PEPTIDE DESIGN

Comprehensive computational design of ordered peptide macrocycles


Mixed-chirality peptide macrocycles such as cyclosporine are among the most potent therapeutics identified to date, but there is currently no way to systematically search the structural space spanned by such compounds. Natural proteins do not provide a useful guide: Peptide macrocycles lack regular secondary structures and hydrophobic cores, and can contain local structures not accessible with α-amino acids. Here, we enumerate the stable structures that can be adopted by macrocyclic peptides composed of L- and D-amino acids by near-exhaustive backbone sampling followed by sequence design and energy landscape calculations. We identify more than 200 designs predicted to fold into single stable structures, many times more than the number of currently available unbound peptide macrocycle structures. Nuclear magnetic resonance structures of 9 of 12 designed 7- to 10-residue macrocycles, and three 11- to 14-residue bicyclic designs, are close to peptide macrocycle structures. Stable structures that can be adopted by macrocyclic peptides composed of L- and D-amino acids by near-exhaustive backbone sampling followed by sequence design and energy landscape calculations. We identify more than 200 designs predicted to fold into single stable structures, many times more than the number of currently available unbound peptide macrocycle structures. Nuclear magnetic resonance structures of 9 of 12 designed 7- to 10-residue macrocycles, and three 11- to 14-residue bicyclic designs, are close to the computational models. Our results provide a nearly complete coverage of the rich space of structures possible for short peptide macrocycles and vastly increase the available starting scaffolds for both rational drug design and library selection methods.

The high stability, diverse functionality, and favorable pharmacokinetic properties of macrocyclic peptides make them promising starting points for targeted therapeutics (1–4). However, there are few well-characterized natural macrocycles, and they are difficult to repurpose for new functions. Thus, most current approaches focus on random library selection methods, which, although powerful (5–7), only cover a small fraction of the vast sequence space that can be accessed by even short sequences of L- and α-amino acids (8, 9) and often yield peptides that are not structured in the absence of target. Methods are needed for designing ordered macrocycles with shapes precisely crafted to bind their targets and with functionalities common in medicinal chemistry, but absent in the natural 20 amino acids, positioned at critical interaction sites. Despite the progress in computational design of proteins (10–14) and constrained peptides as small as 18 residues (15), designing shorter peptide macrocycles had remained an unsolved challenge. The driving force for the folding of larger peptides and proteins is the sequestration of hydrophobic acids by near-exhaustive backbone sampling followed by sequence design and energy landscape calculations. We identify more than 200 designs predicted to fold into single stable structures, many times more than the number of currently available unbound peptide macrocycle structures. Nuclear magnetic resonance structures of 9 of 12 designed 7- to 10-residue macrocycles, and three 11- to 14-residue bicyclic designs, are close to the computational models. Our results provide a nearly complete coverage of the rich space of structures possible for short peptide macrocycles and vastly increase the available starting scaffolds for both rational drug design and library selection methods.

The Monte Carlo simulated annealing sequence design calculations seek a sequence that minimizes the energy of the target backbone conformation, but there is no guarantee that the sequence found maximizes the energy gap between the target backbone conformation and alternative conformations. To assess the energy landscape for low-energy designs (from 21 designs for length 7 to 673 designs for length 10), 10^4 to 10^5 conformations were generated for each sequence, and the energy was minimized with respect to the backbone and side-chain torsion angles. The energy gap and Boltzmann-weighted probability of finding the peptide in or close to the designed main-chain conformation (P^3̂_total) were estimated from the resulting energy landscapes (Fig. 1C). A total of 12, 22, 45, and 145 designs with distinct backbone structures had energy landscapes strongly funneled into the design target structure for 7-, 8-, 9-, and 10-residue macrocycles, respectively (Fig. 1D and figs. S4 and S5). For comparison, in the PDB and CSD (Cambridge Structure Database), we only found four 7- to 10-residue macrocycle structures not bound to a target and composed only from the 20 canonical amino acids and their mirror images without cross-links or backbone modifications such as N-methylation (Fig. 1D and fig. S6). Because of the constraints imposed by the cyclic backbone, the small size, and the presence of L-amino acids, the designs span a local structural space inaccessible or underexplored in native proteins. Recurrent features include hydrogen-bonded turn-like structures and proline-stabilized kinks, some of which are observed rarely or not at all in native proteins (Fig. 2 and fig. S7A), that can be viewed as macrocycle-generating building blocks (fig. S7B and table S1). Stepwise residue insertion preserves some of the building blocks.
and alters others, resulting in a complex propagation of features from the shorter macrocycles to the longer ones (Fig. 2).

It was not feasible to characterize each of the >200 macrocycle designs (database S1) experimentally. Instead, we chemically synthesized (Fig. S8) a subset of 12 peptides [four 7-residue peptides (7mers), two 8mers, three 9mers, and three 10mers; tables S2 to S4] and experimentally characterized their structures by nuclear magnetic resonance (NMR) spectroscopy (table S5 and figs. S9 to S11). Ten of the 12 peptides had well-dispersed one-dimensional NMR spectra, with the number of backbone HN peaks expected for a single conformation (tables S2 to S5). We collected extensive nuclear Overhauser effect (NOE) data (fig. S11) for these peptides and solved their structures using XPLOR-NIH (18, 19) followed by NOE-restrained molecular dynamics (MD) simulations [similar structures were obtained with an independent large-scale enumeration approach (fig. S12)]. As shown in Figs. 3 and 4 and described below, the experimental NMR structures closely matched the design models for nine of these peptides, and in unrestrained MD simulations, eight out of the nine peptides are within 1 Å of the designed structure more than 75% of the time (figs. S13 to S15 and table S6). These data suggest that a large fraction of our >200 macrocycles are structured as designed.

Unlike proteins, macrocycles cannot be stabilized primarily by the hydrophobic effect as they are too small to form a core that can exclude solvent (20). How, then, do the sequences of the designs specify their structures? To address this question, we computed the effect on folding of every single substitution to a different amino acid with the same chirality, and to an alanine with opposite chirality, at each position, for all the designs with NMR-confirmed structures. For each of the 20XNres variants, full energy landscape calculations were carried out through large-scale backbone enumeration (Figs. 3 and 4; details in fig. S16). These computationally intensive calculations were carried out by using cellular phones and tablets of volunteers participating in the Rosetta@Home distributed computing project (http://boinc.bakerlab.org/rosetta). To evaluate the computed sequence-energy-landscape experimentally, we used SLIM (structures for lossless ion manipulations), an ion-mobility mass spectrometry technique that can distinguish different conformations in small molecular structures (21). This technique requires only a small amount of unpurified sample and enables parallel evaluation of the effects of amino acid substitutions on folding. SLIM results from a set of variants with point mutations of design 7.1 at either the dPro4 or Thr2 position (Fig. 3 and fig. S17) were consistent with the sequence-energy landscape calculations: The structure was perturbed more by mutations at the dPro4 position than at the Thr5 position, consistent with the computed $P_{\text{new}}$ values.

Several general principles emerge from the comprehensive landscape calculations and from folding calculations on permuted sequences (Fig. S18). First, l- and d-proline residues play a key role in structure specification: 52% of the positions in which substitutions disrupt the structure are proline residues in the design, and in almost all of the cases, the most destabilizing mutant of a nonproline residue is a substitution to proline (Fig. S19). Proline is the most torsionally constrained amino acid, and placement of l- and d-proline residues favors specific turn and kink structures. Second, side-chain–to-backbone hydrogen bonds that either stabilize a local structural motif, such as Asp2 in design 8.1, or connect structural motifs, such as Glu2 in design 9.9 or Asp2 in design 10.2, are important for structural specification as removal of these interactions substantially reduces the energy gaps. Third, chirality in many cases plays a greater role in structure specification than side-chain identity: Replacing an amino acid residue with its mirror-image isomer is usually more disruptive than changing to a different amino acid with the same chirality. Fourth, for each design, usually fewer than three residues (often proline) are critical to defining the fold, leaving the remainder largely free for future functionalization (figs. S16 and S19). Even after mutation of the remaining residues to Ala (retaining chirality), a number of the sequences still encode the designed structure (fig. S20). Overall, this global analysis of the
amino acid chirality with the exception of dPro5. The structure is largely specified by the designed proline nucleating the 7- to 10-residue macrocycle series.

The energy landscape calculations show that dPro5 plays a critical role (as in the i-Pro/o-Pro in design 7.3, the second proline plays a less critical role). The structure is expanded by insertion of a kink stabilized by Pro5; the remainder of the structure is completed by a tight AAA, i + 3/i, i + 4 turn. Design 10.1 contains a five-residue distorted helix terminated by the critical dPro5. On one face, the structurally critical Glu2 in the middle of the helix makes a long-range side-chain–backbone hydrogen bond to Arg8, and on the other, Ala2, dVal2, and dLeu6 form a non-polar cluster. Design 10.2 contains BX, YA, and the rare YAX building blocks, each beginning with a proline residue; of these, Pro4 in the BX motif is the most critical. As with 10.1, the building blocks are held together by nonpolar interactions (between dVal2 and dAla6) on one face, and a long-range side-chain–backbone hydrogen bond (from Asp3 to Asn5) on the other; both dVal2 and Asp3 are critical for specifying the structure (Fig. 4, panel IV.E).

The entropic cost of folding continues to increase with increasing number of residues, and for 11- to 14-residue macrocycles, additional cross-links to form bicyclic structures were required to obtain single states amenable to NMR structure determination (table S5). We solved the structures of three such designs (Fig. 4, row IV) that feature long-range backbone-backbone hydrogen bonds. Design 11_SS has a i, i + 3/i + 1, i + 4 building block (Fig. 2, third row) with a critical proline (fig. S16) in the first position preceded by a cysteine that forms a critical disulfide to a cysteine preceding a YA turn. Design 12_SS has a rare BXAX, i, i + 4/i, i + 5 turn, which exhibits higher flexibility in the NMR structure, and a disulfide between non–hydrogen-bonding residues. The more compact and complex 14_SS design has a network of interleaved local and nonlocal backbone hydrogen bonds (22), and a n-Cys to i-Cys disulfide bond. The wide variety of shapes spanned by our macrocycle designs, together with their high stability (fig. S26) and high predicted tolerance for sequence mutations (figs. S16 and S20), makes them attractive starting points for developing new therapeutics. One approach to inhibitor design is scaffolding loops at binding interfaces in the PDB; such scaffolding can increase binding affinity by preorganizing the loops in the binding-convenient conformation, enable additional interactions with the target, and improve cell permeability and oral bioavailability (23). We found that 907 of the 1057 “hot loops” identified at protein-protein interfaces by Kritzer and co-workers (24) (database S2) could be scaffolded by one or more of our designs (see fig. S27 for some examples).

As the macrocycle length increases (9 and 10 residues, Fig. 4), so does the entropic cost of folding, and more hydrogen bonds in increasingly diverse patterns (fig. S25) are required to stabilize the peptide in the folded state. Three of six experimentally characterized designs had structures close to computational models, one was disordered, and two had well-dispersed spectra, but the NOE data did not uniquely define the structures (table S5). Design 9.1 contains a YAA i, i + 3/i, i + 4 building block similar to those in the seven-residue macrocycles in which dPro5 plays a critical role (as in the i-Pro/o-Pro in design 7.3, the second proline plays a less critical role). The structure is expanded by insertion of a kink stabilized by Pro5; the remainder of the structure is completed by a tight AAA, i + 3/i, i + 4 turn. Design 10.1 contains a five-residue distorted helix terminated by the critical dPro5. On one face, the structurally critical Glu2 in the middle of the helix makes a long-range side-chain–backbone hydrogen bond to Arg8, and on the other, Ala2, dVal2, and dLeu6 form a non-polar cluster. Design 10.2 contains BX, YA, and the rare YAX building blocks, each beginning with a proline residue; of these, Pro4 in the BX motif is the most critical. As with 10.1, the building blocks are held together by nonpolar interactions (between dVal2 and dAla6) on one face, and a long-range side-chain–backbone hydrogen bond (from Asp3 to Asn5) on the other; both dVal2 and Asp3 are critical for specifying the structure (Fig. 4, panel IV.E).

The entropic cost of folding continues to increase with increasing number of residues, and for 11- to 14-residue macrocycles, additional cross-links to form bicyclic structures were required to obtain single states amenable to NMR structure determination (table S5). We solved the structures of three such designs (Fig. 4, row IV) that feature long-range backbone-backbone hydrogen bonds. Design 11_SS has a i, i + 3/i + 1, i + 4 building block (Fig. 2, third row) with a critical proline (fig. S16) in the first position preceded by a cysteine that forms a critical disulfide to a cysteine preceding a YA turn. Design 12_SS has a rare BXAX, i, i + 4/i, i + 5 turn, which exhibits higher flexibility in the NMR structure, and a disulfide between non–hydrogen-bonding residues. The more compact and complex 14_SS design has a network of interleaved local and nonlocal backbone hydrogen bonds (22), and a n-Cys to i-Cys disulfide bond. The wide variety of shapes spanned by our macrocycle designs, together with their high stability (fig. S26) and high predicted tolerance for sequence mutations (figs. S16 and S20), makes them attractive starting points for developing new therapeutics. One approach to inhibitor design is scaffolding loops at binding interfaces in the PDB; such scaffolding can increase binding affinity by preorganizing the loops in the binding-convenient conformation, enable additional interactions with the target, and improve cell permeability and oral bioavailability (23). We found that 907 of the 1057 “hot loops” identified at protein-protein interfaces by Kritzer and co-workers (24) (database S2) could be scaffolded by one or more of our designs (see fig. S27 for some examples).

As the macrocycle length increases (9 and 10 residues, Fig. 4), so does the entropic cost of folding, and more hydrogen bonds in increasingly diverse patterns (fig. S25) are required to stabilize the peptide in the folded state. Three of six experimentally characterized designs had structures close to computational models, one was disordered, and two had well-dispersed spectra,
Fig. 3. Seven- to eight-residue macrocycle NMR structures are very close to design models. Columns A: Design model. B: Amino acid sequence, torsion bin string, hydrogen bond pattern, and building block composition. C: Observed backbone-backbone (green), backbone–side-chain (purple), and side-chain–side-chain (orange) NOEs. D: Overlay of design model (green) on MD-refined NMR ensemble (gray; the average backbone RMSD to the NMR ensemble is indicated). E: Average decrease in the propensity to favor the designed state (P\text{Near}, see methods) over all mutations at each position. Darker gray indicates larger decreases (P\text{Near} values for each substitution at each position are in fig. S16); positions particularly sensitive to mutation are boxed and indicated by color in the design model in column A. F: Representative energy funnels for mutations at key positions (colored points) as compared to the design sequence (gray points). Row I, column G: Experimental SLIM data. Distribution of peak width at half height for peptide libraries with all amino substitutions at positions 4 and 5; the position 4 library has a broader distribution consistent with the computed energy landscape in column F. Rows II, IV, V, column G: Representative energy landscapes for double substitutions (red) of critical residues overlaid on the original design landscape (gray). Row III, column G: Overlay of design model on alternative structure NMR ensemble (turn flip at bottom right).
increasing the known repertoire of possible macrocycle structures by more than two orders of magnitude. Our results demonstrate that the principles and energy functions developed in recent years to design proteins have quite broad applicability, transferring over to much smaller systems even though (i) the factors dominating the folding of proteins (for example, the hydrophobic effect) differ considerably from those that stabilize conformation of small peptide macrocycles (local hydrogen bonding patterns and intrinsic conformational preferences of amino acid building blocks), and (ii) all designed proteins to date contain regular α-helix or β-sheet structures, whereas small peptide macrocycles lack these and instead contain a wide range of local structures, some of which are rarely or never observed in proteins.

There are two clear paths forward for engineering new macrocyclic therapeutics by exploiting the rigidity and stability of the designs together with the freedom to choose the identities of the nonstructure-specifying positions. The first is experimental: Libraries can be constructed in which, at each position, all residues compatible with the structure are allowed (identified as described above using large-scale energy landscape calculations) and screened for target binding with current library selection methodologies. The second is computational: Each macrocycle can be docked against the target (using, for example, rigid body docking or “hot loop” superposition), and the interface residues designed to maximize binding affinity. Unnatural amino acids can be incorporated in either approach, but the second has the advantage that new functionalities—such as known active site binding groups—can be strategically placed to maximize binding affinity. Beyond binding, the control over geometry and chemistry provided by our approach should contribute to understanding the structural correlates of membrane permeability and other desirable pharmacological properties.
REFERENCES AND NOTES


ACKNOWLEDGMENTS

This work was supported by NIH Ruth Kirschstein F32 award no. F32GM120791-02 (P.H.), Washington Research Foundation (P.H. and T.W.C.), NIH award GM1157490 (F.P.-A.), American Cancer Society postdoctoral fellowship PF-13-056-01-MMC (M.D.S.), Pew Latin-American fellowship in biomedical sciences (D.-A.S.), CONACYT postdoctoral fellowship (D.-A.S.), NIH–National Institute of General Medical Sciences 5R03GM10834 (G.V.), and Howard Hughes Medical Institute (D.B.). The SLIM experiment was supported by Laboratory-Directed Research and Development (LDRD) and was performed in the Environmental Molecular Sciences Laboratory (EMSL), a U.S. Department of Energy (DOE)–Biological and Environmental Research national scientific user facility at Pacific Northwest National Laboratory. Computing resources were provided by the many volunteers who have donated the spare CPU (central processing unit) cycles of their cellular telephones and computers to the Rosetta@Home project; the Hyak supercomputer at the University of Washington; the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830); and Amazon Web Services (AWS) Credits for Research Grant (G.V.). In addition, an award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. We thank A. Roy, H. Park, and many other members of the Baker lab for discussions; R. Flattland, S. Har, W. J. Lloyd, L. Clithero, and A. Bucher for AWS resources and support; and B. Weitzner, A. Nayyeri, S. Ovchinnikov, T. J. Brunette, D. Alonso, L. Goldschmidt, and P. Vecchiato for help with coding and computing resources. Computational sampling was done with the Rosetta simple_cyclopeptide_prediction application and peptide design was carried out with the rosetta_scripts application, both of which are included in the Rosetta software suite. The Rosetta software suite is available free of charge to academic users and can be downloaded from http://www.rosettacommons.org.

Instructions and inputs for running these applications, and all other data and coding necessary to support the results and conclusions, are in the supplementary materials. All the structures presented here are deposited in PDB and Biological Magnetic Resonance Band (BMRB) (design 7.1: PDB ID 6BE9, BMRB ID 30356; design 7.2: PDB ID 6BEW, BMRB ID 30364; design 7.3: PDB ID 6BF3, BMRB ID 30362; design 7.3b: PDB ID 6BF5, BMRB ID 30365; design 8.1: PDB ID 6BEP, BMRB ID 30355; design 8.2: PDB ID 6BEN, BMRB ID 30357; design 9.1: PDB ID 6BED, BMRB ID 30358; design 10.1: PDB ID 6BEQ, BMRB ID 30359; design 10.2: PDB ID 6BER, BMRB ID 30360; design 11.1: PDB ID 6BE5, BMRB ID 30361; design 11.2: PDB ID 6BET, BMRB ID 30362; design 14.5: PDB ID 6BEU, BMRB ID 30363) and the original designs are provided in database S3. D.B., G.B., P.H., and V.K.M. are inventors on a U.S. provisional patent application submitted by the University of Washington that covers computational design of small structured macrocycles. P.H., G.B., V.K.M., and D.B. developed research ideas, the experimental approach, and generated the designed macrocycles. P.H., G.B., V.K.M., and D.B. wrote the manuscript, and all authors contributed suggestions. G.V. and D.B. supervised research.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/358/6369/1461/suppl/DC1

Methods
Figs. S1 to S29
Tables S1 to S7
Databases S1 to S3
References (25–50)
27 August 2017; accepted 15 November 2017
10.1126/science.aap7577
Comprehensive computational design of ordered peptide macrocycles

Science 358 (6369), 1461-1466.
DOI: 10.1126/science.aap7577

Macrocycles by design
Macrocyclic peptides have diverse properties, including antibiotic and anticancer activities. This makes them good therapeutic leads, but screening libraries only cover a fraction of the sequence space available to peptides comprising D and L amino acids. Hosseinzadeh et al. achieved near-complete coverage in sampling the sequence space for 7- to 10-residue cyclic peptides and identified more than 200 designs predicted to fold into stable structures. Of 12 structures determined, nine were close to the computational models. They also sampled and designed 11- to 14-residue macrocycles, but without complete coverage. The designed macrocycles provide a path forward for engineering new therapeutics.

Science, this issue p. 1461