A Graph-Directed Approach for Creation of a Homology Modeling Library: Application to Venom Structure Prediction

Rachael A. Mansbach[1], Srirupa Chakraborty[1,2], Timothy Travers[1,2,†], and S. Gnanakaran[1,*]

[1]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545

[2]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

*Correspondence: gnana@lanl.gov; Tel.: +1 505 665 1923

†Current address: New Mexico Consortium and Pebble Labs Inc., Los Alamos, NM 87544

November 1, 2019

Summary

Many toxins are short, cysteine-rich peptides that are of great interest as novel therapeutic leads and of great concern as lethal biological agents due to their high affinity and specificity for various receptors involved in neuromuscular transmission. To perform initial candidate identification for design of a drug impacting a particular receptor or for threat assessment as a harmful toxin, one requires a set of candidate structures of reasonable accuracy with potential for interaction with the target receptor. In this article, we introduce a graph-based algorithm for identifying good extant template structures from a library of evolutionarily-related cysteine-containing sequences for structural determination of target sequences by homology modeling. We employ this approach to study the conotoxins, a set of toxin peptides produced by the family of aquatic cone snails. Currently, of the approximately six thousand known conotoxin sequences, only about three percent have experimentally characterized three-dimensional structures, leading to a serious bottleneck in identifying potential drug candidates. We demonstrate that the conotoxin template library generated by our approach may be employed to perform homology modeling and greatly increase the number of characterized conotoxin structures. We also show how our approach can guide experimental design by identifying and ranking sequences for structural characterization in a similar manner. Overall, we present and validate an approach for venom structure modeling and employ it to expand the library of extant conotoxin structures by almost 300% through homology modeling employing the template library determined in our approach.

1 Introduction

Toxins have for a long time been considered a rich natural source of therapeutic leads because of their high specificity and binding affinity for various receptors involved in different biological pathways [Zambelli et al., 2016, Verdes et al., 2016]. The drug ziconotide, for example, is a potent analgesic derived from a toxin produced by the aquatic cone snail species *Conus magus* [Miljanich, 2004]. The on-average smaller size of toxins–typically < 100 amino acids along with a sizeable proportion < 30 amino acids long [Dang, 2019]–means they can be employed with relative ease in high-throughput *in silico* screenings to rationally identify candidates for initial scaffolds interacting with a particular receptor of interest. Indeed, this is a fruitful initial line of inquiry for improving drug discovery outcomes and productivity [Romano and Tatonetti, 2019]: in one recent study of note the authors employed a docking approach to identify α-conotoxin BuIA, produced

by species *Conus bullatus*, as a competitive agonist for the lysophosphatidic acid receptor 6, a G-protein coupled receptor involved in the development of several cancers [Younis and Rashid, 2017]. However, such an approach is limited by the necessity of possessing a library of at least moderately-accurate structures of potential toxin candidates [Śledź and Caflisch, 2018]: more structures mean a larger search space and hence a higher likelihood of identifying good initial leads.

Aside from their therapeutic benefit, toxins such as these also pose a threat to biosecurity. The high-throughput evaluation of toxin mode of action as well as the diagnosis and decontamination of disulfide rich toxins, either natural or man-made, is required for public health safety. Rapid advances in synthetic biology have created challenges in determining the health risks posed by natural toxins or modified toxins with even higher pathogenicity [Gomez-Tatay and Hernandez-Andreu, 2019]. Thus, in tandem with high-throughput screening for therapeutic design, it is necessary to also be able to perform high-throughput screening for threat identification and determination of toxin targets and mechanisms of action. Structural characterization stands as a rate-limiting step for high-throughput screening for both therapeutic design and toxin threat characterization, as identified sequences often far outnumber determined structures. For example, only about 3% of sequences isolated from cone snail venom have corresponding experimentally-determined structures [Mansbach et al., 2019].

If the structures of proteins could be rapidly predicted strictly from their sequences, structural determination would not be a bottleneck; however, structure prediction from sequence still remains a challenging proposition [Huang et al., 2016]. Ab initio or de novo modeling approaches for obtaining protein structure predictions by modeling essential folding physics are prohibitively expensive except for small proteins of about 20-62 residues in size [Pitera and Swope, 2003, Ensign et al., 2007, Voelz et al., 2010, Sborgi et al., 2015]. Even for proteins short enough to be de novo modelled in isolation, this can become expensive if a large number of different structures are desired. Structure prediction for a query sequence becomes more tractable when experimentally-resolved structures are available for evolutionarily related sequences: this is referred to as homology modeling [Dill and MacCallum, 2012]. For typical proteins (at least 100 amino acids long), a useful rule of thumb for building a homology model of a protein with unknown structure using a structurally characterized protein as the template is that both proteins should share at least 25% sequence identity [Baker and Sali, 2001, Xiang, 2006]. Of note, several protein structure prediction algorithms use a combination of ab initio and homology modeling approaches. For instance, both I-TASSER [Yang et al., 2015, Zhang, 2008] and ROSETTA [Bradley et al., 2005, Simons et al., 1997] split up the query sequence into fragments that will be searched against a library to identify structure fragments (the homology modeling part), followed by assembly of these fragments into a full-length structural model via molecular dynamics or Monte Carlo simulations using either physics-based or knowledge-based force fields (the ab initio modeling part).

One challenge in applying homology modeling to toxins is that, due to their on-average smaller size, the so-called modeling "safe zone" where structural similarity can be inferred requires a higher sequence identity during structural template selection than the 25% rule of thumb for typical proteins [Krieger et al., 2003, Kong et al., 2004]. This is because at shorter peptide lengths, a sequence identity of 25% is more likely to have arisen by random chance and not due to any evolutionary constraints on the structure. A related challenge occurs in constructing suitable template libraries that contain sufficient information for building accurate homology models for these short peptides. To apply the homology modeling framework for shorter peptides, a reasonable heuristic instead becomes that the alignment length and percent identity fall above the phenomenological curve introduced by Rost [Rost, 1999] (see Eqn. 1 and Fig. S1). The relative steepness of the Rost curve for alignment lengths of less than fifty amino acids provides an illustration of why, for peptides of such lengths, it is important to use the actual functional form, rather than a static cutoff, to assess whether a pairwise alignment contains sufficient information for homology modeling.

In this article, we propose the use of a simple graph-based algorithm for homology modeling of toxins. Graph theory has a long and storied history of usage for sequence-grouping tasks such as homology detection [Santiago et al., 2018], structure prediction [Bolten et al., 2001, Pipenbacher et al., 2002, Yan et al., 2011], protein family identification [Abascal and Valencia, 2002, Enright and Ouzounis, 2000], and even direct homology modeling [Yan et al., 2011]. For large heterogeneous databases, it can be challenging to identify homologues and a number of sophisticated algorithms have been developed for such purposes; we instead focus on the problem of homology modeling a set of cysteine-rich toxins known to be evolutionarily related. In our approach, we employ the number and placement of cysteines within a sequence as a rough initial

estimate of functional and structural relatedness. We apply our approach to the so-called conotoxins, which are small, cysteine-rich peptide toxins produced by the cone snails [Uribe et al., 2018, Mansbach et al., 2019].

In the following sections, we present our graph-based approach and employ it to construct sequence graphs and identify good libraries of templates for homology modeling. We demonstrate that these libraries improve outcomes for structure homology modeling over using the typical 25% flat cutoff and employ them as part of a homology modeling procedure that results in a significantly expanded library of structures for the conotoxins that will be of use in future high-throughput studies. In addition, we use the graph-based approach to construct a set of tables indicating sequences whose experimental or ab initio structural characterization is predicted to be most valuable in creating a broad structure library by using homology modeling.

2 Results

We initialize the algorithm by separating a set of over 2000 known conotoxin sequences into databases containing four, six, eight, and ten cysteines respectively. For each database, we construct graphs of sequences in which an edge between two nodes (i.e., sequences) represents a pairwise alignment that is of sufficient length and percent identity to fall into the safe homology modeling zone above the Rost curve (cf. Eqn. 1 and Fig. S1). Some portion of the sequences have known structures, such that the corresponding nodes are annotated with the relevant PDB ID(s). We employ the graphs thus generated to iteratively add nodes with structures to a library of templates for homology modeling (see Fig. 1 for a schematic illustration of the procedure). We term this set of sequences $\{\mathcal{L}_{\text{ex}}\}$, the set of existing structural library templates (cf. Fig. 2). Nodes are added to $\{\mathcal{L}_{\text{ex}}\}$ in a greedy manner, in order of highest node degree, such that the resulting library will contain enough templates to homology model as many non-structurally-characterized sequences as possible but with small sequence overlap and retaining a number of non-library structures for quality assessment. Since this is approximately the vertex-covering problem of a graph, we cannot find a globally optimal solution, as that problem is NP-complete [Karp, 1972]. We halt the procedure once either we have no further nodes with structures to add or there are no remaining sequences in a given connected component of the graph that are not connected to at least one library template sequence, such that all sequences in that component may be structurally characterized by homology modeling. We refer to the set of sequences that may be homology modeled based on set $\{\mathcal{L}_{\text{ex}}\}$ as set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$ that are covered by $\{\mathcal{L}_{\text{ex}}\}$. We also perform a similar procedure–but without the constraint of structure annotation–on the nodes absent $\{\mathcal{L}_{\text{ex}}\}$ to identify the set $\{\mathcal{L}_{\text{proj}}\}$ that are of interest for experimental or ab initio structural characterization such that they cover the remaining set $\{\mathcal{C}(\mathcal{L}_{\text{proj}}\}$.

In Fig. 2, we present the sequence graphs for sets of conotoxin sequences with four, six, eight, and ten cysteines respectively. We specifically display the set $\{\mathcal{L}_{\text{ex}}\}$ (in orange), which we employ to predict structures for the set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$ (in blue) by homology modeling. We show $\{\mathcal{L}_{\text{proj}}\}$ (in green) whose structural characterization from either experiment or ab initio modeling would lead to coverage by homology modeling of the set $\{\mathcal{C}(\mathcal{L}_{\text{proj}})\}$ (in magenta) that comprises sequences with no characterized structure and not covered by set $\{\mathcal{L}_{\text{ex}}\}$.

These figures demonstrate that we are able to characterize a large number (and moderate proportion) of unknown conotoxin structures, which may be used for high throughput screening. Specifically, out of the 801 sequences with four cysteines, 61 (7.6% of total) currently have experimentally-resolved structures. The graph-based approach selected 49 (6.1% of total) of these structures as comprising the four cysteine template library (set $\{\mathcal{L}_{\text{ex}}\}$; orange circles in Fig. 2a, while the unselected structures are represented in black), which allowed for homology modeling of a further 143 (17.9% of total) sequences (set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$; blue circles in Fig. 2a). This corresponds to an increase of over 230% for the number of structurally characterized sequences over the original 61. In addition, the graph-based approach indicated a further 453 sequences (56.6% of total) would need to be characterized, experimentally or ab initio, to allow for homology modeling of the remaining 151 (18.9% of total). Out of the 1,113 sequences with six cysteines, 44 (4.0% of total) currently have experimentally-resolved structures. The graph-based approach selected 30 (2.7% of total) of these structures as comprising the six cysteine template library, which allowed for homology modeling of a further 148 (13.3% of total) sequences. This corresponds to an increase of over 330% for the number of structurally characterized sequences over the original 44. In addition, the graph-based approach indicated a further 419 sequences (37.6% of total) would need to be

characterized, experimentally or ab initio, to allow for homology modeling of the remaining 509 (45.7% of total). Out of the 190 sequences with eight cysteines, 2 (1.1% of total) currently have experimentally-resolved structures and were selected as comprising the entire template library, which allowed for homology modeling of a further 17 (8.9% of total) sequences. This corresponds to an increase of 850% for the number of structurally characterized sequences over the original 2. In addition, the graph-based approach indicated a further 71 (37.4% of total) would need to be characterized, experimentally or ab initio, to allow for homology modeling of the remaining 101 (53.1% of total). There are no known structures corresponding to ten cysteine sequences so there is no current coverage. The graph-based approach indicated that 19 of the total 53 sequences (35.8%) would have to be characterized to allow for homology modeling of the remaining 34 (64.2%). In Tables 1, 2, and 3, we list all PDB IDS of the structures included in the template set $\{\mathcal{L}_{\text{ex}}\}$ constructed for the four and six cysteine sequences, along with their sequences and the name or names of the corresponding toxins.

In Fig. S2, we present the same sequence graphs used to construct the template libraries, but we color the nodes by relative sequence length instead of set occupation. A significant proportion of isolated sequences (nodes with no connections that therefore cannot be homology modeled) are relatively short (cf. the ring of small red nodes in Fig. S2A and to a lesser extent in Fig. S2B), which demonstrates that a high proportion of isolated nodes may be characterized well through rapid ab initio modeling, particularly for the four and six cysteine sequences. Specifically, 372 of the 453 four cysteine $\{\mathcal{L}_{\text{proj}}\}$ sequences (82.1%) are isolated nodes with no edges; of these 298 (80.1%) are shorter than 20 amino acids, and 353 (94.9%) are shorter than 30 amino acids in length. In addition, 239 of the 419 six cysteine $\{\mathcal{L}_{\text{proj}}\}$ sequences (57.0%) are isolated nodes; of these 86 (36.0%) are shorter than 20 amino acids, and 163 (68.2%) are shorter than 30 amino acids in length. Conversely, 41 of the 71 eight cysteine $\{\mathcal{L}_{\text{proj}}\}$ sequences (57.7%) are isolated nodes, and of these only 5 (12.2%) are shorter than 30 amino acids in length; 10 of the 19 ten cysteine $\{\mathcal{L}_{\text{proj}}\}$ sequences (52.6%) are isolated nodes and of these none are shorter than 30 amino acids in length.

In Fig. 3, we assess the quality of the template libraries for homology modeling, constructed using the graph-based approach employing the Rost cutoff, and compared with a set of template libraries based on a static 25% rule-of-thumb cutoff. In Fig. 3A-B, we constructed homology models for each structure in a library using the other structures in that library and computed the root-mean-square deviation (RMSD) between each modeled structure with the corresponding experimental structure. In Fig. 3C-D, a similar assessment was performed for all structures that were not included in each template library. As expected, there is a statistically significant improvement (downwards shift in the distribution, two-tailed Kolmogorov-Smirnov test with $p < 0.05$) for both in- and out-of-library structures when using the Rost cutoff as compared to the 25% cutoff, which verifies the necessity of our approach. For in-library assessment, the mean of the distribution drops from 4.0 ± 0.7 Å to 1.5 ± 0.2 Å for the four cysteine library and from 3.8 ± 0.6 Å to 2.1 ± 0.2 Å for the six cysteine library, while for out-of-library assessment, the mean of the distribution drops from 1.7 ± 0.1 Å to 1.0 ± 0.2 Å for the four cysteine library and from 1.82 ± 0.09 Å to 1.4 ± 0.1 Å for the six cysteine library.

To illustrate the reason that a static cutoff is less accurate, in Fig. S3, we display an approximation of the distribution of minimum percent identity needed to construct a reliable homology model for a given conotoxin sequence. The minimum required percent identity varies greatly for different conotoxins: although almost none of the conotoxins are long enough to employ the typical 25% cutoff, the relatively large width of the distributions even among conotoxins with the same number of cysteines indicates that choosing a static cutoff is not appropriate. A static cutoff could impact modeling accuracy by either underestimating the needed percent identity for a short sequence or by overestimating the needed percent identity for a long one and thus removing from consideration templates that would otherwise be appropriate, although in the case of a set of short sequences like the conotoxins the primary source of loss of accuracy is expected to be the former.

In Tables 4 and 5 and Tables S1 and S2, we present and rank the set of conotoxins that are of greatest interest for experimental characterization, as availability of experimental structures for these sequences (belonging to set $\{\mathcal{L}_{\text{proj}}\}$) would allow homology modeling of the remainder of the (nonisolated) sequences (belonging to set $\{\mathcal{C}(\mathcal{L}_{\text{proj}}\})$). We present these in order of greatest graph degree, since greater degree in the graph corresponds to the ability to cover a greater number of sequences. Thus, we suggest that experimental structural resolution begin with those sequences listed at the top of their respective tables and work downwards in order to most rapidly and

efficiently structurally characterize the sequence space of the conotoxins.

Finally, in Supporting File `finalmodels.zip`, we attach the set of structures computed by homology modeling, corresponding to sequences in the set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$, with the four, six, and eight cysteine library structures used as templates. Because we divided the sequences into subsets based on the number of cysteines in a sequence, we are able to use keeping the cysteines aligned as an additional criterion during the homology modeling procedure. The average PROCHECK G-factor, which is a log-odds score based on the likelihood of observing the given distributions of ϕ-ψ and χ_1-χ_2 angles in proteins, is 0.086 ± 0.005 for the reported four cysteine models, -0.103 ± 0.007 for the reported six cysteine models, and -0.2 ± 0.1 for the report eight cysteine models. Since this score is not a relative measure and values above -0.5 are generally considered acceptable, this provides evidence that the structures we have computed are physically reasonable. We further assess the quality of the homology modeling protocol by using it to model each structure in the library with templates selected from other structures in that library. The distribution of root-mean-square deviation (RMSD) values of the top three models compared with each experimental structure is shown in Fig. 4A-B. We see that our method performs well: the average RMSD in the four cysteine architecture is 2.00 ± 0.09 Å with at least 80% of the models having less than 3 Å RMSD, and the average RMSD in the six cysteine architecture is 2.3 ± 0.2 Å with 75% of the models having less than 3 Å RMSD. Most of the higher RMSD values are contributed by the flexible loops and coils. When we look at the RMSD distribution after rejecting those atoms that cannot be structurally aligned, as in case of loops and coils, the distributions improve significantly (Fig. S4), with a mean of 1.55 ± 0.09 Å for the four cysteine architecture and a mean of 1.2 ± 0.1 Å for the six cysteine architecture, with 100% of the models for both archtiectures having less than 3.5 Å deviations. A second test for validating our method was performed by checking the distribution of native contacts in the modeled structures (Fig. 4C-D). At least 60% of the native structures were captured in our models, with the distribution means of $80 \pm 1\%$ and $81 \pm 1\%$ for the four and six cysteine architectures respectively. Two pairs of residues were defined to have a native contact if the distance between the Cα atoms in the native experimental structure was less than 8 Å, and the pair was at least 4 residues apart (Cα^{i}– Cα^{i+4}).

3 Discussion

By employing a straightforward graph-based heuristic approach, we have constructed a set of template libraries for homology modeling of conotoxins based on the number of cysteines contained in the sequence that may also be used for homology modeling of other short, disulfide-rich, evolutionarily-related peptides. We demonstrated that libraries constructed to account for the shorter lengths of the conotoxins produce homology models that are more accurate than libraries constructed with the typical static 25% cutoff for most proteins. Currently, sufficient information is not available to homology model any sequences containing more than eight cysteines, as experimental characterization has focused preferentially on the shorter conotoxins.

Next, we employed our libraries to predict a set of structures from sequence using homology modeling, allowing us to expand the library of known conotoxin structures by about 290% overall, although a number of sequences remain without any associated structural predictions. We assessed the quality of these structures through standard techniques to demonstrate they are expected to be reasonably accurate and therefore may be employed for high-throughput screening of conotoxins as novel therapeutics for new receptor targets. In addition, our graph-based approach has allowed us to rank the remaining non-isolated sequences without corresponding characterized structures in an order that would allow for the most rapid expansion of the conotoxin structure library. We also note that of those sequences which were isolated in our graphs–that is, had no edges–80% of those containing four cysteines and 36% of those containing six cysteines were under 20 amino acids long, marking them as good candidates for a high-throughput ab initio modeling procedure, rather than necessarily for experimental characterization, as they will likely be tractable but will not contain any information about other sequences.

One important point about short, disulfide-rich peptides that we have not addressed in this work is the existence of so-called "disulfide isomers." Under certain conditions, there is experimental evidence suggesting that some toxins do not exist as a single set of "native" structures but as a heterogeneous–perhaps metastable–ensemble populated with strikingly different secondary structures corresponding to differing patterns of cysteine connectivity [Paul George et al., 2018, Combelles et al., 2008]. Characterizing multiple possible disulfide isomers is outside the purview

of homology modeling, but it is an important area of future work and sounds a note of caution on the standard interpretation of structure libraries, which generally assume a single "native" structure dictated wholly or primarily by the folding propensities of the amino acid sequence.

Overall, the work in this article presents a rational graph-based algorithm that we employ to expand the repertoire of known conotoxin structures for application in a high-throughput manner as part of the early stages of drug design. We expect that the libraries, the expanded set of structures, and the ranking of sequences in terms of degree of connectedness to other sequences will be valuable resources improving the prospects of conotoxins as novel therapeutic leads and that our approach may be employed for further characterization of other sets of evolutionarily-related toxins.

4 Acknowledgments

T.T and S.G. were supported by the Functional Genomic and Computational Assessment of Threats (Fun-GCAT) program of the Intelligence Advanced Research Projects Activity (IARPA) agency within the Office of the Director of National Intelligence. S.C. and T.T. were also partially supported by the Center for Nonlinear Sciences (CNLS) at LANL. R.A.M gratefully acknowledges a Los Alamos National Laboratory Director's Postdoctoral Fellowship. Triad National Security, LLC (Los Alamos, NM, USA) operator of the Los Alamos National Laboratory under Contract No. 89233218CNA000001 with the U.S. Department of Energy. We thank Dr. Will Fischer for valuable discussions. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.The views and conclusions contained herein are those of the authors, and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Agency (IARPA), Los Alamos National Laboratory (LANL), Department of Energy (DOE), or the US Government.

5 Author Contributions

Conceptualization, R.A.M., S.C., T.T., and S.G.; Methodology, R.A.M., S.C., T.T., and S.G.; Software, R.A.M., S.C., and T.T.; Formal Analysis, R.A.M., S.C. and T.T.; Investigation, R.A.M., S.C. and T.T.; Writing – Original Draft, R.A.M.; Writing – Review & Editing, R.A.M., S.C., T.T., and S.G.; Visualization, R.A.M. and S.C.; Supervision, S.G.; Project Administration, S.G.; Funding Acquisition, S.G.

6 Declaration of Interest

The authors declare no competing interests.

7 STAR Methods

7.1 Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, S. Gnanakaran (gnana@lanl.gov)

7.2 Method Details

For use in construction of the template libraries, we employed a set of 142 conotoxin structures downloaded from the PDB [Berman et al., 2000], which we found by searching "conotoxin" on the PDB. We manually removed several false positives, such as a crystal structure of the acetylcholine-binding protein that was identified due to the title of the associated paper. We also manually removed several sequences that were identical to natural conotoxin sequences but modified by the replacement of disulfide bonds with dicarba bonds. We did not remove redundant sequences

consisting of multiple characterization methods and in a few cases structural isomers resulting from different disulfide-bond connections. In future, further work will be done to properly assess the likelihood of multiple stable or metastable states, but we do not address this consideration further here.

For use in the analysis detailed in this article, we downloaded a set of 6,255 peptide sequences from the Conoserver [Kaas et al., 2012] using the `Tools` > `Download Conoserver's Data` command. We retained only sequences containing four, six, eight, or ten cysteines. We removed anything with the word "precursor" or "patent" in the name, as precursor sequences contain, in addition to the mature peptide sequence that folds into the toxin, a signal sequence and N- and C-terminal pro-regions that are cleaved in the endoplasmic reticulum and Golgi apparatus [Kaas et al., 2010]. A manual inspection of sequences labeled "patent" revealed that many were insufficiently characterized–for example, they noted only the cysteine pattern or they mixed precursor and mature toxin sequences with no indication. We also added to the sequence list any sequence that corresponded to one of the PDB structures that was not already contained in the list. Once the set of all sequences was finalized, we split it into four subsets corresponding to the number of cysteines contained. In the end we retained for analysis a total of 801 unique sequences containing four cysteines, 1,113 unique sequences containing six cysteines, 190 unique sequences containing eight cysteines, and 53 unique sequences containing ten cysteines.

7.2.1 Library template selection procedure

For each subset of sequences corresponding to a different number of contained cysteines, we created an alignment graph as follows. For every sequence we computed a pairwise alignment with every other sequence, using the `PairwiseAligner` class in the `Align` module of the Biopython package [Cock et al., 2009], in global mode, with a gap-open penalty of -10 and a gap-extend penalty of -0.5. Employing the `networkx` Python package [Hagberg et al., 2008], we constructed a graph in which nodes represented sequences and we placed an edge between two nodes whenever the percent identity of the highest-percentage pairwise alignment of the two corresponding sequences was greater than [Rost, 1999],

$$p_{\text{rost}} = n + 480L^{-0.32\left(1+\exp\left(-\frac{L}{1000}\right)\right)}, \tag{1}$$

where L is the length of the alignment in numbers of amino acid residues and we set $n = 5$ (%).

We constructed two different template libraries for each subset of sequences, one from the pairwise alignment graph and one from a static 25%-identity cutoff (with $n = 5$ %). When creating the graph-based libraries (see also Fig. 1), we first identified all connected components in the graph. For each connected component, we chose first the sequence with the highest node degree (number of distinct edges) that corresponded to a structure in the set of 142 structures downloaded from the PDB, added it to the library, and removed that sequence and all sequences it shared an edge with from the graph. We continued this procedure until one of two criteria was satisfied: (i) there were no longer any sequences in the connected component with corresponding structures or (ii) there were no longer any sequences in the connected component without corresponding structures. These criteria corresponded to the following two situations, respectively: (i) there were no other structures available for inclusion in the library or (ii) the entire connected component was able to be homology modeled based on the structures included in the library up until that point. For construction of the static sequence-identity cutoff library, sequences within each data set were clustered and a representative sequence from each cluster chosen by using the `sequence_db.filter` command of MODELLER version 9.20 [Sali and Blundell, 1993, Webb and Sali, 2016], which groups sequences together if their sequence identity is greater than a specified cutoff value. The set of cluster representatives became a library of structures in which between any pair the sequence identity was less than the specified cutoff value.

For computation of the homology modeled structures based on the library templates that were used to assess and compare the quality of the two libraries, we used the `align2d` command followed by the `automodel` procedure from Modeller 9.20 with default parameters. We computed five models for each sequence from each template (except for itself, in the case of library structures being modelled based on other library structures). The best homology model was chosen as the one with the lowest backbone RMSD to the known or experimentally-resolved structure, using the `align` command in PyMOL [Schrödinger LLC, 2015] that superimposes two structures via a structure superposition that is constrained by a prior sequence alignment.

7.2.2 Homology modeling criteria

After assessing the quality of the template libraries, we used the graph-based libraries to construct via homology modeling a database of structures for those conotoxin sequences (set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$ shown in blue in Figs. 2 and S2) that are covered by those libraries (set $\{\mathcal{L}_{\text{ex}}\}$ given in orange in Figs. 2 and S2). The pipeline employed for building these homology-modeled structures is detailed here. A schematic of this pipeline is given in Fig. 5. The 4C subset included 143 such sequences for which structures were computed by homology modeling from 49 library structure templates, while the 6C subset included 148 sequences for which structures were computed by homology modeling from 30 library structure templates. There was only one existing non-isolated library structure template having 8C architecture, and 17 sequences were modeled from it, while no 10C structures could be modeled, since to date there are no structures of conotoxins containing 10 cysteines deposited in the PDB.

Alignment of each of the subject sequences was performed with those sequences that have a structure present in the template library using BLAST [Altschul et al., 1990]. BLOSUM62 substitution matrix [Henikoff and Henikoff, 1992] was used with a gap-open penalty of -10 and a gap-extend penalty of -1. For each sequence, structures were considered possible templates if they fulfilled the following criteria: (i) sequence identity of $\geq 70\%$; (ii) $\geq 70\%$ of sequence length covered; (iii) E-value $\leq 1 \times 10^{-5}$. Additionally, we constrained the cysteines in the sequence to be aligned in the following manner. If there was a one-position shift in the sequence alignment that would allow the cysteines to align, the gap penalties at that position were removed to enforce cysteine alignment. If a greater than one-position shift would be required to allow the cysteines to align, such a template was not considered.

Structural homology modeling was performed using the MODELLER version 9.20 package [Sali and Blundell, 1993, Webb and Sali, 2016]. Multiple templates were used to aid in the modelling process for those subjects where more than one sequence satisfied the above-mentioned criteria. The models were further relaxed by several steps of conjugate gradients and molecular dynamics with simulated annealing as recommended in the thorough Variable Target Function Method (VTFM) optimization of MODELLER [Braun and Go, 1985]. Due to the alignment of cysteines from the template structures, the disulfide bonds could be constrained by patches. Ten such models were generated for each subject sequence. Subjects 107 and 110 from 4C architecture and subject 2 from 6C architecture did not correspond to any templates that satisfied all of our above criteria. Nevertheless, we modeled these sequences based on the best sequence match.

We selected three top models for each subject based on the Discrete Optimized Protein Energy (DOPE) score [Shen and Sali, 2006] and the PROCHECK G-factor [Laskowski et al., 1993]. DOPE, a typical criterion for assessing the quality of a modeled structure, is an atomistic distance-dependent statistical potential calculated from a large set of refined high resolution PDB structures. The PROCHECK G-factor is a log-odds score based on observed distributions of the ϕ-ψ, and χ_1-χ_2 values measuring whether the model is physically reasonable or if it contains unusual stereochemical configurations. In this study, we normalized the DOPE and G-factor scores and used a combined product of probabilities to sort the structures. The top three models selected for each subject are reported in Supplementary File `finalmodels.zip`, along with their DOPE, G-factor, MODELLER optimization function value (molpdf), GA341 scores [John and Sali, 2003], and the Ramachandran plots for each of these models. All RMSD calculations were performed with Pymol [Schrödinger LLC, 2015]. There is only one available non-isolated structure in the 8C extant library. This was used to model all 17 subject sequences. The best three models for each sequence along with their assessment scores are reported in the database.

7.3 Quantification and Statistical Analysis

We use the Kolmogorov-Smirnov two-tailed test as implemented in the SciPy package [Oliphant, 2007] and referred to by the `ks2samp` command to assess whether we may reject the null hypothesis of the RMSDs of experimental structures from homology models based on different template libraries being drawn from the same distribution. We employ a significance level of $p = 0.05$, meaning that we reject the null hypothesis if the KS statistic D returned by the test is such that $D > \alpha\sqrt{\frac{n+m}{nm}}$, where $\alpha = 1.224$ for a significance level of $p = 0.05$, and n and m are the number of samples in each set respectively. The analysis is referred to in Sec. 2.

7.4 Data and Software Availability

Data is available as supplementary files and software is available upon request from the corresponding author (see Next Section).

7.5 Supplemental Information

We provide tables of 8C and 10C sequences in order of projected interest for experimental characterization and four supplementary figures. Structures used as template libraries are provided in the supplementary data folder `libraries.zip`. Homology modeled structures of conotoxins are provided in the supplementary data folder `finalmodels.zip`, along with their scores and the associated Ramachandran plots. Python, Modeller, and Bash analysis scripts for preparation, graph construction and further analysis will be provided upon request.

References

[Abascal and Valencia, 2002] Abascal, F. and Valencia, A. (2002). Clustering of proximal sequence space for the identification of protein families. *Bioinformatics*, 18(7):908–921.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

[Baker and Sali, 2001] Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–6.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242.

[Bolten et al., 2001] Bolten, E., Schliep, A., Schneckener, S., Schomburg, D., and Schrader, R. (2001). Clustering protein sequences–structure prediction by transitive homology. *Bioinformatics*, 17(10):935–941.

[Bradley et al., 2005] Bradley, P., Misura, K. M. S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–71.

[Braun and Go, 1985] Braun, W. and Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm. *J. Mol. Biol.*, 186(3):611–626.

[Cock et al., 2009] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

[Combelles et al., 2008] Combelles, C., Gracy, J., Heitz, A., Craik, D. J., and Chiche, L. (2008). Structure and folding of disulfide-rich miniproteins: Insights from molecular dynamics simulations and MM-PBSA free energy calculations. *Proteins*, 73(1):87–103.

[Dang, 2019] Dang, B. (2019). Chemical synthesis and structure determination of venom toxins. *Chin. Chem. Lett.*, 30(7):1369–1373.

[Dill and MacCallum, 2012] Dill, K. A. and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046.

[Enright and Ouzounis, 2000] Enright, A. J. and Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 16(5):451–457.

[Ensign et al., 2007] Ensign, D. L., Kasson, P. M., and Pande, V. S. (2007). Heterogeneity Even at the Speed Limit of Folding: Large-scale Molecular Dynamics Study of a Fast-folding Variant of the Villin Headpiece. *J. Mol. Biol.*, 374(3):806–816.

[Gomez-Tatay and Hernandez-Andreu, 2019] Gomez-Tatay, L. and Hernandez-Andreu, J. M. (2019). Biosafety and biosecurity in synthetic biology: A review. *Crit. Rev. Env. Sci. Tec.*, 49(17):1587–1621.

[Hagberg et al., 2008] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Gael Varoquaux, T. V. and Millman, J., editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, 89(22):10915–9.

[Huang et al., 2016] Huang, P.-S., Boyken, S. E., and Baker, D. (2016). The coming of age of de novo protein design. *Nature*, 537(7620):320–327.

[John and Sali, 2003] John, B. and Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.*, 31(14):3982–3992.

[Kaas et al., 2010] Kaas, Q., Westermann, J.-C., and Craik, D. J. (2010). Conopeptide characterization and classifications: An analysis using ConoServer. *Toxicon*, 55(8):1491–1509.

[Kaas et al., 2012] Kaas, Q., Yu, R., Jin, A.-H., Dutertre, S., and Craik, D. J. (2012). ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res.*, 40(D1):D325–D330.

[Karp, 1972] Karp, R. M. (1972). *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA.

[Kong et al., 2004] Kong, L., Lee, B. T. K., Tong, J. C., Tan, T. W., and Ranganathan, S. (2004). SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. *Nucleic Acids Res.*, 32:W356–W359.

[Krieger et al., 2003] Krieger, E., Nabuurs, S. B., and Vriend, G. (2003). Homology modeling. *Methods Biochem. Anal.*, 44:509–23.

[Larsson, 2014] Larsson, A. (2014). Aliview: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics*, 30(22):3276–3278.

[Laskowski et al., 1993] Laskowski, R. A., MacArthur, M. W., Moss, D. S., Thornton, J. M., and IUCr (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, 26(2):283–291.

[Mansbach et al., 2019] Mansbach, R. A., Travers, T., McMahon, B. H., Fair, J. M., and Gnanakaran, S. (2019). Snails In Silico: A Review of Computational Studies on the Conopeptides. *Mar. Drugs*, 17(3):145.

[Miljanich, 2004] Miljanich, G. (2004). Ziconotide: Neuronal Calcium Channel Blocker for Treating Severe Chronic Pain. *Curr. Med. Chem.*, 11(23):3029–3040.

[Oliphant, 2007] Oliphant, T. E. (2007). Python for scientific computing. *Comput. Sci. Eng.*, 9(3):10–20.

[Paul George et al., 2018] Paul George, A. A., Heimer, P., Maaß, A., Hamaekers, J., Hofmann-Apitius, M., Biswas, A., and Imhof, D. (2018). Insights into the Folding of Disulfide-Rich μ-Conotoxins. *ACS Omega*, 3(10):12330–12340.

[Pipenbacher et al., 2002] Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D., and Schrader, R. (2002). ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, 18(Suppl 2):S182–S191.

[Pitera and Swope, 2003] Pitera, J. W. and Swope, W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl. Acad. Sci. U.S.A.*, 100(13):7587–92.

[Romano and Tatonetti, 2019] Romano, J. D. and Tatonetti, N. P. (2019). Informatics and Computational Methods in Natural Product Drug Discovery: A Review and Perspectives. *Front. Genet.*, 10:368.

[Rost, 1999] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.*, 12(2):85–94.

[Sali and Blundell, 1993] Sali, A. and Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.*, 234(3):779–815.

[Santiago et al., 2018] Santiago, C., Pereira, V., and Digiampietri, L. (2018). Homology Detection Using Multilayer Maximum Clustering Coefficient. *J. Comput. Biol.*, 25(12):1328–1338.

[Sborgi et al., 2015] Sborgi, L., Verma, A., Piana, S., Lindorff-Larsen, K., Cerminara, M., Santiveri, C. M., Shaw, D. E., de Alba, E., and Muñoz, V. (2015). Interaction Networks in Protein Folding via Atomic-Resolution Experiments and Long-Time-Scale Molecular Dynamics Simulations. *J. Am. Chem. Soc.*, 137(20):6506–6516.

[Schrödinger LLC, 2015] Schrödinger LLC (2015). The PyMOL Molecular Graphics System, Version 1.8. Technical report.

[Shen and Sali, 2006] Shen, M.-y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15(11):2507–2524.

[Simons et al., 1997] Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268(1):209–225.

[Śledź and Caflisch, 2018] Śledź, P. and Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struc. Biol.*, 48:93–102.

[Uribe et al., 2018] Uribe, J. E., Zardoya, R., and Puillandre, N. (2018). Phylogenetic relationships of the conoidean snails (Gastropoda: Caenogastropoda) based on mitochondrial genomes. *Mol. Phylogenetics Evol.*, 127:898–906.

[Verdes et al., 2016] Verdes, A., Anand, P., Gorson, J., Jannetti, S., Kelly, P., Leffler, A., Simpson, D., Ramrattan, G., Holford, M., Verdes, A., et al. (2016). From Mollusks to Medicine: A Venomics Approach for the Discovery and Characterization of Therapeutics from Terebridae Peptide Toxins. *Toxins*, 8(4):117.

[Voelz et al., 2010] Voelz, V. A., Bowman, G. R., Beauchamp, K., and Pande, V. S. (2010). Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.*, 132(5):1526–1528.

[Webb and Sali, 2016] Webb, B. and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. In *Curr. Protoc. Bioinformatics*, volume 54, pages 1–5. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[Xiang, 2006] Xiang, Z. (2006). Advances in Homology Protein Structure Modeling. *Curr. Protein Pept. Sci.*, 7(3):217–227.

[Yan et al., 2011] Yan, Y., Zhang, S., and Wu, F.-X. (2011). Applications of graph theory in protein structure identification. *Proteome Sci.*, 9(Suppl 1):S17.

[Yang et al., 2015] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, 12(1):7–8.

[Younis and Rashid, 2017] Younis, S. and Rashid, S. (2017). Alpha conotoxin-BuIA globular isomer is a competitive antagonist for oleoyl-L-alpha-lysophosphatidic acid binding to LPAR6; A molecular dynamics study. *PloS One*, 12(12):e0189154.

[Zambelli et al., 2016] Zambelli, V., Pasqualoto, K., Picolo, G., Chudzinski-Tavassi, A., and Cury, Y. (2016). Harnessing the knowledge of animal toxins to generate drugs. *Pharmacol. Res.*, 112:30–36.

[Zhang, 2008] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9(1):40.

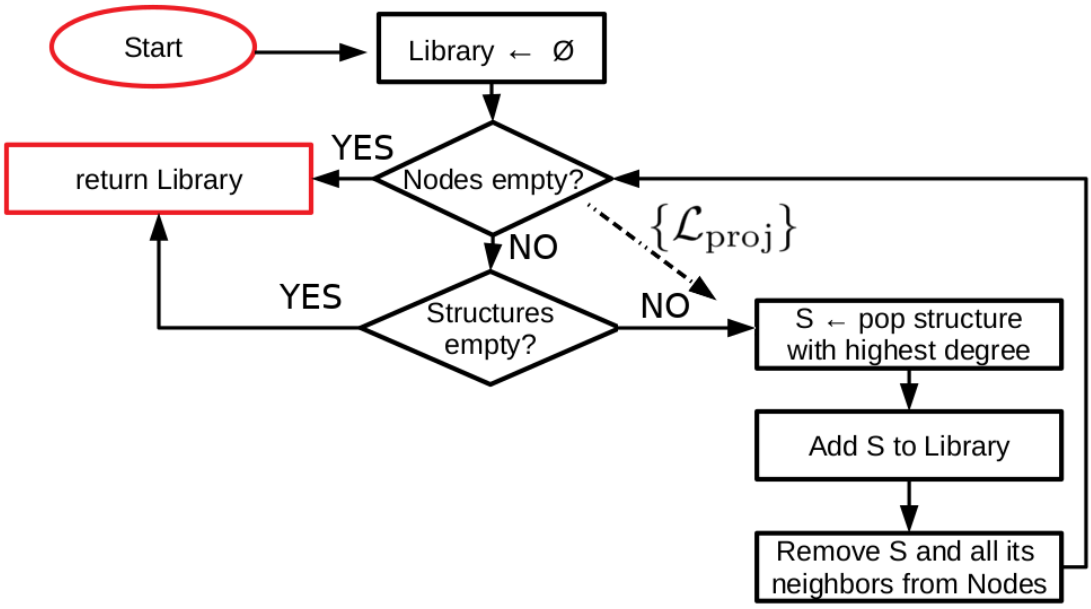

Figure 1: Schematic of a simple graph-based algorithm for constructing a library of structural templates for homology modeling. For each connected component in the graph of sequences, where an edge represents the ability to homology model one sequence based on another, we employ a greedy approach to find a good library of template structures that cover as much of the sequence space as possible. For computation of the sequence set $\{\mathcal{L}_{\text{proj}}\}$ of interest for experimental or ab initio characterization, we skip consideration of the structures and run the algorithm on the subset with structure-associated sequences removed.

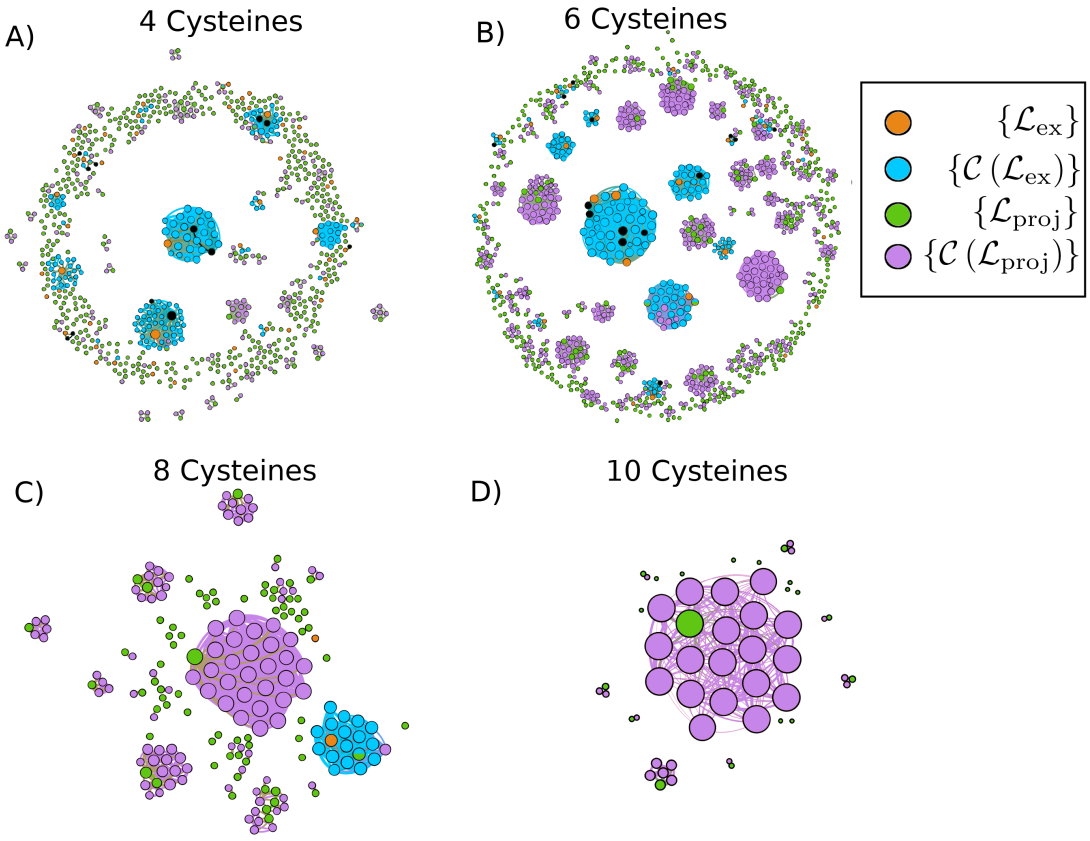

Figure 2: Graph of conotoxins containing (A) four cysteines, (B) six cysteines, (C) eight cysteines and (D) ten cysteines where nodes are sequences and edges exist between sequences with pairwise alignments that have high enough length and percent identity to fall above the Rost curve with $n = 5\%$ (Eqn. 1). We show the set $\{\mathcal{L}_{\text{ex}}\}$ of sequences added to the template libraries in orange, the set of sequences corresponding to unselected structures in black, the set of covered sequences $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$ that we homology model based on the templates included in the library in blue, and the set of projected sequences $\{\mathcal{L}_{\text{proj}}\}$ in green whose structures are in need of characterization in order that the rest of the sequences $\{\mathcal{C}(\mathcal{L}_{\text{proj}})\}$ in magenta may be homology modeled based on some template. The sizes of the nodes corresponds to their degree; that is the number of other sequences that they can be modeled based on or used to model. Node locations and edge lengths were chosen for ease of visualization of separate connectec components. Visualization of the graphs was produced with Gephi 0.9.2 [Bastian et al., 2009].

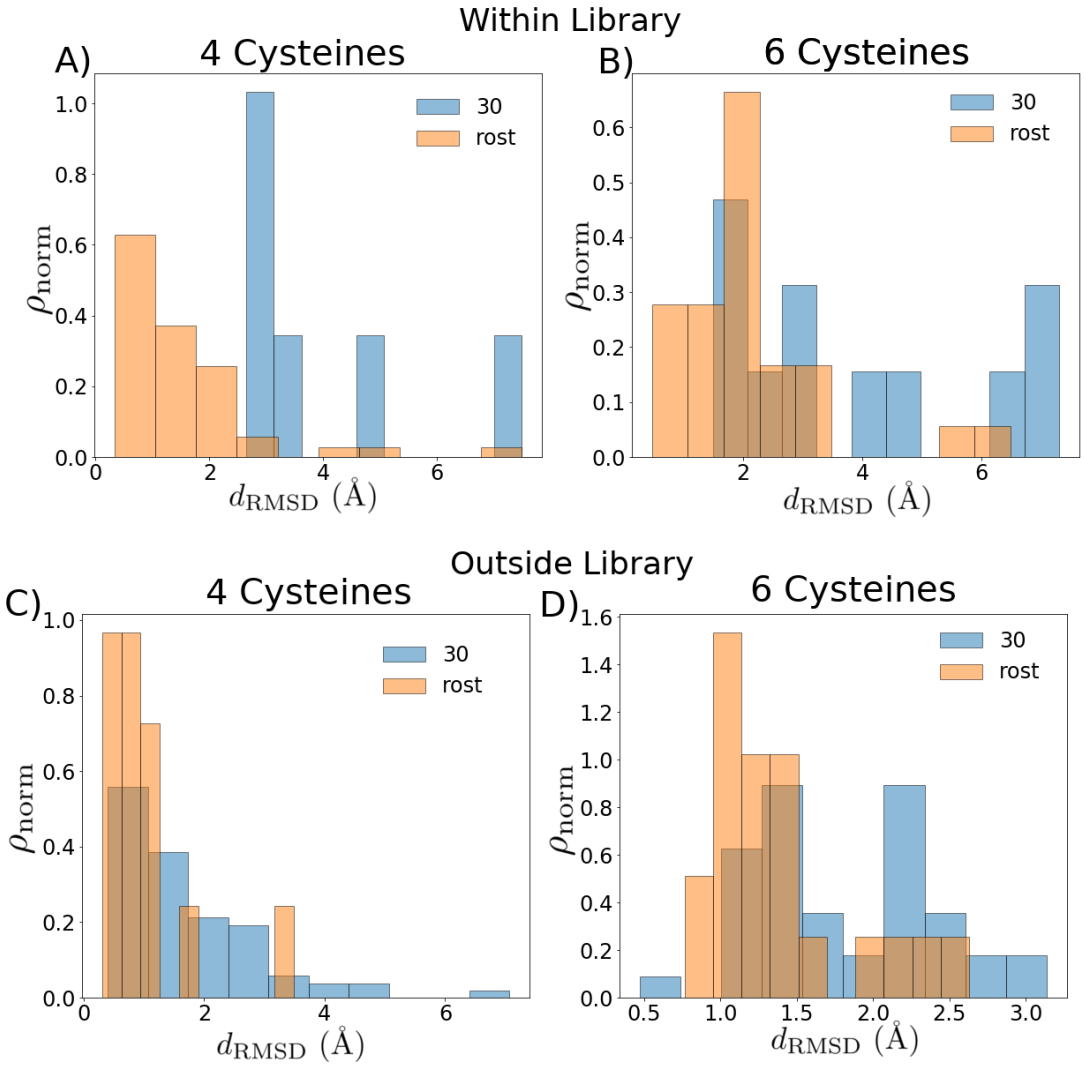

Figure 3: Quality of graph-based template library selection criteria. Comparison of root-mean-square deviation (RMSD) distributions from experimental structures for (A-B) structures within the libraries, with each structure modeled by selecting from all other templates within the given library, and (C-D) structures outside the libraries modeled by selecting from all templates within the given library. For each homology modeled structure, we choose the best fit to experiment. The distributions produced by the simple 30% cutoff libraries are shown in blue; the distributions produced by using the graph-based algorithm are shown in orange.

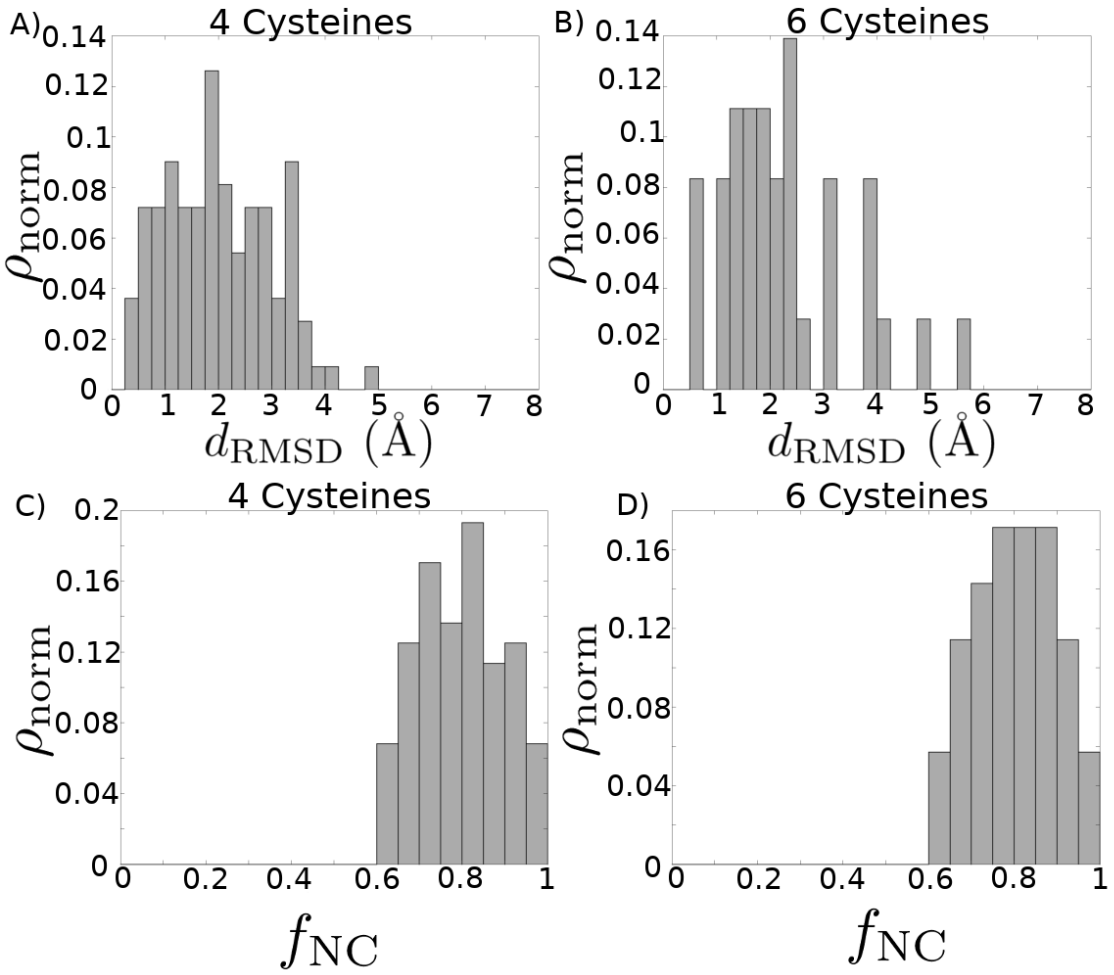

Figure 4: Quality of modeling criteria. (A-B) Distribution of root-mean-square deviation (RMSD) for homology models compared with their corresponding experimental structures, without prior removal of any structural alignment outliers. Each experimental structure present in the library was modeled by selecting from all other templates in the library. The top three models for each structure based on combined MODELLER DOPE and PROCHECK G-FACTOR scores are considered here. (A) Distribution mean = 2.00Å, standard deviation = 0.97Å. (B) Distribution mean = 2.25Å, standard deviation = 1.20Å. (C-D) Distribution of fraction of native contacts present in each of the homology modeled structures, with respect to the experimental structure. Each experimental structure present in the library was modeled by selecting from all other templates in the library. The top three models for each structure based on combined MODELLER DOPE and PROCHECK G-FACTOR scores are considered here. (C) Distribution mean = 0.797, standard deviation = 0.108. (D) Distribution mean = 0.805, standard deviation = 0.097.

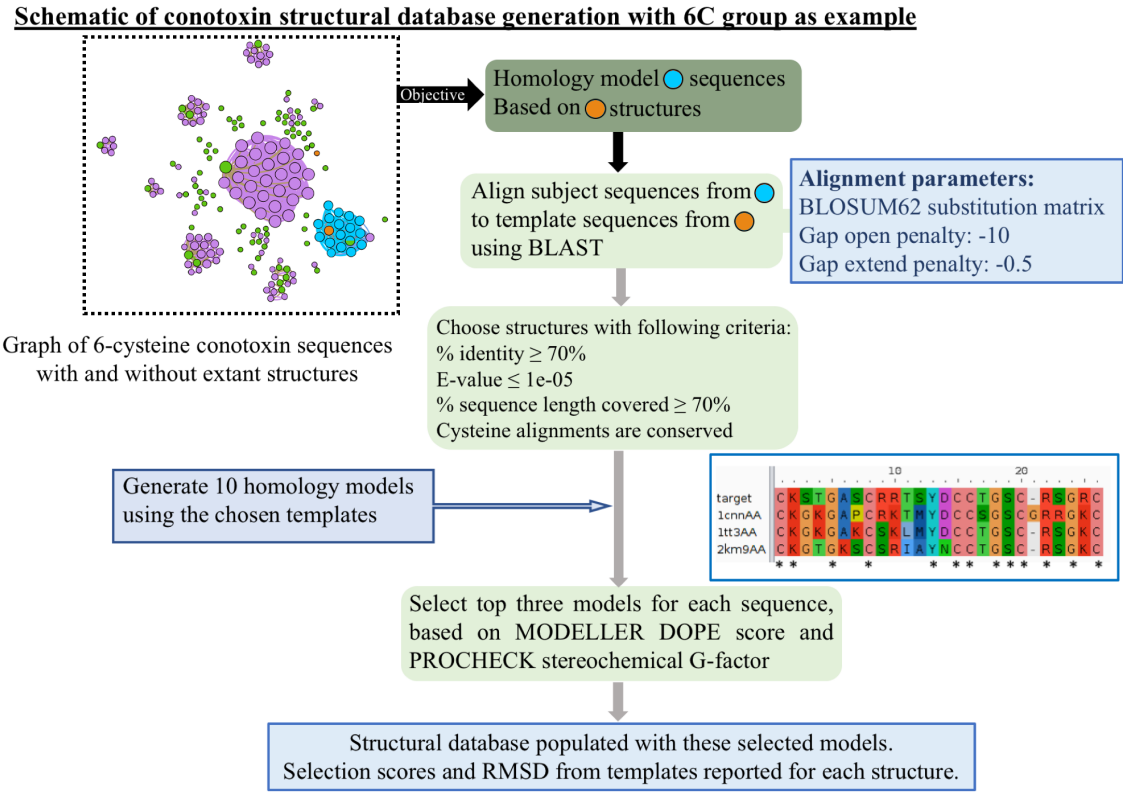

Figure 5: Schematic of procedure for producing homology modeled structures from library templates for conotoxin sequences with unknown structure lying in the set $\{\mathcal{C}(\mathcal{L}_{\text{ex}})\}$. Graph inset of eight cysteine graph is an example. The inset consisting of an example alignment input figure was created using the alignment obtained from BLAST [Altschul et al., 1990] and visualized with Aliview [Larsson, 2014].

Table 1: List of conotoxins with corresponding PDB structure IDs [Berman et al., 2000] comprising 4C library. Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species or has different post-translational modifications.

Name(s)	**PDB ID**	**Sequence**
EpI [sTy15>Y], EpI	1a0m	GCCSDPRCNMNNPDYC
PnIB, PnIB [sTy15Y]	1akg	GCCSLPPCALSNPDYC
CnIA	1b45	GRCCHPACGKYYSC
AuIB, Ac-AuIB, AuIB [ribbon isoform]	1mxp	GCCSYPPCFATNPDC
ImI [R11E]	1e74	GCCSDPRCAWEC
ImI [R7L]	1e75	GCCSDPLCAWRC
ImI [D5N]	1e76	GCCSNPRCAWRC
TIA	2lr9	FNWRCCLIPACRRNHKKFC
MrIB, MrIB C-term amidated	1ieo	VGVCCGYKLCHPC
EI	1k64	RDPCCYHPTCNMSNPQIC
GID, GID*, GID*-NH2, GID*[O16P]	1mtq	IRDECCSNPACRVNNPHVC
SI	1hje	ICCNPACGPKYSC
TXIX	1wct	ECCEDGWCCXAAP
GI	1xga	ECCNPACGRHYSC
Conkunitzin-S1	1y62	RPSLCDLPADSGSGTKAEKRI-YYNSARKQCLRFDYTGQGGN-ENNFRRTYDCQRTCL
PIA, PIA [R1ADMA]	1zlc	RDPCCSNPVCTVHNPQIC
cMII-6	2ajw	GCCSNPVCHLEHSNLCGGAAGG
PlXIVA	2fqc	FPRPRICNLACRAGIGHKYPF-CHCR
GI (SER12)-benzoylphenylalanine	2fr9	ECCNPACGRHYYC
GI (ASN4)-benzoylphenylalanine	2frb	ECCYPACGRHYSC
OmIA	2gcz	GCCSHPACNVNNPHICG
BuIA, BuIA[P6O], BuIA[P7O]	2ns3	GCCSTPPCAVLYC
ImI [P6A]	2ifi	GCCSDARCAWRC
ImI [P6K], ImI [P6K] deamidated	2ifj	GCCSDKRCAWRC
ImI, ImI [C2U,C8U], ImI [C2U,C3U,-C8U,C12U], ImI deamidated, Ac-Im-I, ImI [A9S], ImI [C3U,C12U], ImI [P-60], ImI [P6APro], ImI [P6A(S)Pro], ImI [P6guaPro], ImI [P6betPro], ImI [P6fluoPro], ImI [P6fluo(S)Pro], ImI-[P6phiPro], ImI [P6phi(S)Pro], ImI -[P6benzPro], ImI [P6naphPro], ImI [P-6phi(3S)Pro], ImI [P6phi(5R)Pro]	2bypF	GCCSDPRCAWRC
CMrVIA [K6P], CMrVIA [K6P] amidated	2ih7	VCCGYPLCHPC
CMrVIA, CMrVIA amidated	2b5p	VCCGYKLCHPC
Cyclic MrIA	2j15	NGVCCGYKLCHPCAG
RgIA [P6V]	2juq	GCCSDVRCRYRCR
RgIA [D5E]	2jur	GCCSEPRCRYRCR
RgIA [Y10W]	2jus	GCCSDPRCRWRCR
RgIA	2jut	GCCSDPRCRYRCR
Pc16a	2ler	SCSCKRNFLCC
Midi	2lu6	CNCSRWARDHSRCC
TxIB	2lz5	GCCSDPPCRNKHPDLC
Li1.12, TxID	2m3i	GCCSHPVCSAMSPIC
Ar1248	2m62	GVCCGVSFCYPC

Lo1a	2md6	EGCCSNPACRTNHPEVCD
LvIA	5xgl	GCCSHPACNVDHPEIC
Exendin-4/conotoxin chimera (Ex-4[1-27]/pl14a)	2naw	HGEGTFTSDLSKQMEEEAVRC-FIECLKGIGHKYPFCHCR
Bt1.8	2nay	GCCSNPACILNNPNQC
TXIA(A10L)	2uz6	GCCSRPPCILNNPDLC
CnVA	3zkt	ECCHRQLLCCLRFV
Cyclic Vc1.1	4ttl	GCCSDPRCNYDHPEICGGAAGG
GIC	1ul2	GCCSHPACAGNNQHIC
PeIA, Bt1.4, PeIA[P6O], PeIA[P13O]	5jmeF	GCCSHPACSVNHPELC
Pn10.1	5t6v	STCCGYRMCVPC
LsIA, LsIA#	5t90F	SGCCSNPACRVNNPNIC
VilXIVA	6efe	GGLGRCIYNCMNSGGGLSFIQ-CKTMCY

Table 2: List of conotoxins with corresponding PDB structure IDS [Berman et al., 2000] comprising 6C library. Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species or has different post-translational modifications.

Name(s)	**PDB ID**	**Sequence**
conotoxin-GS	1ag7	ACSGRGSRCPPQCCMGLRCGRGNPQKCIG-AHEDV
PIIIE, PIIIE [K9S], P-IIIE [S17Y,S18N,S20L]	1jlo	HPPCCLYGKCRRYPGCSSASCCQR
MVIIC, S6.6	1omn	CKGKGAPCRKTMYDCCSGSCGRRGKC
TVIIA	1eyo	SCSGRDSRCPPVCCMGLMCSRGKCVSIYGE
TxVII	1f3k	CKQADEPCDVFSLDCCTGICLGVCMW
TxVIA	1fu3	WCKQSGEMCNLLDQNCCDGYCIVLVCT
EVIA	1g1z	DDCIKPYGFCSLPILKNGLCCSGACVGVCADL
GIIIB	1gib	RDCCTPPRKCKDRRCKPMKCCA
GVIA	1ttl	CKSPGSSCSPTSYNCCRSCNPYTKRCY
PIVA	1p1p	GCCGSYPNAACHPCSCKDRPSYCGQ
EIVA	1pqr	GCCGPYPNAACHPCGCKVGRPPYCDRPSGG
PIIIA	1r9i	QRLCCGFPKSCRSRQCKPHRCC
MVIIA[R10K]	1tt3	CKGKGAKCSKLMYDCCTGSCRSGKC
Am2766	1yz2	CKQAGESCDIFSQNCCVGTCAFICIE
MrIIIE	2efz	VCCPFGGCHELCYCCD
FVIA	2km9	CKGTGKSCSRIAYNCCTGSCRSGKC
Im23a, Mr23a	2lmz	IPYCGQTGAECYSWCIKQDLSKDWCCDFV-KDIRMNPPADKCP
BuIIIB	2lo9	VGERCCKNGKRGCGRWCRDHSRCC
KIIIA, KIIA [W8dTrp]	2lxg	CCNCSSKWCRDHSRCC
Ar1446	2m61	CCRLACGLGCHPCC
cGm9a	2mso	SCNNSCQSHSDCASHCICTFRGCGAVNGLP
cBru9a	2msq	SCGGSCFGGCWPGCSCYARTCFRDGLP
Mo3964	2mw7	DGECGDKDEPCCGRPDGAKVCNDPWVCIL-TSSRCENP
MfVIA	2n7f	RDCQEKWEYCIVPILGFVYCCPGLICGPFVCV
cyclic PVIIA	2n8e	CRIPNQKCFQHLDDCCSRKCNRFNKCVLP-ETGGG
conotoxin-muOxi-GVIIJ	2n8h	GWCGDPGATCGKLRLYCCSGFCDSYTKTC-KDKSSA
CnIIIC	2yen	QGCCNGPKGCSSKWCRDHARCC
CcTx	4b1qP	APWLVPSQITTCCGYNPGTMCPSCMCTNTC

Reg12i	6bx9	CCTALCSRYHCLPCC
MoVIB	6ceg	CKPPGSKCSPSMRDCCTTCISYTKRCRKYY

Table 3: List of conotoxins with corresponding PDB structure IDS [Berman et al., 2000] comprising 8C library. Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species or has different post-translational modifications.

Name(s)	**PDB ID**	**Sequence**
G11.1	6cei	CAVTHEKCSDDYDCCGSLCCVGICAKTIAPCK
RXIA, RXIA [Btr33>W]	2p4l	GPSFCKADEKPCEYHADCCNCCLSGICAPSTN-WILPGCSTSSFFKI

Table 4: List of sequences containing four cysteines in order of interest for experimental characterization, based on degree (sequence coverage) in alignment graphs (cf. Fig. 2). Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species or has different post-translational modifications. Node degree corresponds to the number of sequences with pairwise alignments that are long enough and have high enough percent identity to be homology modeled with the given sequence as a template.

Sequence	**Name(s)**	**Degree (# Edges)**
AAKVKYSNTPEECCSNPPCFATHSEICG	Li1.28	10
GCCSDPRCAYDHPEIC	Vc1.1[N9A]	10
GCCSNPVCHLEHSNAC	MII [L15A]	8
AALEDADMKTEKGFLSSIVGNLGTV-GNLVGSVCCQITNSCCPED	Pu5.7	7
RAALEDADMKTEKGVLNAIFSNLGD-LGNLVSSVCCKATTSCCPED	Pu5.9	6
AGLTDADLKTEKGFLSGLLNVAGSV-CCKVDTSCCSNQ	Lt5g	6
GCCSNPVCALEHSNLC	MII [H9A]	6
VPAEQMMEELCPDMCNRGEGEIICT-CVLRRHVVSPSIR	Lt14.4	5
TNEGPGRDPAPCCQHPIETCC	Cal5b	5
RPECCTHPACHVSNPELCS	Mr1.8	4
GCCSRPPCIANNPDLC	TxIA	4
SPGSTICKMACRTGNGHKYPFCNCR	Fe14.1	4
GCCSLPPCALNNPDYC	PnIA [A10L,sTy15Y]	4
YAAVVNRASALMAQAVLRDCCSNPP-CAHNIHCA	Ec1.7	4

NGRCCHPACGKHFSC	Ac1.1b, CnIH, R1.1, Bt1.6-, Mn1.2, C4.3	4
NGRCCHPACGKYFSC	Mn1.5	3
GCCSRAACAGIHQELC	LtIA [A4S]	3
GCCSNPVCHLAHSNAC	MII [E11A,L15A]	3
GCCSHPACSGNNREYCRES	O1.3	3
GGCCSHPVCYFNNPQMCR	Cr1.6	3
GGGCCSHPACAANNQDYC	Gly-AnIB	3
DGCCSSPSCSVNNPDICGG	Eb1.1, Qc1.18	3
LDPCCREPPCASTHTDICT	Li1.4, Sa1.12	3
NECCDNPPCKSSNPDLCDWRS	Qc1.1b, LiC22	3
DECCSNPSCAQTHPEVC	Li1.24, Sa1.6	3
GCCSHPACAGNNPHICS	Li1.11	3
SFRFIPGGIKEIACHRYCAKGIASA-FCNCPDKRDVVSPRI	G14.1	3
VPPEPILEIICPGMCDEGVGKEPFC-HCTKKRDAVSSRI	Vc14.4	3
GCCSYPPCNVSYPEICG	Su1.6	2
AANDKASVQIALTVQECCADSACSL-TNPLIC	Dd1.7, Li1.21	2
TAFGLRLCCKRHHGCHPCGRT	Cal1b	2
AANAKLFDVGQSCCSAPLCALLYMVIC	Sa1.7	2
TVRDACCSDPRCSGKHQDLC	Li1.16, Sa1.3	2
NLQILCCKHTPACCT	S5.3, Eb5.5	2
ECPPWCPTSHCNAGTC	Cl14c	2
GIWCDPPCPKGETCRGGECSDEFNSDV	Cal14.1a	2
GMWDECCDDPPCRQNNMEHCPAS	Lp1.7	2
GRCCHPACGGKYFKC	CnIJ	2
IALIATRECCANPQCWSKNC	Co1.3	2
GCCSHPVCHARHPELC	PeIA[A7V,S9H,V10A,N11R]	2
DGCCSDPACSVNHPDICGG	Qc1.7	2
PPGCCNNPACVKHRCG	Bu1.2	2
LINTRCCPGQPCCRM	Vc5.11	2
NAAANDKASDVIPLALQGCCSNPVC-HVDHPELCL	Cn1.6	2
GCCSHPVCHARHPALC	PeIA[A7V,S9H,V10A,N11-R,E14A]	2
WDVNDCIHFCLIGVVGRSYTECHTMCT	FlfXIVB	2
NGRCCHPACAKYFSC	Mn1.4b	2
NGRCCHPACGGKYVKC	Ac1.2	2
GCCSYPPCFATNSDYC	AuIA	2
DECCAIPLCAKIFPGRCP	Pc1b	1
AANLMALLQESLCPPGCYPSCTNCR-YMFP	Pu14.6	1
GCCAIRECRLQNAAYCGGIY	Ca1.2	1
FLTQQSPRDFAKSVMQLLHYNWIDC-CNYGVSDCCI	Lv5.7	1
APAELILETICPHMCGTGIGEPFCN-CRNKRDVVSSRII	Bt14.3	1

EIVNIIDSISDVAKQICCEITVQCCVLDEE	Vn5.5	1
ECCEDGWCCTAAPLTAP	Vc5.7	1
CCPGWELCCEWDDWW	Mr5.7	1
GCCSFPACRKYRPEMCG	Su1.2	1
DDCCPDPACRQNHPELCST	PuSG1.1	1
APNVKDSKASGSCCDNPSCAVNNSHC	Li1.32	1
YHECCKNPPCRNKHPDLC	Sa1.16	1
GCCSNPACAGSNAHIC	Li1.14	1
GCCVYPPCAVNHPDICRG	Qc1.9	1
VMQLRYYNWIDCCFDGDCCN	Qc5.3	1
TGCCEYPYCAENNPELCG	Co1.4	1
SVEGVISTIKDFAVKVCCSVSLKFC-CPTA	Ts5.5	1
SCCSDSDCNANHPDMCS	Leo-A1	1
SCCPQEFLCCLYLVK	Lp5.1	1
RCCHPACGKNYSC	MI[del1G]	1
QTPGCCWNPACVKNRC	EIIA	1
QGCCSYPACAVSNPDICGG	Qc1.12	1
PECCSDPRCNSTHPELCG	Ai1.2	1
NIQIICCKHTPKCCT	Tx5.5	1
NAWLTPEECCAAPACREMILEFCLA-GEAFAAALDGFRRLPYR	Pu1.5	1
KVYCCLGVRDDWCCAGQIQI	Lt5i	1
IINWCCLIFYQCCL	Sr5.7	1
YCCHPACGKNFDC	SIA	1
GILELAKTVCCSATGISICC	Tx5.13, Tr5.3, Vr5.1	1
GGCCSRPPCILKHPEIC	Qc1.13	1
GCPADCPNTCDSSNKCSPGFP	Cal14a	1
GIRGNCCMFHTCPIDYSRFYCP	Vt1.24	1

Table 5: List of sequences containing six cysteines in order of interest for experimental characterization, based on degree (sequence coverage) in alignment graphs (cf. Fig. 2). Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species or has different post-translational modifications. Node degree corresponds to the number of sequences with pairwise alignments that are long enough and have high enough percent identity to be homology modeled with the given sequence as a template.

Sequence	**Name(s)**	**Degree**
LPPCCSLNLRLCPAPACKYKPCCKS	RIIIJΔ6-11	29
QKGLVPSVITTCCGYDPGTMCPPCR-CTNSCPKKPKKP	S4.4	28
QPWLVPSKITNCCGYNTMEMCPTCM-CTYSCRPKKKKP	Mn4.2	27

DDECEPPGDFCGFFKIGPPCCSGWC-FLWCA	MaIr137, G6.2	20
DCVAGGHFCGFPKIGGPCCSGWCFF-VCA	Vn6.8	20
VCREKGQGCTNTALCCPGLECEGQS-QGGLCVDN	Mi010	20
ECREQSQGCTNTSPPCCSGLRCSGQ-SQGGVCISN	CaHr91	17
TVDEACNEYCEERNKNCCGRTDGEP-VCAQACL	Vi6.7	15
ECTRSGGACYSHNQCCDDFCSTATS-TCV	Eb6.22	15
GCTPPGGACGGHAHCCSQSCNILAS-TCNA	ABVIC	15
TVGEECNEYCEQRNKNCCGKTNGEP-VCAQACL	Tr7.4	15
TATEECEEYCEDEEKTCCGEEDGEP-VCARFCL	Ar6.24	14
EACYNAGTFCGIKPGLCCSAICLSF-VCISFDLIDVFSSP	M6.2	13
TTEECHEYCEDQNKNCCGLTDGEPR-CAGMCL	Tr7.3	13
MTMGCTHPGGACGGHYHCCSQSCNT-AANSCN	MIL3-b (partial)	12
VPEECEESCEEEEKTCCGLENGQPF-CSRICW	Ar6.28	12
DECYPPGTFCGIKPGLCCSERCFPF-VCLSLEF	Ac6.2	12
CLDAGEVCDIFFPTCCGYCILLFCA	TxO1	11
DCTPPDGACGFHYHCCSKFCITISSTCN	MIL2-a	11
CIDGGEICDIFFPNCCSGWCIILVCA	Mr6.8	11
TTAESWWEGECLGWSNGCTHPSDCC-SNYCKGIYCDL	Mr6.16	11
GCTHPGGACGGHHHCCSLFCNTAAN-ACN	MIL3-f	11
CLGSGETCWLDSSCCSFSCTNNVCF	Vn6.15	11
SIAGRTTTEECDEYCEDLNKNCCGL-SNGEPVCATACL	Ts6.7	10
CLDAGEMCDLFNSKCCSGWCIILFCA	Mr6.1	10
GCLEVDYFCGIPFVNNGLCCSGNCV-FVCTPQ	Pn6.7	10
SCGEEGEGCYTRPCCPGLKCIGTAH-GGLCREE	Pu6.7	10
DGCYNAGTFCGIRPGLCCSEFCFLW-CITFVDS	MVIA, Cn6.1	10
NCCNGGCSSKWCRDHARCC	SIIIA[del1]	9
RHGCCKGPKGCSSRECRPQHCC	TIIIA	8
DCGEQGQGCYTRPCCPGLHCAAGAT-GGGSCQP	Conotoxin-1	8
CLAGSAPCEFHRGYTCCSGHCLIWVCA	Cal6.1d	8
KTTAESWWEGECYGWWTSCSSPEQC-CSLNCENIYCRAW	TsMEKL-03	8

KTTAESWWEGECRTWYAPCNFPSQC-CSEVCSSKTGRCLTW	Vn6.5	7
CTPGGEACDATTNCCFLTCNLATNK-CRSPNFP	ABVIL	7
CRPPGMVCGFPKPGPYCCSGWCFAV-CLPV	MaIr193	7
WWEGECRGWSNGCTTNSDCCSNNCD-GTFCKLW	Vn6.3	7
CCSRDCWVCIPCCPNGSA	Lv3-IP01	7
WWWGGCTWWFGRCSTDSECCSNSCD-QTYCELYRFPSRY	Vc6.26	7
YECYSTGTFCGINGGLCCSNLCLFF-VCLTFS	CnVIA, St6.2	7
CCSQDCSVCIPCCPN	Co3-IP02, Ts3-IP07, Vr3-I-P08, Rt3-IP03, Ca3-IP02,-Ec3-IP03	6
FPCNPGGCACRPLDSYSYTCQSPSS-STANCEGNECVSEADW	Cl9.4	6
CTVDSDFCDPDNHDCCSGRCIDEGG-SGVCAIVPVLN	Ar6.19	6
CYDSGTSCNTGNQCCSGWCIFVSCL	Tx6.3	6
CCSQDCWVCIPCCPN	Eu3.2	6
ECIEGSEPCEVFRPYTCCSGHCIIFVCA	Cal6.1h	6
VCVDGGTFCGFPKIGGPCCSGWCIF-VCL	Ar6.2	6
GPPCCLYGSCRPFPGCSSASCCRK	PIIIF [Y17S,N18S,L20S]	5
CCGVPNAACHPCVCNNTC	OIVA [K15N]	5
DCQEKWDYCPVPFLGSRYCCDGFIC-PSFFCA	Da6.6, Tx6.6	5
CTPRNGYCYYRYFCCSRACNLTIKRCL	Ml6.2	5
CTPCGPDLCCEPGTTCDTVLHHTHF-GEPSCSY	Fla6.15	5
CCSQDCRVCIPCCPN	Ts3.1	5
QCTPVGGSCSRHYHCCSLYCNKNIG-QCLATSYP	Ar6.17	5
VKPCSEEGQLCDPLSQNCCRGWHCV-LVSCV	Da6.2	5
CLNDGDDCDTGDDCCSGLCIFDEYF-SYCDDSDPYYDDYDEYYY	Mi029	5
SCGNLHESCSAHRCCPGLKCIGTAH-GGLCRE	Pu6.15 (partial)	5
WWDGECRLWSNGCRKHKECCSNHCK-GIYCDIW	VeG52	4
CCGKPNAACHPCVCNGSCS	G4.1	4
MGYILPALSQQTCCVRPWCDGACDC-CVDS	Co3-D01	4
MKLMLSALRQQECCKPSTCDGGCYH-CC	Lv3-YH04	4
CIPQFDPCDMVRHTCCKGLCVLIAC-SKTA	Pn6.3	4
SCGNLHESCSAHRCCPGLMCFTLPT-PICIW	Pu6.17	4

STSCMEAGSYCGSTTRICCGYCAYF-GKKCIDYPSN	SO5	4
GGCTPCGPNLCCSEEFRCGTSTHHQ-TYGEPACLSY	Ca6.2	4
FAVIFTCTPPGSHCTGHSDCCSDFC-STMSDVCQ	Co6.1	4
CIEQFDPCEMIRHTCCVGVCFLMACI	King-Kong 1	4
CLGFGEACLMLYSDCCSYCVALVCL	Ep6.1	4
TCSPAGEVCTSKSPCCTGFLCTHIG-GMCHH	LvVIA 2	4
CTPSGGACYVASTCCSNACNLNSNKCV	M1	4
CCGVPNAACHPCVCTGKC	PeIVA	4
ATDCIEAGNYCGPTVMKICCGFCSP-FSKICMNYPQN	Ac6.5	4
GCTPRNGACGYHSHCCSNFCHTWAN-VCL	LvVID	3
DVCELPFEEGPCFAAIRVYAYNAKT-GDCEQLTYGGCEGNGNRFATLEDC-DNACARY	Cal9.1d	3
GCGYLGEPCCVAPKRAYCHGDLECN-SVAMCVN	Mr2	3
ECTPPEGACNHPSHCCEDFCDRGRN-RCM	At6.7	3
KFCCDSNWCHISDCECCY	Tx3h	3
WWEGDCTDWLGSCSSPSECCYDNCE-TYCTLW	Lt7b	3
CRSSGSPCGVTSICCGRCYRGKCT	SVIA	3
CKAESEACNIITQNCCDGKCLFFCI-QIPE	Pn6.5	3
CTPPSGYCYHPYYCCSRACNLTRKRCL	At6.2	3
CKSPGTPCSRTMRDCCTSCLSYSKKCR	G6.12	3
PCKTPGRKCFPHQKDCCGRACIITICP	P2a	3
CVPYEGPCNWLTQNCCDELCVFFCL	Gm6.3	3
SKQCCHLPACRFGCTPCCW	Mr3.4	3
CCKYGWTCWLGCSPCGC	PnIVB	2
EIILHALGTRCCSWDVCDHPSCTCC	Vr3-T05	2
CCHWNWCDHLCSCCGS	Mr3.8	2
CAGIGSFCGLPGLVDCCSGRCFIVCLP	Bt6.4, ErVIA	2
CCQAACSPWLCLPCC	Eu3.3, Bt3.3	2
CTQSSEFCDVIDPDCCSGVCMAFFCI	Vc6.40	2
CIPFLHPCTFFFPDCCNSICAQFICL	VcVIC	2
CTVNGVVCDPGNHNCCSGSCLDDED-TPVCGIHVEIQHVHMLS	Pu6.23	2
CCDDSECDYSCWPCCMF	Gm3-WP04	2
DAINVAPGTSITRTETDQECIDTCK-QEDKKCCGRSNGVPTCAKICL	Di6.11	2
CLAPQRWCSMHDDSLHDDNCCKTCI-ILWCS	Pu6.20	2
CIVGTPCHVCRSQSKSCNGWLGKQR-YCGYC	Im9.11	2
CNNRGGGCSQHPHCCSGTCNKTFGVCL	VxVIA, MgJ42	2

CCDRPCSIGCVPCCLP	Ca3-VP01, Cp3-VP05	2
YWTECCGRIGPHCSRCICPGVVCPKR	Bu25	2
WFGHEECTYWLGPCEVDDTCCSASCESKFCGLW	RVIIA	2
QCEDVWMPCTSSHWECCSLDCEMYCTQI	Mr6.29	2
QCPYCVVHCCPPSYCQASGCRPP	Vc7.4	2
QGCCNVPNGCSGRWCRDHAQCC	MIIIA	2
TCSSSSDCPTGQECCPDKLDEPEGSCANECIIT	Pu6.37	2
SCSDDWQYCEYPHDCCSWSCDVVCS	Vc6.12	2
TCNTPTRYCTLHRHCCSLHCHKTIHACA	Pu6.30	2
TTSTRKCKGPLVFCPENHECCSKFCDFIDIPLRYCSTP	Br7.9	2
MTKHCTPPEVGCLFAYECCSKICWRPRCYPS	ABVIE	2
VCCPFGGCHELCLCCD	MrIIIF	2
RCCISPACHDDCICCIT	S3-I05	2
RCCISPACHEECYCCQ	S3-Y01	2
VSIWFCASRTCSTPADCNPCTCESGVCVDWL	Lt9a variant 2	2
QCLPPLSLCTMDDDECCDDCILFLCLVTS	Ar6.5	2
STDDCSTAGCKNVPCCEGLVCTGPSQGPVCQPLA	Vn6.18	2
GCCDPQWCDAGCYDGCC	Qc3-YDG01	2
GCWLCLGPNACCRGSVCHDYCPS	Cal6.4c	2
GCSDFGSDCVPATHNCCSGECFGFEDFGLCT	Pu6.25	2
STDCNGVPCQFGCCVTINGNDECRELDC	Mr6.23	2
RCCTWQECDGNCHCCQ	Cp3-H02	2
RCCVHPACHDDCICCIT	Bt3-I03, Vx3-I03	2
WWGENDCSWTGPCTVNAECCLGVCDETC	Tx7.31	2
GCCHPSTCHVRKGCSRCCS	Tx3g, Vt3-SR01	2
SSDEECVGLSGYCGPWNNPPCCSWWECEVYCAVPGPSF	Mi034	2
SCCNAGFCRFGCTPCCY	Tx3e, Vt3-TP01, Ec3-TP01-2	2
TCDPYYCNDGKVCCPEYPTCGDSTGKLICVRVTD	Im6.7	1
TCLEIGEFCGKPMMVGSLCCSPGWCFFICVG	Pc6b	1
CGGYSTYCEVDSECCSDNCVRSYCTLF	TxVIIA	1
GCCCNPACGPNYGCGTSCSRPSEP	S1.7	1
TRGCKSKGSFCWNGIECCGGNCFFACVY	Cl6.6b	1
CFESWVACESPKRCCSHVCLFVCT	Pn6.6	1
WREGSCTSWLATCTDASQCCTGVCYKRAYCALWE	TxMEKL-022/TxMEKL-021	1

YCSDSGGWCGLDPELCCNSSCFVLC	Cl6.8	1
YCSDDWQPCSHFYDCCKWSCNNGYCP	Vc6.25	1
CCDDSECSYSCWPCCY	TxMMSK-02, Cp3-WP03, -Vr3-WP04, S3-WP01, Rt3-WP01	1
WRVDSECISFWGSCTVDADCCFNSC-DETYGYC	Tx7.30	1
CCDWPCTIGCVPCCLP	TsMMSK-021	1
CCFWPMCRGCDCCYL	Lv3-D02	1
CCGPTACLAGCKPCCY	Tx3-KP03	1
CESYGKPCGIYNDCCNACDPAKKTCT	Conotoxin-3	1
VQPSECKLPAAKGPCKGKYRKVYFN-NFKKQCRMFTYGGCGGNGNKFRNA-KECYHKCAYGV	conkunitzin-G1	1
VCCSFGSCDSLCQCCD	Mr3.16	1
CCLWPECGGCVCCYL	Lv3-V02	1
TRGCKTKGTWCWASRECCLKDCLFV-CVY	Cl6.10	1
CCSVSICQSPPVCECCA	S3-E03	1
CCVVCNAGCSGNCCS	Ts3-SGN01	1
SCSGSGYGCKNTPCCAGLTCRGPRQ-GPICL	Vn6.16	1
RCCIWPECGSCVCCL	Cp3-V08	1
SCGNLHEMCNYHLPCCRPWRCRASR-TGTRCLNKPRYRPV	Pu6.13	1
RDCRPVGQYCGIPYEHNWRCCSQLC-AIICVS	PuIA	1
GCCGSFACRFGCVPCCV	MrIIIA	1
GCCHLLACRMGCTPCCW	Tx3-TP01	1
GCCIEPLCYQYDCDCCRYL	Cp3-D03	1
ECSSPDESCTYHYNCCQLYCNKEEN-VCLENSPEV	LtVIB	1
ECRGYNAPCSAGAPCCSWWTCSTQT-SRCF	Vc6.10	1
GCCPIGPCMQSVCSPCCP	Vr3-SP01	1
GMWGKCKDGLTTCLAPSECCSGNCE-QNCKMW	TxMEKL-011, LeD51	1
GVWSECSDWLAGCSSPSECCSEKCD-TFCRLW	G6.8	1
GWDTPAPCRYCQWNGPQCCVYYCSS-CNYEEAREEGHYVSSHLLERQ	Cal6.3a	1
DECCEPQWCDGACDCCS	LtIIIA	1
KFILHALGQWQCCTMQWCDKACYCCE	Vc3.4	1
DDCTTYCYGVHCCPPAFKCAASPSC-KQT	Cal6.5a	1
KTCQRRWDFCPGSLVGVITCCGGLI-CFLFFCV	Om6.6	1
LCPDYTEPCSHAHECCSWNCYNGHC-TG	Gla(3)-TxVI	1
MQGKISSEQHPMFDPIEGCCTQSCT-TCFPCCLI	Lt3.6	1
DCCSMSACVPPPACECC	Mi3-E04	1

DCCPLPACPFGCNPCCGWPALLSGP-HQVMNNE	Mr020	1
DCCGVKLEMCHPCLCDNSCKNYGK	PIVE	1
DAMQKSKGSGSCAYISEPCDILPCC-PGLKCNEDFVPICL	LtVIA	1
NPKLSKLTKTCDPPGDSCSRWYNHC-CSKLCTSRNSGPTCSRP	LiCr95	1
QCADLGEECYTRFCCPGLRCKDLQV-PTCLLA	Ar6.10	1
QCCDSNSCEYPKCLCCN	Tx3-L02, Vr3-L01, Vt3-L01-, S3-L02	1
CVEDGDFCGPGYEECCSGFCLYVCI	Pu6.2	1
QKCCGKGMTCPRYFRDNFICGCC	CnIIIG	1
QQCCPPVACNMGCEPCC	TxMMSK-04, Vt3-EP01	1
RCCGEGASCPVYSRDRLICSCC	CnIIIE	1
RCCISPACNDTCYCCQD	Vr3-Y02, Vt3-Y01, Ts3-Y01	1
CPNTGELCDVVEQNCCYTYCFIVVCPI	Mr6.2	1
RCCTGKKGSCSGRACKNLKCCA	SxIIIA	1
APWTVVTATTNCCGITGPGCLPCRC-TQTC	A4.4	1

A Supplemental Information

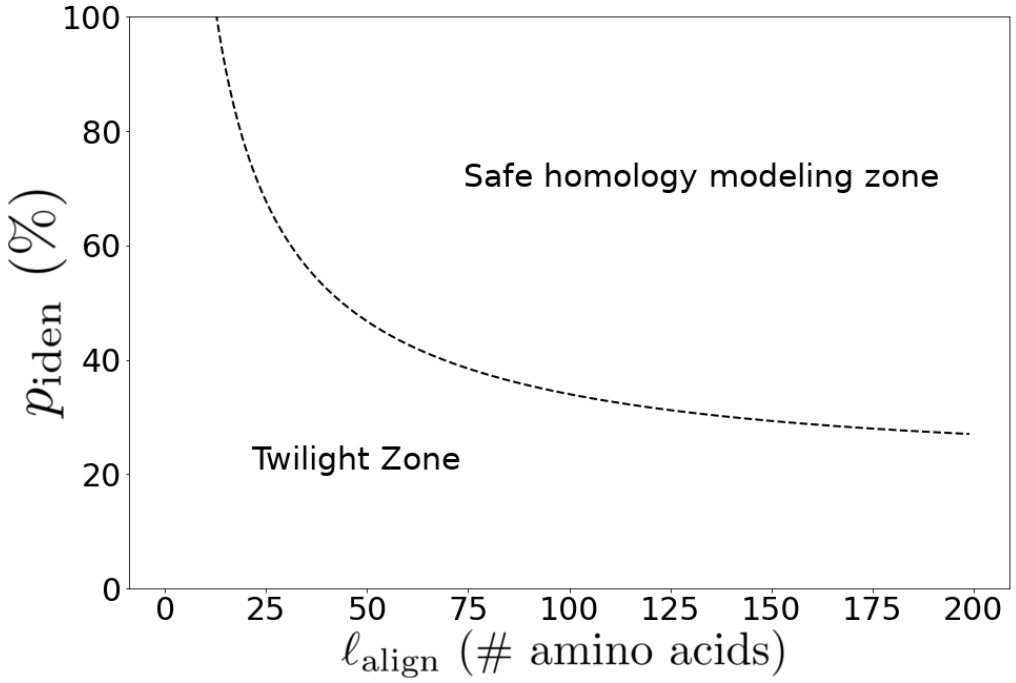

Figure S1: Rost's phenomenological curve (Eqn. 1) of minimum percentage identity for homology modeling as a function of pairwise alignment length with $n = 5\%$ padding as employed in this work. As the length of the alignment decreases, the minimum percent identity for homology modeling increases, and there is a particularly rapid increase below alignments of about 25 amino acids, where a fairly large proportion of toxins reside.

Table S1: List of sequences containing eight cysteines in order of interest for experimental characterization, based on degree (sequence coverage) in alignment graphs (cf. Fig. 2). Name or names of sequences are taken from the Conoserver database [Kaas et al., 2012]. Multiple names for the same sequence indicate the same sequence is produced by different species. Node degree corresponds to the number of sequences with pairwise alignments that are long enough and have high enough percent identity to be homology modeled with the given sequence as a template.

Sequence	**Name(s)**	**Degree (# Edges)**
TDVCKKSPGKCIHNGCFCEQDKPQG-NCCDSGGCTVKWWCPGTKGD	Cal12.1p2	28
GHVPCGKDGRKCGYHADCCNCCLSG-ICKPSTSWTGCSTSTVQLTR	R11.10	18
QCTPKNQICEEDGECCPNLECKCFT-RPDCQSGYKCRP	Vr15b	14
CFPPGVYCTRHLPCCRGRCCSGWCR-PRCFPRY	Cp1.1	10

QCTQQGYGCDETEECCSNLSCKCSG-SPLCTSSYCRP	Cap15a	9
SCDSEFSSEFCEQPEERICSCSTHV-CCHLSSSKRDQCMTWNRCLSAQTGN	Gla-MrII, Eu12.4	9
SRCFPPGIYCTPYLPCCWGICCGTC-RNVCHLRF	Em11.8	8
DKWGTCSLLGKGCRHHSDCCWDLCC-TGKTCVMTVLPCLFLSLIVRWT	Mr11.1	6
TCSLPGDGCIRDFHCCGHMCCQGNK-CVVTVRRCFNFPY	Pu11.5	6
YDAPYCSQEEVRECQDDCSGNAVRD-SCLCAYDPAGSPACECRCVEPW	Cal22d	5
GTCSGRGQECKHDSDCCGHLCCAGI-TCQFTYIPCK	Tx11.3	5
GTCSYLGEGCKRDSDCCGHFCCGGK-TCVITARPCKV	Vc11.4	5
RGVCSTPEGSCVHNGCICQNAPCCH-PSGCNWANVCPGYLWDKN	Cal12.2c	4
TCSDLGQACVHESDCCAQMCCLNKK-CAMTMPPCNFY	Vc11.1	3
CLSEGSPCSMSGSCCHKSCCRSTCT-FPCLIP	Ep11.12	2
TCSNKGQQCGDDSDCCWHLCCVNNK-CAHLILLCNL	M11.2	2
RCSDDTGATCSNRFDCCESMCCIGG-HCVISTVGCP	Im11.14	1
CRLEGSSCRRSYQCCHKSCCIRECK-FPCRWV	Vi11.5	1
TRSFADLPDDWGMCSDIGEGCGQDY-DCCGDMCCDGQICAMTFMACMF	Vc11.6	1
CLRDGQSCGYDSDCCRYSCCWGYCD-LTCLIN	Im11.1	1
CNGRGEWCSTHRSCCDSGDVCCITT-PVGPICTRGCSGRIIPQRRGAQLRHFF	Pu11.9	1
CRAEGTYCENDSQCCLNECCWGGCG-HPCRHP	BtX, Sx11.2	1
CTSEGYSCSSDSNCCKNVCCWNVCE-SHCRHPGKR	Lt11.3	1
CRSGKTCPRVGPDVCCERSDCFCKL-VPARPFWRYRCICL	Mr15.2	1
DCPTSCPTTCANGWECCKGYPCVRQ-HCSGCNH	De13b	1
EGGYVREDCGSDCMPCGGECCCEPN-SCIDGTCHHESSPN	Mi045	1
SCRNEGAMCSFGFQCCKKKCCMSHC-TDFCRNP	Vt11.3	1
WPRLYDSDCVRGRNMHITCFKDQTC-GLTVKRNGRLNCSLTCSCRRGESC-LHGEYIDWDSRGLKVHICPKPWF	Mr22.1	1
MCLSLGQRCGRHSNCCGYLCCFYDK-CVVTAIGCGHY	Bt11.4	1
ASICYGTGGRCTKDKHCCGWLCCGG-PSVGCVVSVAPC	Ca11.3	1

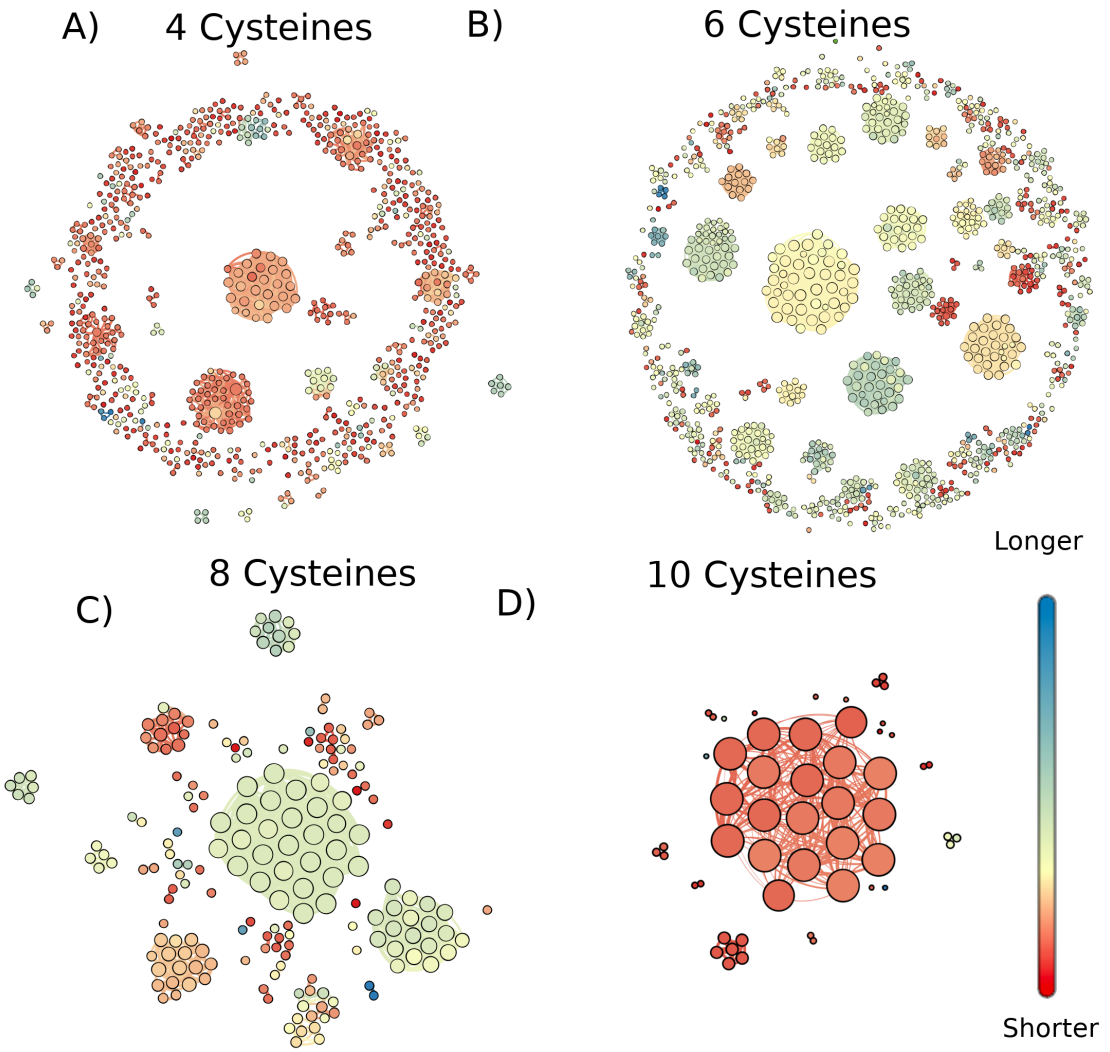

Figure S2: Graph of conotoxins containing (A) four cysteines, (B) six cysteines, (C) eight cysteines and (D) ten cysteines where nodes are sequences and edges exist between sequences with pairwise alignments that have high enough length and percent identity to fall above the Rost curve with $n = 5\%$ (Eqn. 1). Colors show the relative sequence lengths of each graph, but the color scale of each graph is independent of the others. The sizes of the nodes corresponds to their degree; that is the number of other sequences that they can be modeled based on or used to model. Node locations and edge lengths were chosen for ease of visualization of separate connectec components. Visualization of the graphs was produced with Gephi 0.9.2 [Bastian et al., 2009].

Table S2: List of sequences containing ten cysteines in order of interest for experimental characterization, based on degree (sequence coverage) in alignment graphs (cf. Fig. 2) Name of sequences are taken from the Conoserver database [Kaas et al., 2012]. Node degree corresponds to the number of sequences with pairwise alignments that are long enough and have high enough percent identity to be homology modeled with the given sequence as a template.

Sequence	**Name(s)**	**Degree (# Edges)**
DRDVQDCQVSTPGSKWGRCCLNRVC-GPMCCPASHCYCVYHRGRGHGCSC	Cp20.1	19
LHCYEISDLTPWILCSPEPLCGGKG-CCAQEVCDCSGPACTCPPCL	Lt15.6	5
YNRQCCIDKTYDCLKKYRGRENTFA-SVCQQEAAVYCGAWDEAEGCCYGY-SHCMSMYAQQSGLDVAHNGCKDRK-CDNP	Vc21.1	2
QCTLVNNCDRNGERACNGDCSCEGQ-ICKCGYRVSPGKSGCACTCRNA	Ac8.1	2
GCSGTCRRHRDGKCRGTCECSGYSY-CRCGDAHHFYRGCTCTC	Ca8c	2
TCDPTPDCRTTVCETDTGPCCCPHG-YNCQTTNSGRRACVLVCPHNCPP	Pu19.1	1
SGSTCTCFTSTNCQGSCECLSPPGC-YCSNNGIRQRGCSCTCPGT	G8.3	1
GCTRTCGGPKCTGTCTCTNSSKCGC-RYNVHPSGWGCGCACS	GVIIIA	1
GCTISCGYEDNRCQGECHCPGKTNC-YCTSGHHNKGCGCAC	Tx8.1	1

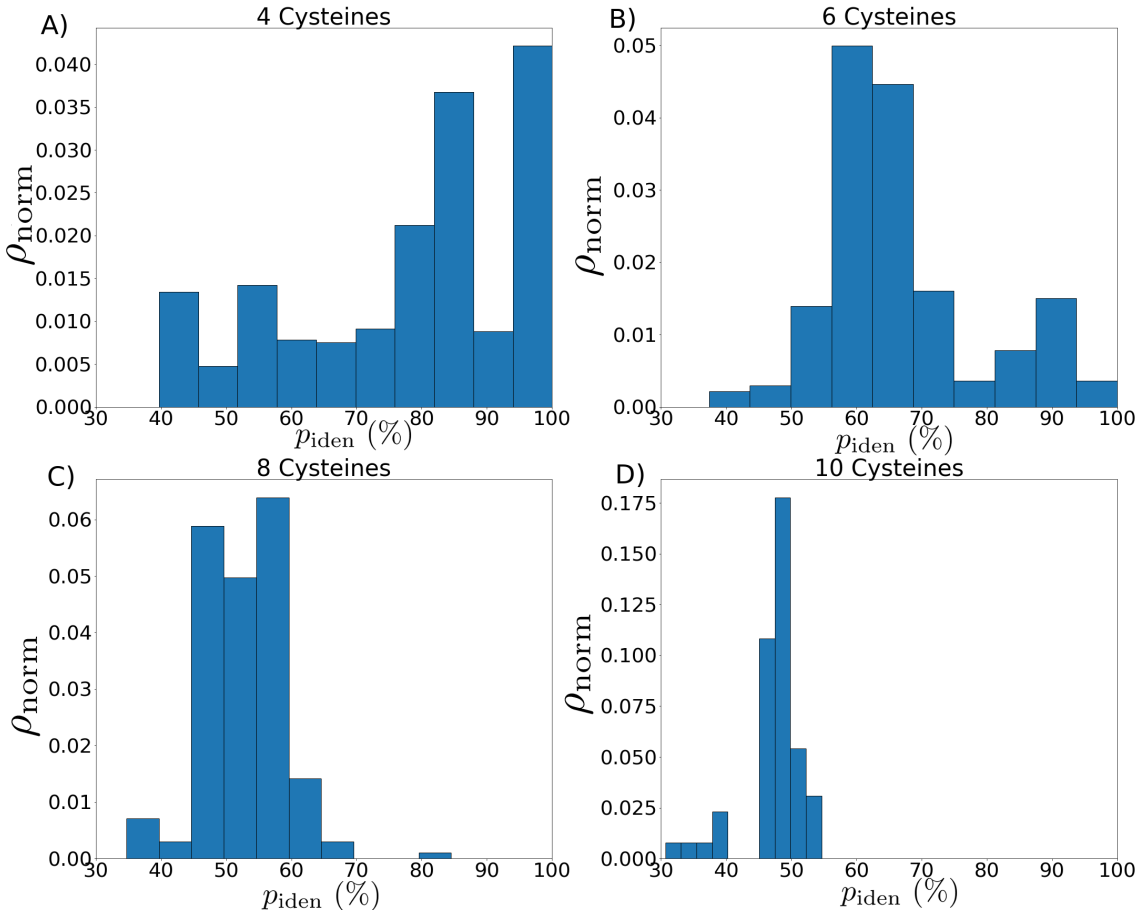

Figure S3: Distributions of approximate minimum percentage identity cutoffs for the conotoxins containing A) four, B) six, C) eight, and D) ten cysteines. We assume for demonstration purposes that any alignment is the length of the peptide itself. We employ Rost's curve (see Eqn. 1 and Fig. S1) with a padding of $n = 5\%$. The distribution shifts significantly downward as the number of cysteines and concomitantly the overall length of the peptides under consideration increases. Note too the presence in panels A) and B) of a bin going up to 100%, which demonstrates the existence of peptides so short among the conotoxins that it is impossible to reliably predict their structure via homology modeling, which comprise a large proportion of the isolated nodes in the graphs (cf. Fig. S2)

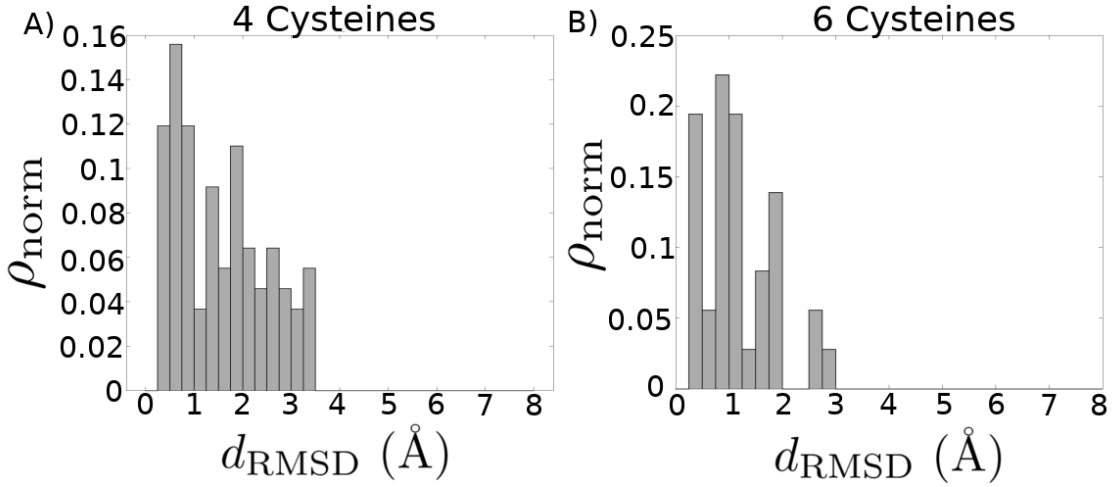

Figure S4: Distribution of root-mean-square deviation (RMSD) for homology models compared with their corresponding experimental structures, after refinement involving rejection of structural alignment outliers. Each experimental structure present in the library was modeled by selecting from all other templates in the library. The top three models for each structure based on combined MODELLER DOPE and PROCHECK G-FACTOR scores are considered here. (A) Distribution mean = 1.55Å, standard deviation = 0.92Å. (B) Distribution mean = 1.17Å, standard deviation = 0.67Å.