toxicol Appl Pharmacol* **279**, 220–229 (2014).
57. Phillips, D. D. *et al*. Engineering anthrax toxin variants that exclusively form octamers and their application to targeting tumors. *J Biol Chem* **288**, 9058–9065 (2013).
58. Hormozi, K., Parton, R. & Coote, J. Adjuvant and protective properties of native and recombinant Bordetella pertussis adenylate cyclase toxin preparations in mice. *FEMS Immunology & Medical Microbiology* **23**, 273–282 (1999).
59. Shewell, L. K. *et al*. The cholesterol-dependent cytolysins pneumolysin and streptolysin O require binding to red blood cell glycans for hemolytic activity. *Proc Natl Acad Sci U S A* **111**, E5312–5320 (2014).
60. Spaan, A. N. *et al*. The staphylococcal toxins gamma-haemolysin AB and CB differentially target phagocytes by employing specific chemokine receptors. *Nat Commun* **5**, 5438 (2014).
61. Bork, P. & Koonin, E. V. Protein sequence motifs. *Current opinion in structural biology* **6**, 366–376 (1996).
62. Sigrist, C. J. *et al*. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic acids research* **38**, D161–166 (2010).
63. Bailey, T. L. *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202–208 (2009).
64. Attwood, T.K. *et al*. The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database: the journal of biological databases and curation* **2012**, bas019 (2012).
65. Lu, W., Negi, S. S., Oberhauser, A. F. & Braun, W. Engineering proteins with enhanced mechanical stability by force-specific sequence motifs. *Proteins* **80**, 1308–1315 (2012).
66. Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. Swiss-Prot: juggling between evolution and stability. *Briefings in bioinformatics* **5**, 39–55 (2004).
67. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann Math Stat* **22**, 79–86 (1951).
68. Ivanciuc, O. *et al*. Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-food. *J Agric Food Chem* **51**, 4830–4837 (2003).
69. Ivanciuc, O. *et al*. Structural analysis of linear and conformational epitopes of allergens. *Regulatory toxicology and pharmacology: RTP* **54**, S11–19 (2009).
70. Garcia, T. I., Oberhauser, A. F. & Braun, W. Mechanical stability and differentially conserved physical-chemical properties of titin Ig-domains. *Proteins* **75**, 706–718 (2009).
71. Schein, C. H., Ivanciuc, O. & Braun, W. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. *J Agric Food Chem* **53**, 8752–8759 (2005).
72. Ivanciuc, O., Schein, C. H. & Braun, W. SDAP: database and computational tools for allergenic proteins. *Nucleic acids research* **31**, 359–362 (2003).

Acknowledgements

These studies were conducted as a collaborative effort of the HESI (Health and Environmental Sciences Institute) Protein Allergenicity Technical Committee (PATC). The PATC is a public-private collaboration of scientists with a shared interest in improving the assessment of novel proteins in the context of safety evaluation. HESI is a global non-profit scientific organization that facilitates public-private partnerships to address contemporary issues in human and environmental health and safety. The project was supported by a research grant between UTMB and HESI to W. Braun. Additional grant support by the NIH (AI109090) to W. Braun is gratefully acknowledged. We thank Michael Koch, Andre Silvanovich, Kevin Glenn and Steven Gendel for valuable discussions.

Author Contributions

S.N. performed the bioinformatics analysis of the toxin sequences and prepared the figures and tables, G.L., H.M., P.C., J.R., and S.P. provided gene identifiers of toxins, C.H.S. manually analyzed annotations and homogeneity in toxin clusters, S.N., and W.B. designed the project and wrote the first version of the paper, C.H.S., G.L., J.R., J.K., L.S., S.P., J.J., and L.M. contributed to the discussion and edited the paper. All authors reviewed the final version.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-13957-1.

Competing Interests: The authors employed by the biotechnological companies Dupont, Bayer SAS, and Dow AgroSciences LLC, contributed to this study in the normal course of their employments. The employment affiliation of the authors is shown on the cover page. J.M. Jez and L. Sterck participated in the Protein Toxin Task Force as employees of their academic institutions. J. Kough is employed by the United States Environmental Protection Agency; the contents of this paper reflect the thoughts and opinions of the author and do not represent an official policy statement of the Agency. Any mention of a product does not constitute an endorsement by the United States Federal government.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.