

θ-Spiral Discovery Reveals a New Class of Biological Building Block

Michael Garton

michael.garton@utoronto.ca

University of Toronto <https://orcid.org/0000-0002-5829-6532>

Biological Sciences - Article

Keywords: protein discovery, noncanonical protein secondary structure, molecular dynamics, NMR spectroscopy, theta-spiral, θ-spiral

Posted Date: August 26th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4850060/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

Our entire living world is constructed from just *seven* major protein secondary structures: α -helices, β -sheets, 3_{10} -helices, π -helices, turns, bends, and *polyPro*-helices. All other polypeptide is classified as *unstructured coil*. Because only a tiny fraction of theoretically possible protein sequences have ever been sampled by evolutionary processes, it is conceivable that other secondary structures remain undiscovered. Using physical and mathematical modelling, we identified an unprecedented structure with spiral geometry. Upon using this to search uncharted polypeptide sequence space *in silico*, we discovered sequences that appeared to support the spiral structure. *Theta*-spirals (ϑ -spirals) were confirmed *in vitro* using NMR spectroscopy, becoming the *eighth* protein secondary structure, and revealing a new class of biological building block: *noncanonical protein secondary structures*.

Main

Discovery of protein α -helix and β -sheet secondary structures were made as early as 1951 based on chemical bond theory, stereochemistry and early crystallographic studies of small polypeptides and amino acids (1, 2). Since then, a number of α -helices and β -sheets, sub-types and other important secondary structures have been discovered. The secondary structure classification system *Define Secondary Structure of Proteins* (DSSP), first developed in 1983 (3) classifies seven types of secondary structure, specifically: α -helix, β -sheet (parallel & anti-parallel), 3_{10} -helix, π -helix, turn, bend, and *polyPro*-helix. Everything else is classified as *unstructured coil* by default, which is not considered a secondary structure. An additional feature identified by DSSP, β -bridges, do not represent standalone secondary structures, but are a pair of hydrogen bonded residues that form part of some β -sheets. Ramachandran diagrams show distinct clusters of dihedral angles for α -helix and β -sheet types, and, in addition to distinct backbone hydrogen bonding, this repetition of similar angles defines them as *regular* secondary structure. For the same reason 3_{10} -helix and π -helices are also defined as *regular* secondary structures. Conversely, turns, bends, *polyPro*-helices, which do not have both repeating dihedral angles and distinct hydrogen bonding networks are classified as *irregular* secondary structures.

Some variation in precise secondary structure definitions exists across different classification systems, most prominently DEFINE (4), STRIDE (5), and SST (6). However, *regular* structures are universally defined by regular dihedral and hydrogen bond patterns (e.g., α -helices have a pattern of $i \rightarrow i + 4$, meaning hydrogen bond acceptors interact with amide donors three positions away). While *individual* hydrogen bonds are often crucial for the formation of *irregular* secondary structures, they do not typically form a recognizable repeating pattern (9) and therefore shape is often used for classification. Irregular secondary structures such as *turns* are also commonly classified by terminus separation length (e.g.: δ -turn, γ -turn, β -turn, α -turn, π -turn, which have $i \rightarrow i + 1$ to $i \rightarrow i + 5$ patterns respectively (7)). Longer turns can be classified as Ω -loops, a motif between 6 to 16 residues in length, with a shape which resembles the upper-case Greek letter Ω (8). Given the minuscule fraction of protein sequence space estimated to ever have been sampled by evolutionary processes (10), it is certainly conceivable that additional secondary structures remain to be discovered.

Discovery of new secondary structures can potentially be facilitated by tools developed for protein design. Protein design has been accelerated by computational methods such as molecular dynamics, fragment-based approaches, and more recently, generative AI (18, 19). *De novo* protein design could be particularly useful for secondary structure discovery, as it allows the elucidation of amino acid sequences that fold into a given target structure. Protein folds with symmetry and repeating structural sub-units have been designed this way, including coiled-coils (11–15), TIM-barrels (16), and β -propellers (17). It can also be used to generate sequences for novel structure targets not previously observed in nature (20–22). More recently, neural networks trained on the PDB have been used to build generative adversarial networks (GAN), variational autoencoders (VAE), and large language models (LLM) that can generate new protein backbone structures (23, 24). However, all novel proteins generated to date have exclusively been constructed from the *seven* secondary structures observed in nature and established since the 1990s.

In this study, we co-opted these well-established computational methods to probe the vast realm of unexplored protein sequence space for undiscovered secondary structures. Physical and mathematical modelling enabled the discovery of a spiral structure unprecedented in nature. Further *in silico* work uncovered amino acid sequences that support spiral folding, and the most stable was used to realize the spiral structure *in vitro* using NMR spectroscopy. We named this eighth secondary structure “*theta*-spiral” (ϑ -spiral) because of its pleasing phonetic flow, spiral-like form, and position of ϑ as the eighth letter of the Greek alphabet. *Theta*-spirals have the potential for size and sequence variation, as well as potential to assemble and interact with canonical secondary structures and other spirals. Their utility and impact are impossible to predict but could be very far reaching.

Spiral polypeptides have secondary structure potential and are not found in nature

In order to create a starting point for the exploration of new secondary structures, we experimented with different shapes not observed in nature (i.e., not in the protein databank (PDB)). A physical poly-alanine model was constructed with removable hydrogen bonds and used to stimulate exploration by whoever passed through our office. A spiral shape emerged as one of the most simple and elegant geometries with secondary structure potential. It had the potential to be formed and stabilized by both hydrogen bonding networks (*regular* secondary structure), and side chain interaction (*irregular* secondary structure). A spiral is a single arc with progressively decreasing curvature as it propagates outward. Allowable dihedral torsion angles φ and ψ were tuned to approximate a spiral backbone structure. Dihedral angles of regular secondary structure elements have been well characterized (25) and visualized using Ramachandran plots. Torsion angles are constrained to a single region on the Ramachandran plot for all the residues that make up a structural element. Inspired by the work of Mannige *et al.* (26) on peptoid nanosheets using two alternating torsion angle zones on the Ramachandran plot, we found that taking a set of dihedral angles and alternating the signs every residue (i.e. Given a φ/ψ set of $50^\circ/-10^\circ$ for one residue, the dihedral angles for the next residue would be

-50°/10°, and then 50°/-10° next, and so forth) induced the backbone to adopt an arc-like structure. Arc curvature could be controlled by varying the magnitude of one of the angles. Initial spiral structures were thereby constructed using the Schrödinger modelling software (Fig. 1).

Spiral structures were made by modifying polypeptide dihedral angles to produce progressively less curvature. Backbone atoms of residues in different concentric rings of the spiral were orientated to face each other (see Fig. 1B) to maximize the chance for hydrogen bond acceptor-donor pair proximity—one of several interaction types that could stabilize novel secondary structures (1, 27). In order to minimize any steric penalties, initial structures were generated using backbone dihedral angles corresponding to the sterically favorable zones of right-handed and left-handed helices on the Ramachandran plot (28). These choices ensure that the requirement of alternating signs to create arc-like chains is satisfied (Fig. 1C & 1E). A set of 20 structures was constructed, each with a length of 34 amino acids. This was the minimum number of amino acids—using favorable phi-/psi- angles—that could produce a reasonable spiral with two concentric rings.

Spiral backbone conformational range explored using molecular dynamics

All 20 structures were simulated for 100 ns using molecular dynamics (MD) to generate a diverse set of spiral backbones. Three replicates for each structure were performed using different force fields, including one specifically designed for simulating intrinsically disordered proteins. Resulting trajectories were clustered to generate a representative structural ensemble of spiral backbone conformations. Diverse conformations allow more diverse amino acid sampling and better chance of uncovering a high quality Rosetta sequence predictions (29). Two different clustering algorithms, DBSCAN (32) and K-means (33) were evaluated using clustering efficiency and accuracy metrics on the MD data. MD trajectories were aggregated and data points were taken every five frames. Resulting data were clustered using DBSCAN and an ensemble of 27 structures representing the conformational space sampled by spirals was determined (Fig. 1F).

Spiral sequence discovered using computational protein design tools

Using the set of 27 spiral backbone structures as queries, novel sequences with the potential to stabilize and fold into them were searched. To avoid canonical secondary structure formation, we narrowed our exploration to sequences that are predicted to be random coil. Each structure from the 27 member conformational ensemble was prepared as an input file for Rosetta's fixbb sequence exploration algorithm (29). One thousand sequences resulted from Rosetta exploration of each of the 27 seed structure queries, for a total of 27,000 sequences predicted to support a stable spiral. Removing duplicates resulted in 16,173 unique sequences. Additional tests were performed to investigate what bias, if any, starting sequence has on the resulting Rosetta identified spiral sequences. Figure 2A logo plots were generated using three different sequence PWMs. *Top*. Made from 16,173 unique sequences

generated from poly-GLU input seeds. *Middle & bottom.* Made from input seeds with two different randomly generated sequences (16,098 & 15,899 sequences respectively). This confirmed that starting sequence does not bias the discovery process.

Spiral sequences predicted to exhibit canonical secondary structure removed

We next sought ways to reduce the sequence set to an experimentally testable number. Among the 16,173 spiral sequences predicted, it was theorized that sequences with any capacity to form canonical secondary structures could compete with spiral formation (35). Secondary structure prediction algorithms were used to filter sequences with potential to compete. Four different secondary structure prediction algorithms were applied serially to identify sequences in which any portion(s) was predicted to form secondary structure. Only sequences where every residue was predicted to form either unstructured coil or not predicted to form anything, were progressed to the next algorithm. Prediction algorithms used in order are: PSSFinder, SSPRED, SPIDER3, and JPRED4 (36–39). This filter produced 207 sequences predicted to have no canonical secondary structure in any of the 34 residue positions (Fig. 2B).

Thermodynamically stable spiral sequences identified using MD.

To further reduce the spiral sequences to an experimentally testable number, all 207 sequences were subjected to MD simulations of increasing length. After an initial round of 50 ns simulation, RMSD values were calculated and trajectories exceeding a stringent 4.0 Å RMSD cut off were discarded (supplementary figure). The remainder were subjected to an additional 50 ns simulation and the pruning process repeated. Sequences still in contention after 100 ns were subjected to a final round of 50 ns simulation with three replicate trajectories. Six sequences remained which consistently had RMSD values lower than 4 Å and were therefore determined to be the most thermodynamically stable. All six candidates had very similar sequences and so an additional sequence that did not meet the final RMSD cutoff was added to the pool for extra sequence diversity (Fig. 2C). Very long (1 μs) MD simulations were run for each of the seven final sequences, with c11_0280 and c11_0869 predicted to be the most thermodynamically stable (Fig. 2D).

CD spectroscopy suggests noncanonical secondary structure character.

All seven of the sequences predicted to be the most thermodynamically stable were synthesized, and their secondary structure characteristics were investigated with circular dichroism (CD) spectroscopy (40). All seven were confirmed to be thermostable. Minimal deviation was observed in the spectra across three temperatures of 5°C, 20°C, and 40°C (Fig. 3A). A negative control sample was made by randomly scrambling the sequence of one of the seven peptide sequences. The random sequence was not predicted to form any canonical secondary structure. CD spectra for canonical secondary structure elements are well established. Predominantly α-helical proteins have two distinct negative peaks at

around 220 nm and 210 nm, and positive ellipticity at wavelengths lower than 200 nm (41). β -sheets produce a single broad negative peak at around 220 nm and positive bands near 195 nm (42). Disordered proteins show close to zero ellipticity above 210 nm and negative values past 200 nm (43). CD spectra for all seven spiral candidate sequences were closer to that of disordered proteins than secondary structure (Fig. 3A). To quantify secondary structure character, CD data were uploaded to the BeStSel (44) webserver. Results in Fig. 3B indicate that spiral candidate sequences are mostly expected to produce a combination of β -sheet, turn and 'other' structural features. This did not rule out the existence of novel secondary structure types with a mix of different canonical structure character.

Spiral structure confirmed in vitro using NMR Spectroscopy

Two-dimensional NMR spectroscopy (46) was used to obtain the three-dimensional structure of candidate sequence c11_0869. This sequence was selected as the most stable in MD simulation (Fig. 2D) that also exhibited the least canonical secondary structure character in CD (Fig. 3B). Homonuclear ^1H 2D Total Correlation Spectroscopy (TOCSY) and Nuclear Overhauser Spectroscopy (NOESY) NMR spectra were obtained. Cross-peaks in the characteristic fingerprint region of the 2D TOCSY spectra, spanning the amide hydrogen band along the f2 axis, and the α -proton to aliphatic proton band along the f1 axis, were first analyzed to assign residue identities. The 2D TOCSY experiment allows the restriction of cross-peak correlations to a single spin system (47), or an amino acid in the context of proteins. Following identification of residue identities via 2D TOCSY, 2D NOESY spectra of corresponding peptides were employed to assign the exact sequence position of the identified amino acid cross-peaks (48, 49). Intensities of the NOE peaks in a NOESY spectrum are correlated with the atomic distance between the two correlating atoms (49). Signal intensities were converted into a list of distance restraints, which were used to model an ensemble of three-dimensional structures (supplementary figure). After assignment of cross-peaks was completed, all fully assigned cross-peaks and a number of cross-peaks that could not be confidently assigned were picked and integrated to calculate their peak volumes. Using the ARIAWeb webserver (50), which implements the ARIA structure calculation program (51), these peak volumes were converted into distance restraints between corresponding atoms. A 2D phase-sensitive COSY (psCOSY) spectrum was also obtained. The multiplet peaks in the psCOSY were processed using NMRPipe (52) and analyzed following the ACME protocol (53) to derive a set of vicinal backbone HN-HA J-coupling constants. These values were used to derive backbone dihedral angles via the Karplus equation during the structure calculation process (54). These restraints, along with a simulated annealing protocol were used to calculate the structural model. Figures 3C & 3D show the calculated structure. The best structure from this ensemble (model #4) aligned very closely to the *in silico* structure (backbone RMSD of NMR structure superimposed with the *in silico* starting structure is 3.0 Å). Most of the difference is due to N- and C-termini not fully engaging with the spiral.

Spiral structure dihedral angles and hydrogen bonding indicate irregular class

Hydrogen bonding analysis was carried out on the NMR spectroscopy determined spiral structure. Figure 3D shows that just two hydrogen bonds were identified, both resulting from a single donor atom

(yellow lines). This was confirmed by MD simulations of the NMR structure (supplementary figure). Lack of repetition of similar dihedral angles (Fig. 3E & 3G) or distinct hydrogen bonding network (Fig. 3D) suggest this spiral should be classified as an *irregular* secondary structure. It is comparable to the Ω -loop irregular secondary structure, which is characterized by a single hydrogen bond that holds the loop in place and does not have a hydrogen bond network or repeated dihedral angle pattern.

NMR spiral structure and sequence confirmed as having no precedent in nature

To assess if spiral structures already exist in the known repertoire of natural and engineered proteins, *in silico* and NMR spiral structures were screened for similarity against every structure in the PDB using Distance Matrix Alignment (via the DALI webserver). DALI evaluates 3D structural similarity between protein structures by performing a pair-wise comparison between user-specified input structures and the entire PDB. This method involves a flexible superimposition of two structures, followed by calculation and comparison of intra-molecular distances. DALI returned *zero* matches in the PDB for both *in silico* and NMR determined spiral structures. Running spiral sequences on the NCBI PBLAST server returned “*no significant similarity found*”. Combined with the DALI result, we conclude that polypeptide spirals are novel in sequence and structure.

Discussion

We report discovery of a new protein secondary structure that we call a ϑ -spiral. Protein secondary structures are a very small group of fundamental biological building blocks, and as such this addition is a significant discovery. Theta is associated with mystery and the unknown, appropriate for a structure discovered far from the shores of evolutionary exploration to date. To our knowledge, this is the first secondary structure to be discovered outside of observation in nature—in the largely unexplored vastness of protein sequence space. The highly incremental nature of evolution means that known secondary structures cluster as tiny islands of highly similar geometry in this space. It is reasonable to conjecture that, if by chance evolution had begun with different polypeptide sequences, these islands, and thus the living world, may be constructed from entirely different secondary structures.

It has not escaped our notice that ϑ -spirals have the potential to vary in radius by changing sequence length, and that they have potential to accommodate sequence variation. It is possible to envisage ϑ -spirals assembling with each other and/or canonical and yet to be discovered noncanonical secondary structures, to produce highly novel proteins. An obvious example might be ϑ -spiral stacking, by fine-tuning the sequence to generate cooperative inter-spiral interfaces. Controllable repeating units, (perhaps producing novel fibres), could be a natural extension of this. It will also be very interesting to express this structure in living organisms, particularly those that undergo rapid evolution. We hope the community will explore the potential of this new secondary structure in both basic and applied scientific contexts. ϑ -spirals are also the first member of a new class of fundamental biology building blocks,

noncanonical protein secondary structures. This study offers a first glimpse into the potentially huge world of novel secondary structures. Our computational workflow can easily be generalized to further explore this space, by, for example, sampling alternating regions on the Ramachandran plot to discover other unique backbone geometries. While impossible to predict, it is nevertheless true that β -spirals – and noncanonical protein secondary structure exploration generally – hold potential for extraordinarily far reaching impact.

Materials and Methods

Structural design of new backbone conformations based on dihedral angles

3D visualization and manipulation of peptide backbone structure was performed using the Schrodinger Maestro software suite and its Protein Preparation Wizard tool (55).

Molecular dynamics simulations and RMSD analysis

All molecular dynamics simulations were performed on ComputeCanada servers using the AMBER software suite (56). Structural PDB files were generated either with Maestro's Protein Preparation Wizard or Rosetta's fixbb application (29). PDB structures were prepared for AMBER using the tLeap module of AmberTools (57). Structures were explicitly solvated in a truncated octahedral box of TIP3P water surrounding the structure by 10 Å on every side. Sodium, chlorine, or potassium counter-ions were added to maintain charge neutrality. The ff14sb AMBER force field for proteins was used (58). Prior to simulation, all structures were energy minimized to remove bad contacts and relax high-energy clashes using a combination of steepest descent and conjugate descent methods with positional restraints on amino acid atoms gradually removed. All systems were heated from 0 to 300K over 20 ps using constant volume (NVT) ensemble parameters (ntb = 1, ntt = 3, ntp = 0). Systems were then equilibrated for an additional 100 ps using constant pressure ensemble (NPT) parameters (ntb = 2, ntt = 3, ntp = 1). Both heating and equilibration used the SHAKE algorithm to constrain hydrogen bond dynamics (59). Production runs were run in either increments of 50 or 100 ns using NPT ensemble parameters and Berendsen thermostat for temperature regulation. Trajectory information was saved every 4 ps. RMSD analysis of trajectories were carried out on trajectories stripped of water molecules with the CPPTRAJ software module (31).

MD trajectory clustering

The DBSCAN algorithm requires the specification of two parameters. The first is 'minpoints' which specifies the minimum number of points in a cluster, and the second is ϵ , the distance cutoff for forming a cluster. In the method described by Ester et al. (32), a rough estimate of these parameters can be obtained from the K-dist plot. The K-dist plot shows, for each frame (X-axis), the K-th nearest distance to

the next data point (Y-axis), sorted by decreasing distance. This was repeated ten times for values of K from 1 to 10. The analysis was performed only on heavy atom types of the peptide. The K-dist plots display an initial steep slope followed by a flattening of the curve at around 1.2 Å. They also show that the shape of the K-dist plots does not change drastically until K = 7 is reached. From this, we chose the values of minpoints and ϵ to be 6 and 1.2, respectively. For the K-means algorithm, the only input parameter to determine is N, the number of clusters to form. Because we had no particular requirement for N, we opted to scan different values of N from 1 to 10 in increments of 1, and then to 100 in increments of 10, to calculate clustering quality metrics for each. The Davies-Bouldin Index (DBI) metric is a measure of how close points within a cluster are to each other relative to how far apart clusters are to other clusters. Lower values of DBI indicate better clustering quality. The pseudo-F-statistic (psF) is a similar measure of cluster quality, where higher values of psF indicate better clustering quality. Both metrics and their use in the analysis of clustering algorithm performance are described in detail by Shao (34). For the K-means algorithm, the optimal clustering results with N = 1. This was not practical for our purposes; however, as we intended to identify unique structures from different clusters as input seeds to perform sequence design and generate a library with some degree of diversity. Nevertheless, it does indicate that the structure conformations in the trajectory are relatively homogenous, which does indicate a degree of conformational stability.

Circular dichroism sample preparation

Peptides were delivered in lyophilized form. A phosphate buffer was prepared according to the following steps: 1.011 g of $\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$ and 0.17 g of $\text{NaH}_2\text{PO}_4 \cdot \text{H}_2\text{O}$ from Sigma Aldrich were added to 800 mL of distilled water. Solution was adjusted to physiological pH using HCL or NaOH. Distilled water was added to achieve final volume of 1L. Stock peptide solutions were made by dissolving 4.2–4.9 mg of lyophilized peptide to 2 mL of buffer. Samples were diluted to a final concentration of 150 μM .

Circular dichroism experiments

CD measurements were taken using the OLIS DSM 1000 Spectrophotometer. Samples were loaded onto a 0.1 cm path length cuvette. For standard CD spectra, measurements were taken from 260 nm to 200 nm in increments of 1 nm with an integration time of 2 s. Three replicate measurements were taken for each sample at each temperature. For thermal melt experiments, measurements were taken at wavelengths offset by up to 2 nm (ex, 220, 221, 222 nm) with an integration time of 10 s. Temperatures were ramped gradually from 20 degrees Celsius to 80 degrees Celsius and ramped back down. Measurements were taken after each heating step. Resulting data was analyzed with OLIS SpectralWorks software.

Bradford assay for determination of peptide concentration

The ThermoFisher Pierce Detergent Compatible Bradford Assay Kit was used to prepare diluted albumin standards (BSA) from 25 $\mu\text{g}/\text{mL}$ to 2000 $\mu\text{g}/\text{mL}$. Two replicate standards were prepared. Two replicates of 10 μL sample peptides and BSA standards were loaded onto a microplate. To each standard and sample well, 300 μL of Coomassie blue dye was added. Plate was incubated for 10 minutes at room

temperature, and absorbance measurements were taken at 595 nm with a plate reader. Measurements in wells containing blank replicates were subtracted from the standard well measurements. A standard curve was obtained by fitting a fourth-order polynomial to the blank-corrected measurements. Sample concentration was then determined using blank-corrected sample measurements.

BCA assay for determination of peptide concentration

The ThermoFisher Pierce BCA Protein Assay Kit was used to prepare diluted albumin standards (BSA) from 25 µg/ml to 2000 µg/ml. Two replicate standards were prepared. Two replicates of 25 µL sample peptides and BSA standards were loaded onto a microplate. The BCA working reagent (WR) was prepared by mixing 5 mL BCA Reagent A and 1 mL BCA Reagent B. To each standard and sample well, 200 µL of the WR was added. Plate was incubated for 30 minutes at 37 degrees Celsius, and absorbance measurements were taken at 562 nm with a plate reader. Measurements in wells containing blank replicates were subtracted from the standard well measurements. A standard curve was obtained by fitting a fourth-order polynomial to the blank-corrected measurements. Sample concentration was then determined using blank-corrected sample measurements.

NMR spectroscopy experiment acquisition

¹H 1D, ¹H-¹H 2D phase-sensitive COSY, TOCSY, and NOESY NMR data were recorded from a 700 MHz Varian NMR spectrometer at the Centre for Spectroscopic Investigation of Complex Organic Molecules and Polymers (CSICOMP) NMR Facility at the University of Toronto. Samples were prepared in a 15% deuterium oxide solvent with watergate water suppression applied. Data were measured at 27.1 degrees Celsius.

NMR data processing and analysis

All initial data processing was performed using the MestreNova software (MNova). Automatic baseline correction, solvent suppression, 2X zero filling were applied along both f1 and f2 dimensions. Manual zero-order and first-order phase corrections were applied for each spectrum. For phase-sensitive COSY spectra, the Amplitude-Constrained Multiplet Evaluation (ACME) software was used, and processing was performed using NMRPipe (52, 53). Visualization of spectra was done using MNova, POKY, and NMRView (60, 61). Assignment of TOCSY and NOESY cross-peaks was done using MNova and POKY. Identification of TOCSY cross-peaks in the fingerprint region of the spectra was performed by referring to online references and descriptions of the typical cross-peak pattern, as well as the average and standard deviation of chemical shifts for each atom of each residue recorded in the Biological Magnetic Resonance Bank. Assigned TOCSY cross-peaks were identified on the complimentary NOESY spectrum and a NOESY walk was performed by analyzing backbone amide and alpha proton chemical shifts in the fingerprint region, as well as in the NH-NH region. Ambiguous or indeterminate cross-peaks were selected, but not assigned. Peak volumes were obtained from POKY NOESY spectra by manually integrating each selected NOE cross-peak, and exported to a data file along with a file containing the resonance assignment information.

Declarations

Acknowledgments

Funded using a Canadian NSERC Discovery Grant

References

1. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* 37, 205–211
2. Pauling L, Corey RB (1951) Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds. *Proc Natl Acad Sci U S A* 37:729–740
3. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
4. Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure*. *Proteins Struct Funct Bioinform* 3:71–84
5. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Bioinform* 23:566–579
6. Konagurthu AS, Lesk AM, Allison L (2012) Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* 28:i97–i105
7. Toniolo C, Benedetti E (1980) Intramolecularly Hydrogen-Bonded Peptide Conformation. *Crit Reviews Biochem* 9:1–44
8. Leszczynski JF, Rose GD (1986) Loops in Globular Proteins: A Novel Category of Secondary Structure. *Science* 234:849–855
9. Fetrow JS (1995) Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J* 9:708–717
10. Huang P-S, Boyken SE, Baker D (2016) The coming of age of de novo protein design. *Nature* 537:320–327
11. Ho SP, DeGrado WF (1987) Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *J Am Chem Soc* 109:6751–6758
12. O’Neil KT, DeGrado WF (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* 250:646–651
13. Thomson AR et al (2014) Computational design of water-soluble α -helical barrels. *Science* 346:485–488
14. Burton AJ, Thomson AR, Dawson WM, Brady RL, Woolfson DN (2016) Installing hydrolytic activity into a completely de novo protein framework. *Nat Chem* 8:837–844

15. Beesley JL, Woolfson DN (2019) The de novo design of α -helical peptides for supramolecular self-assembly. *Curr Opin Biotechnol* 58:175–182
16. Huang P-S et al (2016) De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* 12:29–34
17. Voet ARD et al (2014) Computational design of a self-assembling symmetrical β -propeller protein. *Proceedings of the National Academy of Sciences* 111, 15102–15107
18. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF (1999) Solution structure and dynamics of a de novo designed three-helix bundle protein. *Proc Natl Acad Sci U S A* 96:5486–5491
19. Lemay JK et al (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 17:665–680
20. Kuhlman B et al (2003) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* 302:1364–1368
21. Marcos E et al (2018) De novo design of a non-local β -sheet protein with high stability and accuracy. *Nat Struct Mol Biol* 25:1028–1034
22. Huang P-S et al (2011) RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS ONE* 6:e24109
23. Anand N, Eguchi R, Huang P-S (2022) Fully differentiable full-atom protein backbone generation (November 22, 2022)
24. Eguchi RR, Choe CA, Huang P-S (2022) Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput Biol* 18:e1010271
25. Lovell SC et al (2003) Structure validation by Ca geometry: ϕ , ψ and C β deviation. *Proteins Struct Funct Bioinform* 50:437–450
26. Mannige RV et al (2015) Peptoid nanosheets exhibit a new secondary-structure motif. *Nature* 526:415–420
27. Cheng P-N, Pham JD, Nowick JS (2013) The Supramolecular Chemistry of β -Sheets. *J Am Chem Soc* 135:5477–5492
28. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99
29. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences* 97, 10383–10388
30. Das R, Baker D (2008) Macromolecular Modeling with Rosetta. *Annu Rev Biochem* 77:363–382
31. Roe DR, Cheatham TE (2013) Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9:3084–3095
32. Ester M, Kriegel H-P, Xu X A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 6
33. Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans Inf Theory* 28:129–137

34. Shao J, Tanner SW, Thompson N, Cheatham TE (2007) Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput* 3:2312–2334
35. Soto MA, Sepúlveda A, Tohá J (1985) Conservation of the secondary structure of protein during evolution and the role of the genetic code. *Orig Life Evol Biosph* 16:157–164
36. Salamov AA, Solovyev VV (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol* 247:11–15
37. Pundhir S, Kumar A (2011) SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems. *Bioinformatics* 6:380–382
38. Heffernan R et al (2018) Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J Comput Chem* 39:2210–2216
39. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 43:W389–W394
40. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 1:2876–2890
41. Holzwarth G, Doty P (1965) The Ultraviolet Circular Dichroism of Polypeptides¹. *J Am Chem Soc* 87:218–228
42. Greenfield NJ, Fasman GD (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* 8:4108–4116
43. Sy V, None B, Ia S, Zm W, Cs Y, Jt (1993) Circular dichroic analysis of denatured proteins: inclusion of denatured proteins in the reference set. *Anal Biochem* 214
44. Micsonai A et al (2018) BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res* 46:W315–W322
45. Micsonai A et al (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences* 112, E3095–E3103
46. Aue WP, Bartholdi E, Ernst RR (1976) Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *J Chem Phys* 64:2229–2246
47. Chitayat S, Smith SP (2013) Total Correlation Spectroscopy (TOCSY) in NMR Protein Structure Determination. In: Roberts GCK (ed) *Encyclopedia of Biophysics*. Springer, pp 2622–2623
48. Wüthrich K (1990) Protein structure determination in solution by NMR spectroscopy. *J Biol Chem* 265:22059–22062
49. Clore GM, Gronenborn AM (1989) Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. *Crit Rev Biochem Mol Biol* 24:479–564
50. Allain F, Mareuil F, Ménager H, Nilges M, Bardiaux B (2020) ARIAweb: a server for automated NMR structure calculation. *Nucleic Acids Res* 48:W41–W47

51. Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316
52. Delaglio F et al (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293
53. Delaglio F, Wu Z, Bax A (2001) Measurement of Homonuclear Proton Couplings from Regular 2D COSY Spectra. *J Magn Reson* 149:276–281
54. Wang AC, Bax A (1996) Determination of the Backbone Dihedral Angles ϕ in Human Ubiquitin from Reparametrized Empirical Karplus Equations. *J Am Chem Soc* 118:2483–2494
55. 55 (2021) Maestro, Schrödinger, LLC, New York, NY
56. Case DA et al (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26:1668–1688
57. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285:1735–1747
58. Maier JA et al (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11:3696–3713
59. Kräutler V, van Gunsteren WF, Hünenberger PH (2001) A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J Comput Chem* 22:501–508
60. Lee W, Rahimi M, Lee Y, Chiu A (2021) POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules. *Bioinformatics* 37:3041–3042
61. Johnson BA, Blevins RA, View NMR (1994) A computer program for the visualization and analysis of NMR data. *J Biomol NMR* 4:603–614

Figures

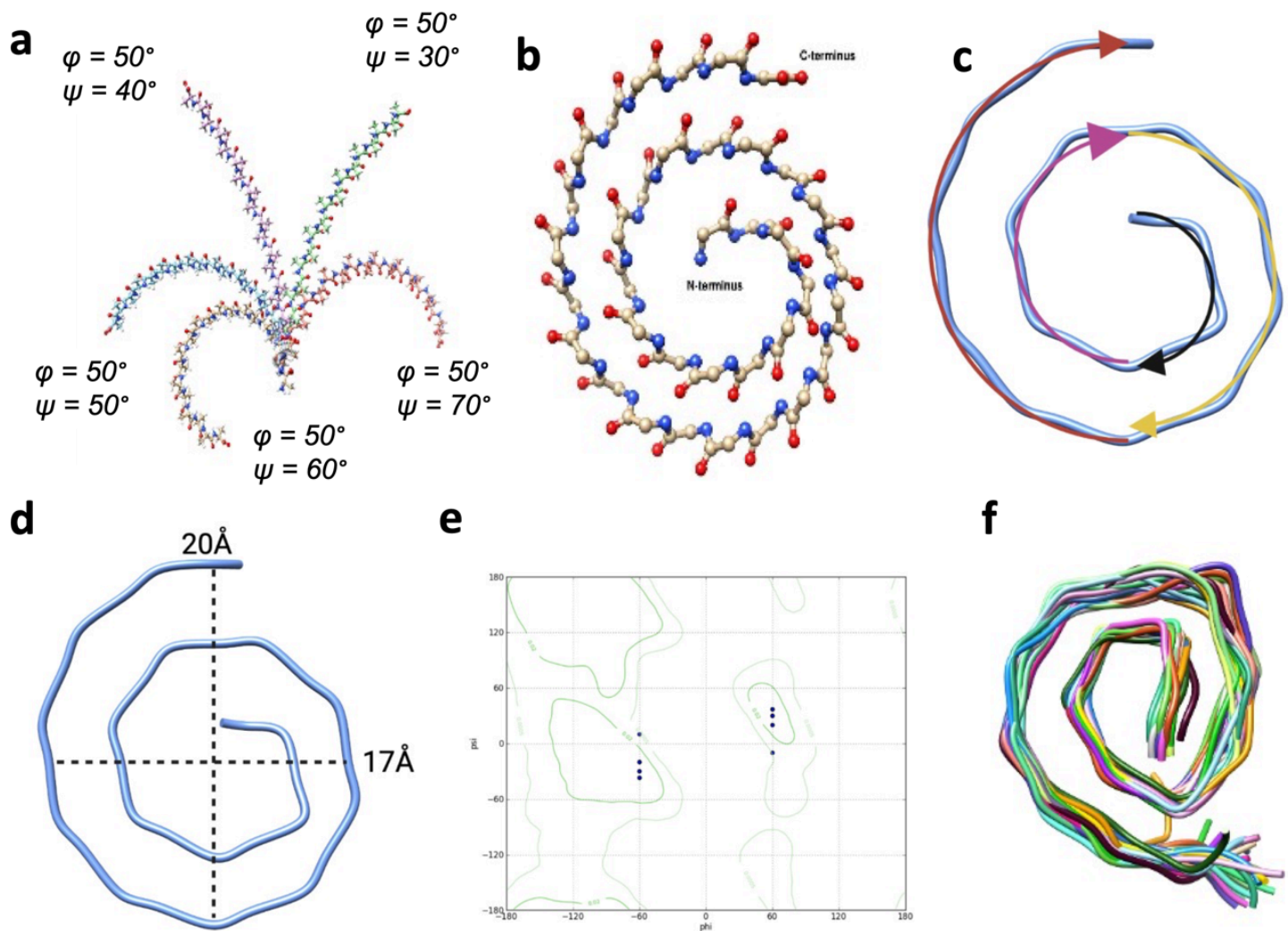


Figure 1

In silico spiral structure modelling. (a) induction of arc-like polypeptide fragments by alternating $\pm \phi$ and ψ pairs, changing the magnitude of ψ & keeping the magnitude of ϕ the same. Magnitude of ϕ is kept constant at 50° , while the magnitude of ψ is varied between 30° to 70° in increments of 10° . (b) initial full spiral structure, consisting of 34 residues, one inner arm and one outer arm (c) illustration of the 4 segments with different sets of alternating ϕ/ψ angles: $0\text{-}\pi$ radians (black arrow), $\pi\text{-}2\pi$ radians (purple arrow), $2\pi\text{-}3\pi$ radians (yellow arrow), $3\pi\text{-}4\pi$ radians (red arrow), (d) Spiral dimensions illustrating one means of classification. (e) Ramachandran plot of spiral structure. (f) All 27 clustered MD representative structures superimposed.

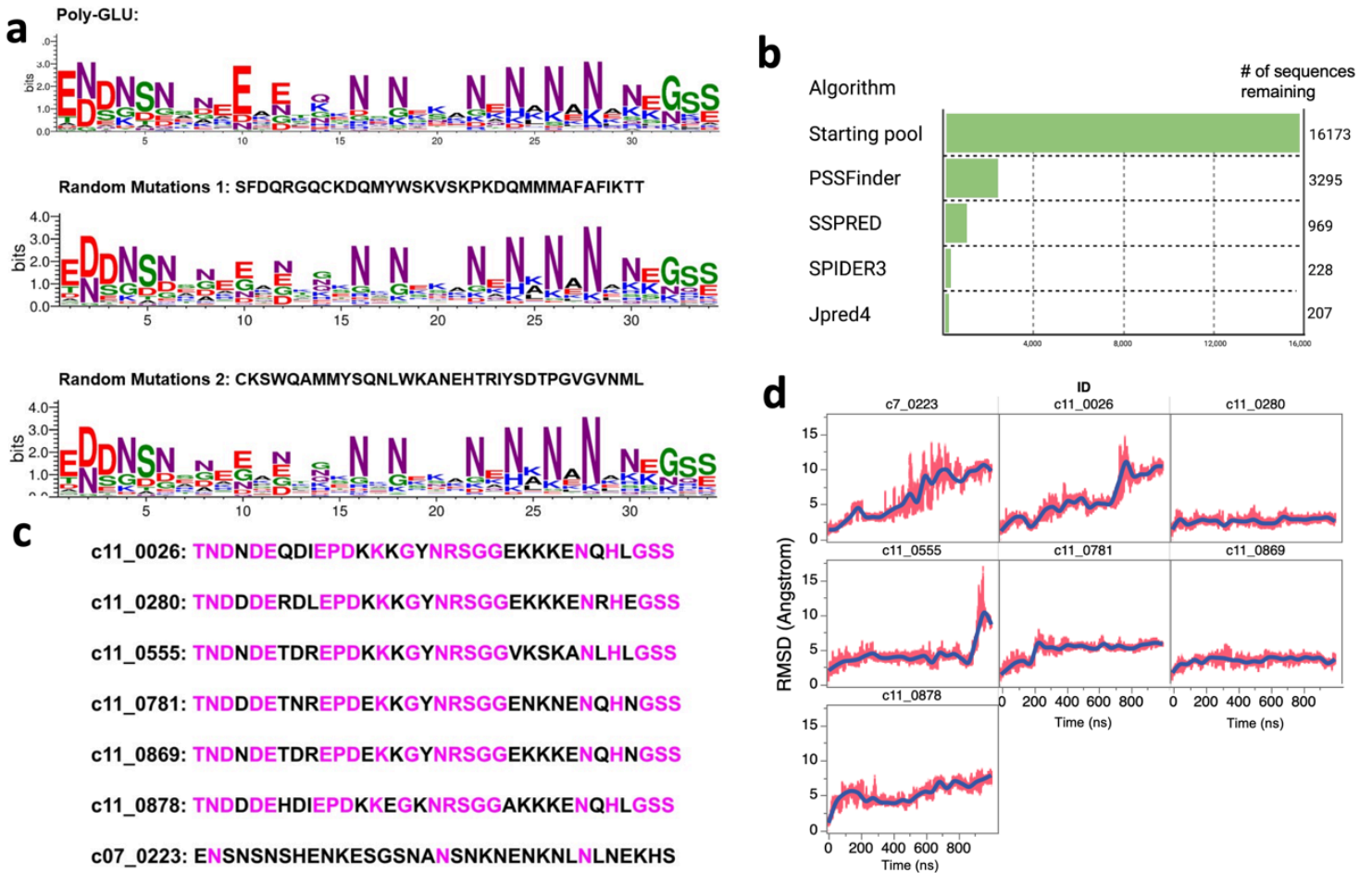


Figure 2

Spiral sequence discovery. (a) Sequence logos showing the distribution of sequences from Rosetta's fixbb algorithm. Top row: generated directly from 27 representative poly-GLU spiral structures. Middle & bottom rows: Same 27 structures randomly mutated to check for starting sequence bias. (b) Serial removal of sequences predicted to have canonical secondary structure using multiple algorithms, reducing the number of candidates from 16,173 to 207. (c) six most stable sequences determined using MD showing significant sequence similarity (conserved residues in pink). Seventh wildcard sequence simulated to increase diversity. (d) RMSD traces of long MD simulations (1 μ s) for the final seven spiral sequences.

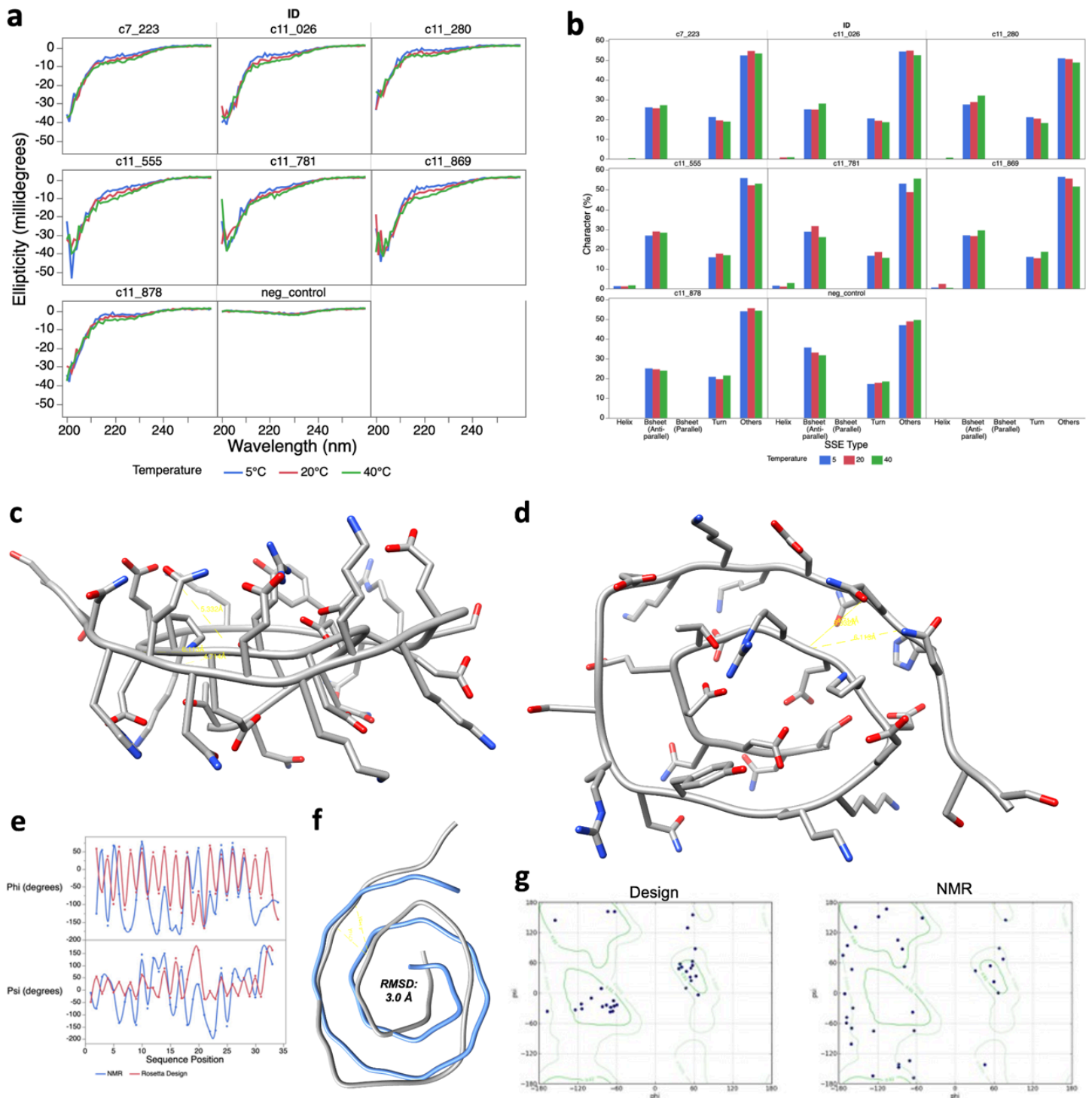


Figure 3

In vitro spiral candidate evaluation (a) Circular dichroism spectra with ellipticity measured from 200-260 nm at temperatures of 5° (blue), 20° (red), 40° (green) celsius for seven candidate peptides and one negative control peptide. (b) Secondary structure character of CD spectra at the three different temperatures (blue = 5°, red = 20°, green = 40°) quantified using the BeStSel method (45). (c) Solved structure of most stable spiral sequence candidate (#c11_0869) using NMR seen from a side/edge view, (d) top down view. Yellow lines indicate backbone hydrogen bonds. (e) Comparison of phi and psi angles per residue position between in silico design and NMR structure, (f) Ribbon representation showing

structural alignment of in silico design (blue) with NMR structure (grey), producing an root mean square deviation of 3.0 Å, (g) Ramachandran plots comparing backbone dihedral angles for the in silico design with NMR structure (c11_869 model #4).