Screening, large-scale production and structure-based classification of cystine-dense peptides

Colin E. Correnti1,5, Mesfin M. Gewe1,5, Christopher Mehlin1, Ashok D. Bandaranayake1, William A. Johnsen1, Peter B. Rupert2, Mi-Youn Brusnika1, Midori Clarke1, Skyler E. Burke1, Willem De Van Der Schueren1, Kristina Pilat1, Shanon M. Turnbaugh1, Damon May1,4, Alex Watson1, Man Kid Chan1, Christopher D. Bahl2,3, James M. Olson1* and Roland K. Strong2*.

Peptides folded through interwoven disulfides display extreme biochemical properties and unique medicinal potential. However, their exploitation has been hampered by the limited amounts isolatable from natural sources and the expense of chemical synthesis. We developed reliable biological methods for high-throughput expression, screening and large-scale production of these peptides: 46 were successfully produced in multimilligram quantities, and >600 were deemed expressible through stringent screening criteria. Many showed extreme resistance to temperature, proteolysis and/or reduction, and all displayed inhibitory activity against at least 1 of 20 ion channels tested, thus confirming their biological functionality. Crystal structures of 12 confirmed proper cystine topology and the utility of crystallography to study these molecules but also highlighted the need for rational classification. Previous categorization attempts have focused on limited subsets featuring distinct motifs. Here we present a global definition, classification and analysis of >700 structures of cystine-dense peptides, providing a unifying framework for these molecules.

Peptides are differentiated from proteins on the basis of their smaller size of generally fewer than ~50 residues. Because of their size, peptides cannot form sufficient cooperative interactions and thus usually do not adopt the stable, defined structures achieved in proteins through well-packed hydrophobic cores. Exceptions include peptides that alternately organize around cores of tightly packed disulfides (Fig. 1a), which often confer extreme thermal, chemical and proteolytic stability1–4. These peptides can also include analgesic, antihelminthic, antimicrobial, antitumor, and proteolytic stability1–4. These peptides can also include analgesic, antihelminthic, antimicrobial, antitumor, and proteolytic stability1–4. Archetypes of such peptides, with cores of at least three cystines, include ‘inhibitor cystine knot’ peptides, or knottins, and the closely related ‘cyclic cystine knot’ peptides, or cyclotides5–8. Examples include venom toxins from cone snails, spiders and scorpions; plant protease inhibitors; and antimicrobial defenses. Knottins and cyclotides are topologically pseudo-knotted, with one cystine (the ‘knotting’ cystine) crossing through the macrocycle formed by the other two cystines (the ‘bracketing’ cystines) and the interconnecting backbone, which often contains additional accessory cystines (Fig. 1b). Proteins can also be pseudo-knotted through intrachain cystines, e.g., ‘growth-factor cystine knot’ (GFCKs; Fig. 1c). However, GFCK intrachain cystines do not dominate the fold of the protein, which includes a conventional hydrophobic core distinct from the structures of knottins and cyclotides.

Natural knottins and cyclotides have demonstrated many properties beyond stability, which are useful in clinical applications; these properties include the potential for oral delivery, cell penetration and tumor homing3–6. Inherent pharmacological properties can also include analgesic, antihelminthic, antimicrobial, antitumor, insecticidal or ion-channel-modulatory activities1. These peptides are intermediate in size between protein biologics and conventional small-molecule drugs; hence, they are potentially small enough to penetrate a variety of tissues and solid tumors but large enough to enable protein-like ligand specificity and affinity. Approved drugs exploiting these properties include ziconotide (Prialt), based on a cone-snail venom knottin; linacotide (Linzess), based on Escherichia coli heat-stable enterotoxin; and plecanatide (Trulance), based on human uroguanylin10.

Overall, the minimal common elements defining this class of molecules are short sequences constituting independent folding domains and having a high density of cystines (i.e., at least three). We refer to this categorization as ‘cystine-dense peptides’ (CDPs), drawing a distinction between CDPs and larger proteins with cystine-knotted elements, such as GFCKs. Inspired by these peptides’ unique folds and their potential for clinical application—as well as the success of ziconotide, linacotide, plecanatide and tozuleristine (Tumor Paint)11, a tumor-homing, fluorophore conjugate of the scorpion knottin Chlorotoxin (CTX)—we sought to more fully explore and exploit these molecules. However, although these peptides are generally amenable to chemical synthesis, and some have been expressed recombinantly, a major impediment to exploiting this unique pharmacomolecular space has been the lack of a reliable high-throughput expression platform. Here we report biologic production of over 700 CDPs and exhaustive biochemical characterization of 46 CDPs demonstrating unique properties. We optimized crystallographic methods for elucidating CDP structures. Structural analyses highlighted another impediment to understanding this fold space: the absence of a unified, global scheme for classifying CDPs. Because there are many examples of structured cystine-rich peptides that are not knottins or cyclotides, or even knotted, we propose a structure-based classification scheme unifying and framing an analysis of a complete catalog of available CDP structures.
Results

CDP sequence-based definition. Inspection of the Protein Data Bank (PDB)\(^{12}\) and homologous sequences from natural sources suggested a CDP-defining motif comprising: (i) six or more cysteine residues in a span from 13 to 81 residues long and not recognizable as a cytoplasmic protein or domain, a zinc-finger protein or a GFCK; and (ii) a constrained distribution of cysteines, Cys-\(X_{13-15}\)-Cys-\(X_{13-15}\)-Cys-\(X_{13-15}\)-Cys-\(X_{13-15}\)-Cys (where \(X\) represents any amino acid). To confirm formation of cystines, the candidate CDP should be embedded in a sequence with a recognizable leader peptide, e.g., by using SignalP\(^{13}\); annotated as a secreted or integral membrane protein; or experimentally shown to contain specific cystines. CDPs may be embedded in larger proteins or in tandem arrays (some examples have more than 40 (refs \(^{14,15}\)) but should comprise an independent folding unit.

Applying these rules to the PDB yielded a continuum of structures, though adding the criterion of a minimal ‘cysteine density’, with at least 12% cysteine content in the span including the bounding cystines, satisfactorily separated CDPs from small proteins with emergent hydrophobic cores. This threshold density is approximately ten-fold higher than the average for all proteins\(^{16,17}\). In the PDB, as of April 2017, there were 775 experimentally determined structures with domains conforming to this motif, excluding wholly synthetic, designed sequences and including 422 knotted CDPs, 203 nonknotted CDPs with three cystines and 150 nonknotted CDPs with more than three cystines (Supplementary Table 1).

CDP biologic production. To develop a biologic expression platform for CDPs, we selected a target set of 100 proteins (Supplementary Table 2). We concentrated on CDPs similar or

Fig. 1 | CDPs versus GFCKs. a–c. Crystal structures of the archetypical scorpion-venom knottin Chlorotoxin (CTX) (a), a distinct knotted peptide, \(\alpha\)-KTx 3.10 (BoiTxl) (b) and a nonknotted cysteine-containing peptide, Elafin (c), all determined as part of this work. The structures, shown in backbone representation, with cysteine side chain sulfur atoms shown as spheres, demonstrate the degree to which cystines dominate the cores of these structured peptides. The backbone and cysteine side chain atoms are colored from blue to red from the N to C terminus, highlighting the cystine connectivity and pseudoknot topology characteristic of knottins and related peptides. The topology is represented as \(u-v, [w-x], y-z\), where \(w-x\) is the knotting cystine, and \(u-v\) and \(y-z\) are the bracketing cystines. CTX, like many knotted CDPs, has an accessory cystine, circled in red, in addition to the three core cystines defining the pseudoknot element. d. A typical dimeric GFCK, human nerve growth factor (NGF), shown in cartoon representation, with one monomer colored gray and the other monomer colored as in a.

Fig. 2 | CDP biochemical and crystallographic analyses. a. Example comparative RPC analyses of two scorpion potassium-channel toxins, CDP #11 (left) and #17 (right), produced with the Daedalus-based biologic production platform, under nonreducing conditions (blue), reducing conditions (red; through addition of 100 mM Tris(2-carboxyethyl)phosphine (TCEP)) and thiol-blocking treatment after reduction (dashed; through addition of 40 mM iodoacetamide (IAA) after addition of TCEP), mAU, milliabsorbance units. b. Normalized CD spectra from all 46 successfully expressed CDPs, showing a wide range of secondary-structure compositions. c. Superposition of two structures of CTX (target #28), shown as licorice-stick representations of the peptide backbone plus cysteine side chains, colored by atom type, determined by NMR (PDB 1CHL\(^{15}\), periwinkle carbons) or by crystallography (reported here, gray carbons). d. Superposition of all 18 independent views of the structure of Elafin (target #4, a human class 1–3, 2–5, 4–6 nonknotted CDP) in the crystal structure reported here, shown as a cartoon ribbon representation of the peptide backbone plus licorice-stick representation of all side chains, colored by atom type. The structure of Elafin was previously determined by crystallography (PDB 1FLE\(^{14}\)), though only in complex with elastase, and alone by NMR (PDB 2REL\(^{16}\)). e. Superposition of all five independent views of the tetramer of the potassium-channel toxin \(\gamma\)-KTx 2.2 hitchin from the venom of the Manchurian scorpion (target #48), from two different crystal forms, shown in cartoon representation and colored by tetramer.
identical to previously studied CDPs, largely spider- and scorpion-venom components, to validate success; however, we also included some simpler two-cystine, non-CDP sequences (e.g., Hefutoxin) to phase in the magnitude of the challenge. We abandoned initial attempts to produce CDPs in bacterial expression systems in favor of mammalian secretion-pathway-based systems, which incorporate folding chaperones and extensive quality-control machinery and thus dramatically improved success rates. CDPs produced in mammalian cells were also free of contaminating endotoxins, thereby streamlining in vivo applications. CDPs were ultimately most successfully produced through stringently produced peptides. We chose to apply the most rigorous characterization. RPC is extremely sensitive to any chemical heterogeneity in a preparation, and it can also detect the presence of chemically identical but conformationally distinct isoforms. Expression of a CDP was considered successful only if the final RPC traces showed identically identical but conformationally distinct isoforms. Expression of a CDP was considered successful only if the final RPC traces showed

Because isolated CDPs displayed anomalous behavior in reduced/nonreduced comparative PAGE and size-exclusion chromatography, reverse-phase chromatography (RPC) was used for preparative purification after cleavage, final purity assessment and biochemical characterization. RPC is extremely sensitive to any chemical heterogeneity in a preparation, and it can also detect the presence of chemically identical but conformationally distinct isoforms. Expression of a CDP was considered successful only if the final RPC traces showed only a single dominant peak under both oxidizing and reducing conditions, thus indicating a single folding state, and the absence of proteolysis or heterogeneity of any kind, including conformational heterogeneity (Fig. 2a and Supplementary Dataset 1). Prior studies of CDP synthesis (e.g., ref. 21) have yielded preparations that were biologically functional but displayed complex RPC traces, which have been suggested to indicate the presence of multiple conformers that do not interchange on the RPC timescale. Distinct metastable conformers of a CDP may not, for instance, all interact equivalently with a target ion channel, thus complicating analyses of less stringent ly produced peptides. We chose to apply the most rigorous

Identical to previously studied CDPs, largely spider- and scorpion-venom components, to validate success; however, we also included some simpler two-cystine, non-CDP sequences (e.g., Hefutoxin) to phase in the magnitude of the challenge. We abandoned initial attempts to produce CDPs in bacterial expression systems in favor of mammalian secretion-pathway-based systems, which incorporate folding chaperones and extensive quality-control machinery and thus dramatically improved success rates. CDPs produced in mammalian cells were also free of contaminating endotoxins, thereby streamlining in vivo applications. CDPs were ultimately most successfully produced through stringently produced peptides. We chose to apply the most rigorous characterization. RPC is extremely sensitive to any chemical heterogeneity in a preparation, and it can also detect the presence of chemically identical but conformationally distinct isoforms. Expression of a CDP was considered successful only if the final RPC traces showed only a single dominant peak under both oxidizing and reducing conditions, thus indicating a single folding state, and the absence of proteolysis or heterogeneity of any kind, including conformational heterogeneity (Fig. 2a and Supplementary Dataset 1). Prior studies of CDP synthesis (e.g., ref. 21) have yielded preparations that were biologically functional but displayed complex RPC traces, which have been suggested to indicate the presence of multiple conformers that do not interchange on the RPC timescale. Distinct metastable conformers of a CDP may not, for instance, all interact equivalently with a target ion channel, thus complicating analyses of less stringent ly produced peptides. We chose to apply the most rigorous

### Table 1 | Data collection and refinement statistics (molecular replacement)

| Target #4 (PDB 6ATU) | Target #28 (PDB 6ATW) | Target #34 (PDB 6AVA) | Target #48, form #1 (PDB 6AUP) | Target #48, form #2 (PDB 6AUY) | Target #49 (PDB 6AVC) | Target #83 (PDB 6AV8) |
|----------------------|----------------------|----------------------|-------------------------------|-------------------------------|----------------------|----------------------|
| **Data collection**  |                      |                      |                               |                               |                      |                      |
| Space group          | P4_1                 | P2_2,2_1             | P4_2,2                         | P2_2,2_1                      | C21                  | C3_21                |
| Cell dimensions      | a, b, c (Å)          | 71.3, 71.3, 214.4   | 22.5, 26.3, 48.0               | 41.4, 41.4, 88.1              | 58.9, 80.4, 94.2     | 50.2, 48.1, 50.3     |
|                      | α, β, γ (°)          | 90.0, 90.0, 90.0     | 90.0, 90.0, 90.0               | 90.0, 90.0, 90.0              | 90.0, 90.0, 90.0     | 90.0, 90.0, 90.0     |
| Resolution (Å)       | 50.00–2.40 (2.49–2.40)| 50.00–1.53 (1.56–1.53)| 50.00–2.20 (2.28–2.20)        | 50.00–1.95 (1.98–1.95)        | 50.00–1.90 (1.93–1.90)| 50.00–1.89 (1.95–1.88)|
| Rmerge               | 0.72 (0.49)          | 0.11 (0.13)          | 0.16 (0.37)                    | 0.16 (0.64)                  | 0.07 (0.23)          | 0.10 (0.41)          |
| Rfree                | 0.08 (0.53)          | 0.10 (0.13)          | 0.16 (0.37)                    | 0.16 (0.65)                  | 0.08 (0.21)          | 0.10 (0.42)          |
| Completeness (%)     | 99 (97)              | 87 (11)              | 100 (100)                      | 98 (80)                      | 83 (17)              | 85 (38)              |
| Redundancy           | 7.2 (6.8)            | 13.1 (2.3)           | 58.4 (59.5)                    | 48.3 (23.7)                  | 3.2 (1.5)            | 57.7 (40.4)          |
| **Refinement**       |                      |                      |                               |                               |                      |                      |
| Resolution (Å)       | 50.00–2.40 (2.49–2.40)| 50.00–1.53 (1.56–1.53)| 50.00–2.20 (2.28–2.20)        | 50.00–1.95 (1.98–1.95)        | 50.00–1.90 (1.93–1.90)| 50.00–1.89 (1.95–1.88)|
| No. reflections      | 39,448 (2,852)       | 3,857 (52)           | 4,048 (293)                    | 32,452 (2,732)               | 7,568 (195)          | 2,574 (110)          |
| Rmerge / Rfree       | 0.20 / 0.25          | 0.13 / 0.18          | 0.17 / 0.24                    | 0.19 / 0.24                  | 0.20 / 0.22          | 0.20 / 0.26          |
| No. atoms            | Protein 6,304        | 289                  | 556                            | 4,561                        | 1147                 | 272                  |
|                      | Sulfate              | —                    | —                              | 100                          | 29                   | —                    |
|                      | Glycerol             | —                    | —                              | 24                           | 3                    | —                    |
|                      | Water                | 312                  | 55                             | 77                           | 120                  | 43                   |
|                      | Protein              | 4.354                | 1.84                            | 15.84                        | 23.07                | 15.95                |
|                      | Sulfate              | —                    | —                              | 26.05                        | 44.74                | —                    |
|                      | Glycerol             | —                    | —                              | 24                           | 3                    | —                    |
|                      | Water                | 36.8                 | 28.1                           | 24.28                        | 13.36                | 24.27                |
| r.m.s. deviations    | Bond lengths (Å)     | 0.013                | 0.016                          | 0.009                        | 0.018                | 0.014                |
|                      | Bond angles (°)      | 1.12                 | 1.54                           | 1.33                         | 1.97                 | 1.62                 |

Values in parentheses are for highest-resolution shell.

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.
threshold for success, requiring absolute homogeneity. Of the 100 targets, 46 were successfully produced under these stringent criteria (Supplementary Dataset 1). For example, CDP #11 showed near-complete resistance to reduction, but CDP #17 was a doublet under nonreducing conditions and hence did not meet our quality-control criteria (Fig. 2a). This system is amenable to high-throughput, pilot-scale expression screening on automated robotic platforms as well as affordable preparative-scale production, with final yields of up to 10 mg/L at a 2-L culture scale. The high-throughput platform processed hundreds of CDPs per week, yielding up to 20 μg at the 1-mL culture scale. These amounts were sufficient to evaluate proper folding through analytical RPC by applying the same stringent criteria as above. We determined that another 678 CDPs (55%) out of an additional list of 1,232 targets were expressible, by using the high-throughput approach (Supplementary Table 3). These additional targets more broadly sampled CDPs from plants, arthropods and other taxa; 872 had less than 75% identity to a CDP in the PDB.

CDP biochemical and structural characterization. To further explore the previously reported extreme biochemical stability of CDPs4,6,7, we tested the 46 successfully produced CDPs under a battery of conditions that would be expected to denature or degrade conventional globular proteins, including extended incubation under reducing conditions or at high temperatures, and proteolytic digestion (Supplementary Dataset 1). Most of the 46 CDPs showed resistance to some combination of these conditions, but several showed truly exceptional stability, including two knotted CDPs that were completely resistant to all conditions tested (Supplementary Dataset 1). CD spectra were collected from the 46 successfully produced CDPs to evaluate secondary-structure content (Fig. 2b and Supplementary Dataset 1). The range observed included spectra showing very limited secondary-structure content, in agreement with known homologous structures.

The ultimate confirmation of proper cystine formation in a CDP is experimental determination of its three-dimensional structure. Most isolated CDP structures available in the PDB were determined by NMR, partly because of the perception that CDPs are inherently difficult to crystallize. Crystallography has previously been used primarily, but very effectively, to determine structures of complexes between CDPs and binding-partner proteins. However, access to an efficient expression system allowed for large-scale production of highly purified CDPs that could be highly concentrated (to ≥80 mg/mL), thereby enhancing crystallizability. Of the 46 successfully expressed CDPs, 12 were crystallized (26% success rate, one in two crystal forms) and used to determine structures (Tables 1 and 2, Supplementary Table 4 and Table 2 | Data collection and refinement statistics (sSAD)

| Target #19 (PDB 6ATS) | Target #56 (PDB 6AVD) | Target #60 (PDB 6ATN) | Target #63 (PDB 6ATL) |
|-----------------------|-----------------------|-----------------------|-----------------------|
| **Data collection**    |                       |                       |                       |
| Space group           | C121                  | P121                  | P41,2,2                | P12,1                 |
| Cell dimensions       |                       |                       |                       |                       |
| a, b, c (Å)           | 46.1, 21.3, 20.4      | 21.5, 26.2, 31.1      | 46.6, 46.6, 44.4      | 27.7, 23.2, 46.3     |
| α, β, γ (°)           | 90.0, 92.6, 90.0      | 90.0, 99.2, 90.0      | 90.0, 90.0, 90.0      | 90.0, 94.4, 90.0     |
| Resolution (Å)        | 50.00-1.95 (1.98-1.95)| 50.00-1.80 (1.83-1.80)| 50.00-1.76 (1.79-1.76)| 50.00-1.80 (1.83-1.80)|
| Rmerge                | 0.04 (0.05)           | 0.06 (0.13)           | 0.05 (0.19)           | 0.11 (0.31)          |
| Rfree                 | 0.04 (0.04)           | 0.07 (0.13)           | 0.07 (0.26)           | 0.11 (0.32)          |
| I/σ(I)                | 41.8 (25.8)           | 14.2 (6.2)            | 25.0 (2.7)            | 35.3 (12.7)          |
| CC1/2                 | (0.99)                | (0.98)                | (0.94)                | (0.98)               |
| Completeness (%)      | 90 (45)               | 83.9 (67.9)           | 44.7 (2.2)            | 99.9 (100)           |
| Redundancy            | 2.7 (1.6)             | 1.8 (1.2)             | 2.1 (1.0)             | 11.6 (1.0)           |
| **Refinement**        |                       |                       |                       |                       |
| Resolution (Å)        | 23.03-1.90 (1.95-1.90)| 30.70-1.80 (1.84-1.80)| 32.93-1.76 (1.81-1.76)| 24.58-1.77 (1.83-1.77)|
| No. reflections       | 1,370 (51)            | 2,623 (125)           | 4,611 (158)           | 4,274 (73)           |
| Rwork / Rfree         | 0.14 / 0.16           | 0.17 / 0.23           | 0.17 / 0.20           | 0.14 / 0.18          |
| No. atoms             | Protein 197           | 289                   | 283                   | 538                  |
|                       | Ligand/ion            |                       |                       |                       |
|                       | Sulfate —             | 6                     | —                     | 15                   |
|                       | Citric acid —         | —                     | —                     | 11                   |
|                       | Water 33              | 40                    | 17                    | 67                   |
|                       | Protein 14.8          | 10                    | 25.1                  | 15.8                 |
|                       | Ligand/ion —          | 29.4                  | —                     | 15.8                 |
|                       | Water 30.1            | 26.2                  | 36.13                 | 29.4                 |
| r.m.s. deviations     | Bond lengths (Å)      | 0.014                 | 0.009                 | 0.013                | 0.016                |
|                       | Bond angles (°)       | 1.26                  | 1.52                  | 1.44                 | 1.97                 |

Values in parentheses are for highest-resolution shell.
Fig. 3 | Structure-based CDP-classification scheme. a. The 15 potential connectivities linking six cysteines pairwise are shown, arranged by the five possible pairings in the first cystine (i.e., 1–2, 1–3, and so forth, with the focus set of six cysteines numbered from 1 to 6 from the N to C terminus). Subsequent pairings in the remaining cystines are shown in descending rows. Connectivities with corresponding experimentally determined structures in the PDB are indicated in black or green; there are no natural CDPs with a 1–6, 2–3, 4–5 connectivity in the PDB (indicated in red). The 1–4, 2–5, 3–6 connectivity pattern (green) was by far the most commonly observed: 312 examples, 298核查ed and 14 nonknotted. Percentage class distributions of the 621 knotted CDPs, plus nonknotted three-cystine CDPs, are shown in parentheses. Cystine connectivity could not be determined for the subgroup of 150 nonknotted CDPs with more than three cystines. b. Outline of the overall CDP-classification scheme, showing the hierarchical relationship among cysteine density (GFCKs versus CDPs), pseudoknotting (the knotted CDP subset of CDPs) and type classifications, based on cysteine connectivity plus knotting topology. Common connectivities are grouped within boxes, thus highlighting that only 5 of the 15 possible connectivities were observed among known knotted CDP structures. Example cartoon schematics of cysteine connectivity/topology (numbered yellow circles indicate cysteines; number observed in the PDB is indicated in parentheses) are shown for the canonical shankins, hitchins and knottins, and the simplest nonknotted CDP type, (1–6, 2–5, 3–4; upper right). Only types observed in the PDB as of April 2017 are indicated, showing the nonrandom distribution of knotted-CDP types: only 12 of the possible 45 (15 connectivities ×3 pseudoknotted topologies) were observed. c. Tabulation of the class and type distribution of the 775 CDP structures, knotted (red) and nonknotted (black), highlighting the dominance of knottins and hitchins among knotted CDPs. All CDPs in the PDB and their classification are shown in Supplementary Table 1.

Supplementary Fig. 3). Initial phases were determined either by molecular replacement (MR) or sulfur single-wavelength anomalous diffraction (sSAD)\(^2\), by using copper K\(\alpha\) radiation to maximize the anomalous signal. With optimized procedures, sSAD proved to be an extremely effective approach for CDP structural analysis, providing completely unbiased initial phases and more direct, detailed information about sulfur substructure.

These new structures all showed the expected overall cystine connectivity and topology, on the basis of comparisons with previously determined homologous structures, thus validating the expression platform. However, many detailed structural differences were found when these crystal structures were compared with previous NMR structures (e.g., CTX; Fig. 2c), particularly in the arrangement of the cystine core, the key defining element of CDP structure. The disparity in structural details is also likely to explain the inability to use the as-deposited CTX NMR structure as an MR search model to phase the crystallographic data, which required computational remodeling with Rosetta\(^2\) to generate a successful search model. A number of CDP crystals yielded multiple independent views of their structures (Supplementary Table 4), thus providing additional information about structural rigidity/flexibility (Fig. 2d) and quaternary structure (Fig. 2e). In the most extreme cases, the whey acidic protein–type, four-disulfide-core peptide Elafin (target #4, a human, nonknotted CDP) crystallized with 18 copies in the asymmetric unit (AU), thus demonstrating the extraordinary structural rigidity possible in CDPs. The \(\gamma\)-KTx 2.2 potassium–channel toxin (target #48, a knotted CDP from the venom of the Manchurian scorpion) crystallized in the same tetrameric state, with 20 copies in total in the AUs of two different crystal forms. Although prior NMR structural analyses have suggested inherent flexibility or the presence of multiple conformers in solution for some CDPs (e.g., Supplementary Fig. 3), the high degree of structural conservation observed among multiple views in CDP crystal structures conversely favors the adoption of a single rigid structure for these examples.

CDP ion-channel modulation. Many natural venom and toxin CDPs have been reported to modulate ion-channel activity\(^7\), and our initial target set of CDPs included many related molecules. To assess channel activity, both to verify production of functional CDPs and to identify additional specificities/activities, the selectivity profiles of 37 successfully expressed CDPs were assessed with a commercial electrophysiological assay on a panel of 20 human ion channels (Supplementary Fig. 4). All showed some activity toward at least a subset of ion channels, predominantly potassium channels. Although many of the 37 had not been previously characterized or had been tested on a more focused set of channels, the agreement between our results and the limited prior results was very good, thus reinforcing the utility of our CDP production/characterization platform.

CDP structure-based classification. Coupling optimized CDP crystallography with robust expression platforms enables full determination of the boundaries of CDP fold space, if CDPs can be identified on the basis of sequence and classified on the basis of structure. However, we encountered several practical problems in applying our CDP-defining motif to larger databases to fully catalog candidate CDPs on the basis of sequence. To ensure cysteine formation, the CDP-defining motif to larger databases to fully catalog candidate CDPs on the basis of sequence. To ensure cysteine formation, the human, nonknotted CDP) crystallized with 18 copies in the asymmetric unit (AU), thus demonstrating the extraordinary structural rigidity possible in CDPs. The \(\gamma\)-KTx 2.2 potassium–channel toxin (target #48, a knotted CDP from the venom of the Manchurian scorpion) crystallized in the same tetrameric state, with 20 copies in total in the AUs of two different crystal forms. Although prior NMR structural analyses have suggested inherent flexibility or the presence of multiple conformers in solution for some CDPs (e.g., Supplementary Fig. 3), the high degree of structural conservation observed among multiple views in CDP crystal structures conversely favors the adoption of a single rigid structure for these examples.

CDP ion-channel modulation. Many natural venom and toxin CDPs have been reported to modulate ion-channel activity\(^7\), and our initial target set of CDPs included many related molecules. To assess channel activity, both to verify production of functional CDPs and to identify additional specificities/activities, the selectivity profiles of 37 successfully expressed CDPs were assessed with a commercial electrophysiological assay on a panel of 20 human ion channels (Supplementary Fig. 4). All showed some activity toward at least a subset of ion channels, predominantly potassium channels. Although many of the 37 had not been previously characterized or had been tested on a more focused set of channels, the agreement between our results and the limited prior results was very good, thus reinforcing the utility of our CDP production/characterization platform.

CDP structure-based classification. Coupling optimized CDP crystallography with robust expression platforms enables full determination of the boundaries of CDP fold space, if CDPs can be classified on the basis of sequence and classified on the basis of structure. However, we encountered several practical problems in applying our CDP-defining motif to larger databases to fully catalog candidate CDPs on the basis of sequence. To ensure cysteine formation, the cluster of cysteines defining a candidate CDP should be localized in...
a protein ectodomain. Although the presence of recognizable leader peptides worked well for type I transmembrane proteins, problems were encountered with reliably localizing CDPs to ectodomains in type II and type III transmembrane proteins, and annotations lacked sufficient standardization to allow for confident substitution. Identifying the bounding cysteines in a cluster of cysteines in a sequence was also problematic, because the number of cysteines in CDPs can be highly variable, and can even be an odd number when some CDPs covalently link to another peptide through an interchain cystine (e.g., PDB 1BUN). Likewise, no sequence-based rules were discerned that could rigorously identify the full sequence clustering or correlating sequence with structure. However, limiting the analysis to knotted CDPs showed much better sequence clustering, at identity levels supporting potential linkage of sequence with structure. However, the sequence diversity in CDP space was broad enough to preclude useful clustering or relating sequence to structure. The locations of pacifastin (yellow arc, labeled ‘P’), and hitchin (black-to-gray arcs, labeled by cluster number) clusters defined by structural homology (Fig. 5) are indicated, showing strong sequence similarity between members of the pacifastin and hitchin cluster #1 subsets, but weaker similarities between other hitchin clusters and knotted CDP types in general. The sequence identity between the knotting cystines in the sub-branch associated with the cluster #1 hitchins and the cluster #1 hitchins was ~26%, and the minimal pairwise identity within the cluster #1 hitchins was ~50%. Knotted CDP types are interspersed throughout the phylogram, with types typically clustering only very locally.

The second level of classification, applicable to knotted CDPs, was based on cysteine topology, i.e., among the three core cystines comprising the knotting element, ignoring variable accessory cysteines, which cystine pseudoknots the fold (Fig. 3b). In any connectivity class, denoted \( u-v, w-x, y-z \), where the knotting cystine is indicated by brackets. Nonknotted CDPs can be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.

The first level of our proposed classification was determined on the basis of cysteine connectivity (Fig. 3a). Numbering the cysteines in the three-cystine core or knotting element sequentially from 1 to 6 yields 15 theoretically possible connectivity classes, and archetypical knottins and most GFCKs fell into the 1–4, 2–5, 3–6 class. This class was, by far, the most frequently observed among knotted CDP structures in the PDB (298 examples) and thus is referred to as ‘canonical’. Four other connectivity classes were observed in deposited knotted CDP structures, with variable representation, and nine additional connectivity classes were observed exclusively in deposited nonknotted CDPs with three cystines (Fig. 3b). One connectivity class (1–6, 2–3, 4–5) was not observed in any natural CDPs, although it was found in wholly synthetic designed CDPs (e.g., PDB 5JI4 (ref. 31)). Nonknotted CDPs with more than three cystines cannot be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.

The second level of classification, applicable to knotted CDPs, was based on cystine topology, i.e., among the three core cystines comprising the knotting element, ignoring variable accessory cysteines, which cystine pseudoknots the fold (Fig. 3b). In any connectivity class, denoted \( u-v, w-x, y-z \), where the knotting cystine is indicated by brackets. Nonknotted CDPs can be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.

The second level of classification, applicable to knotted CDPs, was based on cystine topology, i.e., among the three core cystines comprising the knotting element, ignoring variable accessory cysteines, which cystine pseudoknots the fold (Fig. 3b). In any connectivity class, denoted \( u-v, w-x, y-z \), where the knotting cystine is indicated by brackets. Nonknotted CDPs can be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.

The second level of classification, applicable to knotted CDPs, was based on cystine topology, i.e., among the three core cystines comprising the knotting element, ignoring variable accessory cysteines, which cystine pseudoknots the fold (Fig. 3b). In any connectivity class, denoted \( u-v, w-x, y-z \), where the knotting cystine is indicated by brackets. Nonknotted CDPs can be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.

The second level of classification, applicable to knotted CDPs, was based on cystine topology, i.e., among the three core cystines comprising the knotting element, ignoring variable accessory cysteines, which cystine pseudoknots the fold (Fig. 3b). In any connectivity class, denoted \( u-v, w-x, y-z \), where the knotting cystine is indicated by brackets. Nonknotted CDPs can be assigned to comparable connectivity classes, because the focus subset of three cystines cannot be defined and numbered in the same way in the absence of a knotting element and hence were included together in a separate class, ‘\( \cdot \)’.
structures represented as carbon backbones, colored by secondary structure (α, red; β, yellow; coil, green), and with cystines shown in licorice-stick representation, are paired with derived MEME-identified sequence motifs, shown as sequence-logo plots. The red line is an averaged CD spectrum calculated from the set of 14 hitchin spectra and from all 16 successfully expressed hitchins (α, hitchin subset (Fig. 5g)). Sequence conservation for the other three hitchin subsets (Fig. 5l). In contrast to the knottins, a cluster (#2) of βββ four-disulfide-containing hitchins from scorpions (Fig. 5d). A cluster (#3) of βββ four-disulfide-containing hitchins from plants (r.m.s.d. = 0.46 Å; Fig. 5d) and a cluster (#4) of three-disulfide-containing hitchins from various taxa (r.m.s.d. = 0.65 Å; Fig. 5e). These results echo findings from previous studies identifying knotted CDP-subtype structural clusters, e.g., the Möbius and Bracelet clusters of the type 1–4, 2–5, 3–6 cyclotides and, more broadly, the ‘cysteine-stabilized αβ defenses’. These results also validated our proposed classification focused on core cystines by revealing hitchin structural homologies despite the presence of variable accessory cystines.

Structure-based sequence alignments yielded motifs potentially useful for prospectively identifying candidate structural homologs of the pacifastins, which are type 1–4, 2–6, 3–5 knotted CDP serine protease inhibitors from arthropods (global superposition r.m.s. deviation (r.m.s.d.) = 0.35 Å; Fig. 5a). In addition to the pacifastin cluster, available CDP structures were aligned to group CDPs into four more recognizable clusters: a cluster (#1) of αβ three-cystine hitchins from scorpions (r.m.s.d. = 0.55 Å; Fig. 5b); a cluster (#2) of βββ four-disulfide-containing hitchins from scorpions (r.m.s.d. = 0.60 Å; Fig. 5c); a cluster (#3) of βββ four-disulfide-containing hitchins from plants (r.m.s.d. = 0.46 Å; Fig. 5d); and a cluster (#4) of three-disulfide-containing hitchins from various taxa (r.m.s.d. = 0.65 Å; Fig. 5e).
these hitchin-subset clusters all displayed conserved secondary-structure content, with very similar αβ or βαβ folds, as echoed in their respective CD spectra (Fig. 5m,n). Normalized CD spectra from all 14 successfully expressed knottins showed a wide range of secondary-structure compositions, and the spectra from all 16 successfully expressed hitchens showed a fairly narrow range of secondary-structure compositions, in agreement with the greater degree of structural homology observed among hitchens in general.

Taxonomy-based phylogenies (Supplementary Fig. 5), although likely to be affected by experimenter selection bias, showed very uneven distributions among knotted CDPs, examples from arachnids, magnoliopsids, mammals and gastropods, at the class level, predominated. Nonknotted CDPs were predominately from mammalian (class) and primate (order) sources, whereas knotted CDPs were predominately from arachnids, mostly scorpions.

Discussion

We developed an efficient, reliable platform for large-scale production and high-throughput expression screening of endotoxin-free CDPs. The pipeline incorporates multiple steps to stringently validate proper folding (including structure determination by X-ray crystallography) and to assay biological function (including ion-channel inhibition). Starting from a purely sequence-based CDP definition, we also developed a proposed robust, purely structure-based classification system for CDPs, framing an exhaustive analysis of these molecules. The classified CDPs encompassed a wide range of diverse molecules, including epidermal growth factor–like domains, low density lipoprotein–like domains, tumor necrosis factor receptor–like domains, transforming growth factor receptor–like domains, low density lipoprotein–like domains, tumor necrosis factor receptor–like domains, transforming growth factor receptor–like domains, trefollin/plexin domains, notch-repeat-like domains, resistin-like domains, osmotin-like domains, thiamatin-like proteins, disintegrins, anaphylatoxins, insect antifeedance proteins and chitin-binding penaeidins.

The most unexpected result of the analyses of CDP sequence/structure relationships was the very limited correlation between CDP sequence and structure type; this lack of correlation severely restricts the prospective mapping of structures on the basis of sequence alone. The next challenges will be to develop more sophisticated sequence-based tools to completely catalog CDP sequence space, and to apply these tools to guide broader sampling of CDP-structure space, more evenly across taxa, to confidently identify its boundaries. Focused studies to parse the roles of the few conserved residues in determining CDP folds are now possible with high-throughput platforms for expressing panels of sequence variants to deeply sample effects on folding. These tools together enable future exploration of CDP space for advancing basic science, through the study of an exceptional protein-fold family, and clinical application, through production and manipulation of molecules with uniquely useful properties.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41594-018-0033-9.

Received: 26 September 2017; Accepted: 23 January 2018; Published online: 26 February 2018

References

1. Molesini, B., Treggiari, D., Dalbeni, A., Minuz, P. & Pandolfini, T. Plant cystine-knot peptides: pharmacological perspectives. Br. J. Clin. Pharmacol. 83, 63–70 (2017).
2. Herzog, V. & King, G. F. The cystine knot is responsible for the exceptional stability of the insecticidal spider toxin ω-hexatoxin-Hv1a. Toxins (Basel) 7, 4366–4380 (2015).
3. Reinwarth, M., Nasu, D., Kolmar, H. & Avrutina, O. Chemical synthesis, backbone cyclization and oxidative folding of cystine-knot peptides: promising scaffolds for applications in drug design. Molecules 17, 12533–12552 (2012).
4. Kolmar, H. Natural and engineered cystine knot miniproteins for diagnostic and therapeutic applications. Curr. Pharm. Des. 17, 4329–4336 (2011).
5. Postic, G., Gracy, J., Pépin, C., Chuche, L. & Gelly, J. C. KNOTTIN: the database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. Nucleic Acids Res. 46, D454–D458 (2018).
6. Kould, A., Ji, Y., Aboye, T. L. & Camarero, J. A. Cyclotides, a novel ultrastable polyepptide scaffold for drug discovery. Curr. Pharm. Des. 17, 4294–4307 (2011).
7. Schwarz, E. Cystine knot growth factors and their functionally versatile proregions. Biol. Chem. 398, 1295–1308 (2017).
8. Iyer, S. & Acharaya, K. R. Tying the knot: the cystine signature and molecular-recognition processes of the vascular endothelial growth factor family of angiogenic cytokines. FEBs J. 278, 4304–4322 (2011).
9. Kintzing, J. R. & Cochran, J. R. Engineered knottin peptides as diagnostics, therapeutics, and drug delivery vehicles. Curr. Opin. Chem. Biol. 34, 143–150 (2016).
10. Al-Salama, Z. T. & Syed, Y. P. Plecanatide: first global approval. Drugs 77, 593–598 (2017).
11. Veiseh, M. et al. Tumor paint: a chlorotoxin: Cy5.5 bioconjugate for intraoperative visualization of cancer foci. Cancer Res. 67, 6882–6888 (2007).
12. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
13. Nielsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–793 (2004).
14. Liou, Y. C., Tocilj, A., Davies, P. L. & Jia, Z. Mimicry of ice structure by surface hydroxyls and water of a β-helix antifeedance protein. Nature 406, 322–324 (2000).
15. Liang, Z., Sottrup-Jensen, L., Aspán, A., Hall, M. & Söderhäll, K. Pacifastin, a novel 155-kDa heterodimeric proteinase inhibitor containing a unique transferrin chain. Proc. Natl. Acad. Sci. USA 94, 6682–6687 (1997).
16. Moura, A., Savageau, M. A. & Alves, R. Relative amino acid composition signatures of organisms and environments. PLoS One 8, e77319 (2013).
17. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169 (2017).
18. Randaraynaye, A. D. et al. Daedalus: a robust, turnkey platform for rapid production of decigram quantities of active recombinant proteins in human cell lines using novel lentiviral vectors. Nucleic Acids Res. 39, e143 (2011).
19. Lindon, K. A. et al. Autoreactivity and exceptional CDR plasticity (but not unusual polyspecificity) hinder elicitation of the anti-HIV antibody 4E10. PLoS Pathog. 9, e1003639 (2013).
20. Blommel, P. G. & Fox, B. G. A combined approach to improving large-scale production of tobacco etch virus protease. Protein Expr. Purif. 55, 53–68 (2007).
21. Cabrita, L. D. et al. Enhancing the stability and solubility of TEV protease using in silico design. Protein Sci. 16, 2360–2367 (2007).
22. Kapust, R. B. et al. Tobacco etch virus protease: mechanism of autoysis and rational design of stable mutants with wild-type catalytic proficiency. Protein Eng. 14, 993–1000 (2001).
23. Cugatatto, F., López-Requena, A., Burrone, O. R. & Petris, G. Engineered tobacco etch virus (TEV) protease active in the secretory pathway of mammalian cells. J. Biotechnol. 212, 159–166 (2015).
24. Wingerd, J. S. et al. The tarantula toxin β/β-TRTX-Pre1a highlights the importance of the S1-S2 voltage-sensor region for sodium channel subtype selectivity. Sci. Rep. 7, 974 (2017).
25. Liu, Q., Liu, Q. & Hendrickson, W. A. Robust structural analysis of native biological macromolecules from multi-crystal anomalous diffraction data. Acta Crystallogr. D Biol. Crystallogr. 69, 1314–1332 (2013).
26. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545–574 (2014).
27. Kwong, P. D., McDonald, N. Q., Sigler, P. B. & Hendrickson, W. A. Structure of beta 2-bungarotoxin: potassium channel binding by Kunitz modules and targeted phospholipase action. Structure 3, 1109–1119 (1995).
28. Zhu, L. M., Gao, B. & Zhu, S. Y. Origin of neurotoxins from defensins. Sheng Li Xue Bao 67, 239–247 (2015).
29. Tarr, D. E. Establishing a reference array for the CS-α superfamily of defensive peptides. BMC Res. Notes 9, 490 (2016).
30. Wu, Y., Gao, B. & Zhu, S. New fungal defensin-like peptides provide evidence for fold change of proteins in evolution. Biosci. Rep. 37, BSRR20160438 (2017).
31. Bhardwaj, G. et al. Accurate de novo design of hyperstable constrained peptides. Nature 538, 329–335 (2016).
32. Simonet, G., Claeys, I. & Broeck, J. V. Structural and functional properties of a novel serine protease inhibiting peptide family in arthropods. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 132, 247–255 (2002).
33. Lippens, G., Najib, J., Wodak, S. I. & Tartar, A. NMR sequential assignments and solution structure of chlorotoxin, a small scorpion toxin that blocks chloride channels. Biochemistry 34, 13–21 (1995).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.
34. Tsunemi, M., Matsuura, Y., Sakakibara, S. & Katsube, Y. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 Å resolution. Biochemistry 35, 11570–11576 (1996).
35. Francart, C., Dauchez, M., Alix, A. J. & Lippens, G. Solution structure of R-elafin, a specific inhibitor of elastase. J. Mol. Biol. 268, 666–677 (1997).
36. Rost, B. Twilight zone of protein sequence alignments. Protein Eng. 12, 85–94 (1999).
37. Baker, D. & Sali, A. Protein structure prediction and structural genomics. Science 294, 93–96 (2001).
38. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36 (1994).
39. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100 (1990).
40. Janssen, B. J., Schirra, H. J., Lay, F. T., Anderson, M. A. & Craik, D. J. Structure of Petunia hybrida defensin 1, a novel plant defensin with five disulfide bonds. Biochemistry 42, 8214–8222 (2003).

Acknowledgements
We thank J. Carter for substantial logistical support, and J. Simon, A. Mhyre and N. Nairn for helpful discussions. The project was supported by NIH grants R01CA135491-07 (J.M.O.) and T32-H600035 (C.D.B.), Project Violet, the Wissner-Slivka Foundation, the Kismet Foundation, the Sarah M. Hughes Foundation, Strong4Sam, Yahn Bernier and Beth McCaw, Len and Norma Klorfine, Anne Croco and Pocket Full of Hope.

Author contributions
C.E.C., M.M.G., C.M., M.-Y.B., D.M., J.M.O. and R.K.S. designed experiments, analyzed data, administered the project and wrote the manuscript. C.E.C., C.M., A.D.B., W.A.J., W.d.v.d.S. and S.M.T. developed the CDP production and analysis pipeline. M.M.G., P.B.R. and R.K.S. performed crystallographic structure determinations and analyses. C.E.C., M.M.G., M.-Y.B., D.M. and S.E.B. performed phylogenetic and taxonomic analyses. C.E.C., A.D.B., W.A.J., M.C., S.E.R., W.d.v.d.S., S.M.T., A.W. and M.K.C. produced and analyzed CDPs. C.E.C., M.M.G., K.P., C.D.B. and R.K.S. designed and developed SuperTEV.

Competing interests
J.M.O. is a founder and shareholder of Blaze Bioscience, Inc.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41594-018-0033-9.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.M.O. or R.K.S.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

CDP production and purification. For expression of CDPs in mammalian cell lines, the Daedalus\% expression cassette was modified to include, from the N to C terminus: a mouse IgG leader peptide sequence, an optimized FLAG epitope sequence, a hexahistidine tag, a purification-tag sequence, the Scn fusion-partner sequence, the TEV scission sequence (ENLYFQ), a short glycine/serine spacer sequence and the CDP sequence. A minimum of three amino acids before the first CDP cysteine residue was found to be essential for efficient protease cleavage, thus necessitating a short glycine/serine spacer. Because CDPs are found in nonmammalian hosts, but the Daedalus system is based on production in mammalian cells, the targeted CDP sequences were checked and corrected for cryptic N-glycosylation sequences, e.g. in E. coli heat-stable enterotoxin, by mutation to generate fully functional peptides. Difficulties in defining CDP boundaries outside of the bounding cysteines probably also limited the overall success rate. CDP-encoding Daedalus lentiviruses were produced by transient transfection, using psPAX2 (Addgene 12260), pMD2.G (Addgene 12259) and Scn/CDP fusion-encoding vectors, by using linear 25-kDa polyethyleneimine (PEI; Polysciences). Cells were tested for mycoplasma contamination (MycoProbe; R&D Systems). For preparative-scale production of CDPs, transfected cells were cultured in 5 mL of FreeStyle 200 Medium Expression Medium (Thermo Fisher), and the supernatant was fed with 2 mL of FreeStyle Medium supplemented with 6 mM valproic acid (Sigma-Aldrich) after 24 h. Lentivirus was harvested 44 h after medium was filtered through a 0.45-μm Sterilpore filter (Millipore). HEK293F cells (Thermo Fisher) were transduced with 1 mL of lentivirus stock added dropwise in 125-mL shake flasks with 1 × 10^6 cells in 9 mL of FreeStyle 200 Medium supplemented with 6 mM valproic acid. Transduced cells were expanded until a total culture size of 4 L at ~5 × 10^6 cells/mL was reached or viability began to drop. CDP fusion proteins were purified by immobilized metal-affinity chromatography (IMAC)\% with HisTrap FF crude columns (GE) on an AKTA Pure FPLC system (GE). The CDP was cleaved from the Scn fusion partner by protease digestion, separated from unwanted digestion products by size exclusion on Superdex 75 10/300GL or by column chromatography on Phenyl-Sepharose (Amersham) in 20 mM sodium phosphate, 300 mM NaCl, pH 7.4. CDPs were resuspended at a target concentration of 100 μM in PBS. To determine resistance to high temperatures, CDPs were incubated at 0.5 mM in PBS at 75 °C or 100 °C for 1 h and pelleted, and the supernatant was analyzed by high-performance liquid chromatography (HPLC). The incorporated TEV-cleavage sequence in the fusion construct leaves an exogenous GS-disulfide bond at the linker of the N terminus of the recombinant CDP. Expressed CDPs were confirmed by direct-infusion electrospray mass spectrometry (ES–MS) on an LTQ-Orbitrap mass spectrometer (Thermo Electron). For MS–MS, CDPs were dissolved in water at 1 mg/mL, then desalted and purified through C18 ZipTip chromatography (EMD Millipore). Cystine formation was confirmed by analysis of the m/z monoisotopic distribution and determination of net charge, and by crystallography. Calculated and observed m/z values are listed in Supplementary Table 2.

High-throughput pilot-scale expression screening was carried out in an analogous but scaled fashion, with viral production performed in 2-mL deep-well blocks (Axygen), producing 1 mL of lentivirus stock per well. 50–100 μL of the lentivirus stock was used to transduce ~2 × 10^6 HEK293F cells in 1 mL of FreeStyle Medium supplemented with 6 mM valproic acid. Transduction was confirmed by flow cytometry after 36 h, with a NovoCyte cytometer (ACEA), and cultures were fed with 6 mM valproic acid after 120 h. Culture supernatants were harvested after 7 d and transferred to a Protino purification plate (Machery-Nagel) containing 100 μL of Ni–NTA IMAC resin (GE) per well, washed and eluted. Eluted CDPs were cleaved and analyzed by RPC, as described above. For high-throughput expression screening of the broader set of 1,232 CDPs, termini were chosen three residues beyond the binding cysteines in the embedding sequence, or the native termini of the peptide if the sequence did not extend that far.

Biochemical characterization. To determine resistance to high temperatures, CDPs were incubated at 0.5 mM in PBS at 75 °C or 100 °C for 1 h and pelleted, and the supernatant was analyzed by RPC. To determine resistance to proteolytic digestion, CDPs were mixed with 50 U of porcine pepsin (Sigma-Aldrich P7012) in 100 μL of PBS at pH 2. CDPs were incubated at 37 °C for 1 h and pelleted, and the supernatant was analyzed by RPC. Oxidized and reduced forms (prepared through addition 10 mM DTT) were compared. CD

Data availability. Crystal-structure atomic coordinates and diffraction data have been deposited in the PDB under accession codes 6ATL, 6ATN, 6ATU, 6ATW, 6AU7, 6AVP, 6AVH, 6AVA, 6AVC and 6AVD Other data supporting this study are available upon request.

References

41. Knippik, A. & Plüchkhu, N. An improved affinity tag based on the FLAG peptide for the detection and purification of recombinant antibody fragments. Biotechniques 17, 754–761 (1994).
42. Kim, Y. et al. High-throughput protein purification for X-ray crystallography and NMR. Adv. Protein Chem. Struct. Biol. 75, 85–105 (2008).
43. Wang, J., Yadav, V., Smart, C. P. & Bartlett, W. Toward oral delivery of biopharmaceuticals: an assessment of the gastrointestinal stability of 17 peptide drugs. Mol. Pharm. 12, 966–973 (2015).
44. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. 276, 307–326 (1997).
45. McCoy, A. J., Jensen, B. M. & Wlodawer, A. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674 (2007).
46. Winn, M. D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D. Biol. Crystallogr. 67, 235–242 (2011).
47. Sheldrick, G. M. Experimental phasing with SHEXL/D/E: combining chain tracing with density modification. Acta Crystallogr. D. Biol. Crystallogr. 66, 479–485 (2010).
48. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D. Biol. Crystallogr. 60, 2126–2126 (2004).
49. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).

50. Davis, I. W. et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).

51. Graef, J. D. et al. Validation of a high-throughput, automated electrophysiology platform for the screening of nicotinic agonists and antagonists. *J. Biomol. Screen.* **18**, 116–127 (2013).

52. Gillie, D. J., Novick, S. J., Donovan, B. T., Payne, L. A. & Townsend, C. Development of a high-throughput electrophysiological assay for the human ether-a-go-go related potassium channel hERG. *J. Pharmacol. Toxicol. Methods* **67**, 33–44 (2013).

53. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

54. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

55. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

56. Theobald, D. L. & Steindel, P. A. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics* **28**, 1972–1979 (2012).

57. Theobald, D. L. & Wuttke, D. S. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics* **22**, 2171–2172 (2006).

58. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).

59. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol. Biol.* **1079**, 263–271 (2014).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.
**Experimental design**

1. **Sample size**
   Describe how sample size was determined.  
   - Not applicable.

2. **Data exclusions**
   Describe any data exclusions.  
   - No data were excluded from analyses.

3. **Replication**
   Describe the measures taken to verify the reproducibility of the experimental findings.  
   - Experiments were replicated where possible. X-ray diffraction data will be available through the PDB upon publication.

4. **Randomization**
   Describe how samples/organisms/participants were allocated into experimental groups.  
   - Not applicable.

5. **Blinding**
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.  
   - The CRO performing the ion channel assays was blinded to sample identity. Blinding is not applicable in other experiments.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. **Statistical parameters**
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   - The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)  
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  
   - A statement indicating how many times each experiment was replicated  
   - The statistical test(s) used and whether they are one- or two-sided  
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons  
   - Test values indicating whether an effect is present  
   - Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.  
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)  
   - Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.
### Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

- Rosetta for molecular modeling, CLUSTALW and PROMALS3D for sequence alignments;
- Geneious for phylogenetic analyses, MEME for generating sequence motifs, Theseus for structure superpositions, IonWorks for electrophysiology, and HKL200, PHASER, SHELX, Coot, REFMAC, and MolProbity for crystallography (citations provided). Microsoft Excel was also used.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

### Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

Materials produced in these studies are freely available under standard MTA terms, protecting commercialization rights and indemnifying against liability.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

- a. State the source of each eukaryotic cell line used.
  - HEK293, HEK293T, and CHO cells were obtained directly from the ATCC; HEK293F cells were obtained directly from Thermo Fisher.

- b. Describe the method of cell line authentication used.
  - Cell lines were directly sourced from the ATCC or a commercial vendor (Thermo Fisher).

- c. Report whether the cell lines were tested for mycoplasma contamination.
  - Cell lines all tested negative for mycoplasma contamination.

- d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.
  - No commonly misidentified cell lines were used.

### Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

No human subjects were used.