Title: A systematic analysis of the beta hairpin motif in the Protein Data Bank

Running Title: The beta hairpin motif in the Protein Data Bank

Cory D. DuPai[1,2], Bryan W. Davies[1,3], Claus O. Wilke[2†]

[1]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA

[2]Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA.

[3]Center for Systems and Synthetic Biology, John Ring LaMontagne Center for Infectious Diseases, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA

[†]Correspondence: Claus O. Wilke, wilke@austin.utexas.edu

Manuscript: 22 Pages

Figures: 4

Tables: 1

Supplementary Figures: 3

Abstract

The beta hairpin motif is a ubiquitous protein structural motif that can be found in molecules across the tree of life. This motif, which is also popular in synthetically designed proteins and peptides, is known for its stability and adaptability to broad functions. Here we systematically probe all 49,000 unique beta hairpin substructures contained within the Protein Data Bank (PDB) to uncover key characteristics correlated with stable beta hairpin structure, including amino acid biases and enriched inter-strand contacts. We also establish a set of broad design principles that can be applied to the generation of libraries encoding proteins or peptides containing beta hairpin structures.

Keywords

Beta hairpin, computational biology, PDB, protein design

Importance

The beta hairpin motif is a common protein structural motif that is known for its stability and varied activity in diverse proteins. Here we use nearly fifty thousand beta hairpin substructures from the Protein Data Bank to systematically analyze and identify key characteristics of the beta hairpin motif. Ultimately, we provide a set of design principles for the generation of synthetic libraries encoding proteins containing beta hairpin structures.

Introduction

Beta hairpins, one of the simplest stable protein structural elements, consist of two antiparallel beta-sheets joined by a short loop region. Despite their simplicity in form, beta hairpins are highly adaptable in function. Beta strands are known to participate in protein–protein interactions that are often facilitated by specific amino acid orientations[1] and beta hairpin motifs are no different.[2–4] Indeed, these motifs are a core feature in a diverse array of bioactive molecules, from large beta barrel proteins that transport cargo through cellular membranes[5–7] to substantially smaller antimicrobial peptides and peptide derivatives.[8–10] Whether through self-aggregation,[11,12] target binding,[13] or amphipathic structure formation,[6,14] beta hairpin motifs facilitate a range of different biological functions.

In addition to its prevalence in nature, the beta hairpin motif is stable in even small structures and extensively adaptable to specific functions, making it a popular choice in engineered protein structures. Efforts to design such structures have benefited from several decades of research aimed at identifying how beta hairpins form[15–17] and what factors influence their stability and specific activity.[2,18–21] Examples of synthetic proteins that have successfully adapted the beta hairpin motif for specialized functions include hydrogels,[9] antimicrobial peptides,[22] and various molecules with material science applications.[8]

Although largely successful, beta hairpin engineering efforts are typically limited to testing relatively small libraries involving derivatives of a stable scaffold structure or existing protein via peptidomimetics.[2,4,19,23,24] With the increasing availability of high throughput screening platforms to test for activity in large libraries of de novo sequences[25–27] there is an obvious need for broader

design principles that can be applied to the generation of libraries with millions of diverse beta hairpin containing proteins. Knowledge of amino acid propensities throughout known beta hairpin sub-structures could inform such design principles but existing catalogs are too broadly focused on beta sheets, outdated, or limited in scope.[16,20,28–31] An up-to-date characterization of amino acid distributions at specific positions within beta hairpins does not exist.

Using a systematic analysis of sequence and structural data from all beta hairpin containing proteins in the Protein Data Bank (PDB), we derived key sequence factors and patterns common to beta hairpins. Important features include amphipathic faces created by the periodic alternation of hydrophilic and hydrophobic amino acids within beta strands, the high prevalence of aspartic acid/asparagine caps at the N-terminal end of beta strands, and specific residue contacts that are over (e.g. cysteine-cysteine, salt bridges) and under (e.g. proline-lysine) represented. These findings give us a broader understanding of naturally occurring beta hairpins and will aid future efforts in the design of bioactive molecules containing the beta hairpin motif.

Results

General approach

To identify and classify motifs we used the following process (see Materials & Methods for further detail). We first collected all PDB structures[32] and their corresponding amino acid sequences filtered to 90 percent similarity. We then used DSSP-derived secondary structure annotations[33] to identify potential beta hairpin substructures consisting of two antiparallel beta-sheets joined by a short loop region (Fig. 1). After determining contacting residues between beta strands, we

excluded any structures with less than four contacts from further analysis. This process identified nearly 50,000 unique beta hairpin motifs from some 24,000 independent protein structures. Using these structures, we calculated average amino acid frequencies within structural regions and observed amino acid contacts between hairpin beta strands. We then classified and divided motif structures based on turn length and orientation of beta strand faces. Using these groupings, we determined average amino acid frequencies at each position of the beta hairpin motif.

Secondary structure explains average amino acid frequencies

It has long been known that different secondary structural elements tend to favor the inclusion of certain amino acids over others.[29,30,34,35] This is exactly what we see with our analysis of beta hairpin motifs (Fig. 2), with a clear difference in average amino acid frequencies between beta strands, the turn region, and background levels across all included protein structures. Our analysis agrees with previous work illustrating a strong preference for glycine, asparagine, and aspartic acid in flexible turn regions.[29,30] While proline is also more common in the turn region than in either beta strand, we see no difference in turn region prevalence when compared to background levels. This is in contrast to previous findings that saw significant enrichment of proline in turn regions.[8,36] This lack of proline enrichment and the relatively low average proline abundance in the turn region is particularly surprising given the known role of such residues in stabilizing beta turns.[36,37]

When looking at amino acid levels in the beta strands, there appears to be little to no difference in prevalence between strands. Both strands show an increased occurrence of isoleucine, valine, and several other chiefly hydrophobic residues in beta sheet structures, supporting previous research.[38] Additionally, both strands show a greater tolerance for positively

charged residues as is commonly observed with anti-parallel beta strands as opposed to their parallel counterparts.[7,10,39] We further probed for differences across domains of life but saw no strong trends in individual amino acids (Supp. Fig. 1A). There were, however, taxa specific differences in turn region preference for polar and negatively charged amino acids (Supp. Fig. 1B).

Residue positional biases are linked to flexibility, stability, and hydrophobicity

Beta hairpins, especially those in membrane interacting structures such as beta barrels and some antimicrobial peptides, are known to incorporate amphipathic beta sheets that periodically alternate between hydrophilic and hydrophobic amino acids, creating two distinct faces[40,41] (Fig. 1). To account for these faces in our analysis, we divided our dataset based on the presentation of an initial polar or hydrophobic face for both the N and C terminal beta strands (see Materials and Methods). After accounting for these amphipathic faces as well as differences in turn region length, clear patterns emerged in all regions of the beta hairpin motif (Fig. 3). The most obvious pattern observed was the alternating preference for charged/polar and hydrophobic residues in both beta strands (Fig. 3A-B). While hydrophobic residues appear to be more favorable in either beta strand on average (Fig. 2), polar and charged residues are well tolerated when oriented correctly.

On a more granular level, we further surveyed for differences in amino acid frequencies at specific locations within the larger hairpin motif. In contrast to their average beta strand frequencies, hydrophobic amino acids are also less tolerated at the C-terminal edge of either beta strand regardless of orientation. In their place, aspartic acid and (to a lesser extent) asparagine are over-represented at these loci, with this effect being particularly strong for the N-terminal beta strand where the last residue is one of these two amino acids in nearly 20% of observed hairpins.

This frequency is roughly that observed for these two amino acids, on average, in the turn region (Fig. 2, Fig. 3C), although other common turn and cap-associated residues, namely glycine and proline, do not show an over-representation at these positions. Interestingly, aspartic acid residues at the C-terminal end of either beta strand also correlate with increased frequencies of bulky aromatic amino acids (i.e. tyrosine, tryptophan, and phenylalanine) at the N-terminally adjacent position and a preference for glycine at the first N-terminal strand residue (Supp. Fig. 2A).

Although proline showed no enrichment in the average turn region compared to background levels (Fig. 2), proline frequencies are slightly higher than background in the first residue of turns with three to four amino acids and substantially higher than background in the second residue of turns with five amino acids (Fig. 3C). These findings largely agree with existing evidence on the prevalence and importance of prolines in the beginning of turn regions[42–44] but the nearly four-fold enrichment for residue two prolines in hairpin structures with five amino acid long turn regions when compared to background levels is particularly surprising. In combination with the fact that over half of all fourth residues in five amino acid long turn regions are glycines, these findings suggest that beta hairpins with longer turn regions may have very specific physiochemical requirements that limit amino acid diversity.

Amino acid contacts between strands favor stabilizing interactions

As the overall beta hairpin structure is stabilized by interactions between the two beta strands, we sought to identify enriched amino acid pairings between strands to see if certain interactions were more common than expected. Pairings between residues with similar electrostatic properties, that is two hydrophobic residues or a polar residue and a polar/charged residue, were largely more

common than expected (Fig. 4, Supp. Fig. 3). This data agrees with our previous findings regarding the grouping amino acids into beta strand faces based on similar physiochemical properties. In a similar vein to the pairing of electrochemically similar residues, oppositely charged residues tended to pair together in electrostatically favorable salt bridges that are known to stabilize protein structures.[45–48] Such salt bridges represented some of the most enriched amino acid pairings.

The most enriched amino acid pairing between beta strands is that of cysteine with itself to create a structurally stabilizing di-sulfide bond. Such pairings are often used to stabilize engineered peptide structures[49,50] and cysteine coupling is so preferential in nature that many organisms possess a proteome-wide bias towards even numbers of cysteine residues.[51]

In contrast to enriched contact pairings, several classes of interactions, typically those between electrochemically dissimilar residues, were observed much less than expected. The low observance of inter-strand contacts between polar/charged and hydrophobic amino acids (Fig. 4) is intuitive given the strong repulsive nature between such residues which could destabilize overall protein structure.

Design Principles

Taken altogether, our work provides a strong foundation of general principles that can be applied to the design of functionally diverse high throughput beta hairpin libraries (Table 1). First, libraries should seek to incorporate beta strands with amphipathic faces as seen in our analysis of beta strand positional biases (Fig. 3A-B). Second, aspartic acid and asparagine should be favored at C-terminal beta strand residues, especially in the beta strand preceding the turn region. Next, proline and glycine should be utilized in residues two and four of five residue turn regions given their

overwhelming enrichment in these positions (Fig. 3C). Fourth, average secondary-structure amino acid preferences should inform design choices, especially within the turn region. While residues in both hairpin beta strands show positionally specific frequency deviations from secondary structure averages (Fig. 2, Fig. 3A-B), there is much less deviation within the turn region (Fig. 3C). Lastly, stabilizing interactions should be favored between beta strands. Such interactions include salt bridges, disulfide bonds, and the pairing of certain biochemically similar residues (i.e. hydrophobic-hydrophobic and polar-polar pairings) (Fig. 4). These simple guidelines are specific enough to inform design choices while flexible enough to allow for applications across broad research areas.

Table 1: Design principles
1. Incorporate amphipathic beta strand faces
2. Favor aspartic acid/asparagine at C-terminal beta strand residues
3. Incorporate T2 proline and T4 glycine in five residue turns
4. Account for secondary structure biases, especially in the turn region
5. Favor salt bridges and di-cysteine interactions to provide stability

Discussion

By analyzing the composition of beta hairpin motifs across all proteins within the PDB we have identified key characteristics of this versatile structure. Expanding on existing knowledge of secondary structure biases, we outline the preference for the amphipathic orientation of amino acids within beta strands to create two faces with different physiochemical properties. We further identify key positional preferences for specific amino acids in all regions of the hairpin motif. Lastly, we highlight the importance of stabilizing interactions between residues in the N and C terminal beta strands of the hairpin.

Our results integrate and expand upon existing knowledge of protein amino acid biases and intra-protein interactions to provide a systematic framework and novel insights to describe the

beta hairpin motif. We find that stable beta hairpin structures tend to possess site-specific amino acid preferences and to incorporate amphipathic character in both hairpin beta strands. While existing secondary-structure-specific amino acid distributions [29,30] are accurate and informative, such averages prove inadequate to capture the inherent nuances of the beta hairpin motif. For instance, while our analysis finds that an average hairpin beta strand would consist of only hydrophobic residues (Fig. 2), a beta hairpin containing two such average strands without any amphipathic character would be statistically improbable (Fig. 3A-B) and highly unlikely to fold correctly [21], let alone function biologically [10].

Position-specific amino acid biases need to be considered to help form stable protein structures. Our observation that prolines are less enriched in turn regions (Fig. 2) than previously observed [8,36] is perhaps best explained by the extreme position-specific preference of proline residues in turn regions of a given length (Fig. 3C). Thus, certain proline residues are enriched within and likely to stabilize hairpin turn regions even though there is no strong trend when averaged across all turn residues. Outside of the turn region, hairpin beta strands also exhibit amino acid biases at key loci as well as a strong proclivity to incorporate stabilizing inter-strand contacts. We find that asparagine and aspartic acid residues are much more common at the C-terminal end of either hairpin beta strand (Fig. 3A-B, Supp. Fig. 2). These residues may participate in a beta capping phenomenon to block the continuation of beta structure into a turn region [16]. A beta capping role may also explain our observation of an increased prevalence of bulky aromatic residues preceding terminal aspartic acids (Supp. Fig. 2) as aromatic residues are known to stabilize beta hairpin structures [18,19]. Lastly, appropriate contacts between hairpin beta strands are imperative to provide structural stability. As an example, we identified cysteine pairings as being

particularly enriched in beta hairpin substructures (Fig. 4). Such pairings have long been used to stabilize engineered peptide structures [49,50], are so preferential in nature that many organisms possess a proteome-wide bias towards even numbers of cysteines [51].

While our analysis of amino acid preferences within beta hairpin secondary structures across the domains of life showed no strong differences (Supp. Fig. 1A) there were some interesting minor trends as well as a notable difference in turn region composition between taxa (Supp. Fig. 1B). Cysteines, which are fairly uncommon across proteins in general, appear twice as often in Eukaryotic beta hairpins than in Prokaryotic or Archaeaotic beta hairpins. This observation agrees with previous data showing the same trend of increasing cysteine occurrence in proteomes of more complex organisms [52–55]. Of greater note is the inverse relationship between polar and negative amino acid propensities within beta hairpin turn regions across taxa. Frequencies for negatively charged amino acids within the turn region decrease from Archaea to Bacteria, Eukarya, and finally Viruses while polar amino acids show the opposite trend. This difference is likely explained by protein adaptations to harsh environments in Archaea/Bacteria [56] that are less commonly encountered by Eukaryotic or viral proteins. This trend is not seen in either beta strand of the hairpin as turn structures are some of the most accessible protein regions [57] and would likely experience more selective pressure in harsh environments than less exposed beta strands.

One major limitation of our approach is that we were only able to establish broad general properties of beta hairpins that might influence overall structure or function. This is in contrast to prior work that has focused on identifying key design factors for specific beta hairpin scaffolds [23,24,42,58] or grouping beta hairpins and related structures into increasingly detailed classifications

[20,31,44]. While the PDB dataset that we analyzed could be used to expand upon these highly focused areas of research, the broad applicability of our results would be compromised.

In combination with prior research efforts, our simple design guidelines (Table 1) can be adapted to the creation of large-scale protein or peptide libraries aimed at almost any functional purpose, from anticancer drugs to biosensors. For example, beta hairpin antimicrobial peptides are known to incorporate multiple disulfide bonds and favor an overall net positive charge while still maintaining amphipathic character [10,13]. Adapting our design principles with these properties in mind would facilitate the construction of a library of positively charged, disulfide stabilized peptides with presumptive beta hairpin structure to test for antimicrobial activity.

In summary, our findings are broadly adaptable to creating large libraries of beta hairpin containing molecules skewed towards a specific functionality and will help engineering efforts keep pace with the ever-expanding capacity of screening assays.

Materials and Methods

Identification of beta hairpin substructures. We defined the beta hairpin motif as an amino acid sequence containing two sets of four to fourteen extended beta strand residues joined by one to five turn, bend, or unannotated residues. A maximum beta strand length of fourteen was selected based on the typical length of beta strands in monomeric beta barrel proteins [59] while the range of turn lengths was selected based on prior research into beta hairpins [17]. We searched DSSP [33] derived secondary structure annotations of all PDB proteins (downloaded from https://cdn.rcsb.org/etl/kabschSander/ss.txt.gz on July 22nd 2020) for this motif. We further filtered

our dataset to include only IDs for representative structures clustered to within 90% sequence identity. Clusters were obtained from PDB on July 22nd 2020 using the RESTful Web Service Interface (https://www.rcsb.org/pdb/software/rest.do). Further manual filtering was applied to exclude redundant and overly similar hairpin sequences, largely from structures of nanobodies, antibodies, and their derivatives.

Identification of contacting residues. To ensure that our analyzed motifs possessed the correct beta hairpin 3D structure, we filtered our dataset to only include structures in which at least four amino acid side chain pairs formed contacts between the N and C terminal beta strands. We defined contacts as any pair of residues in which side-chain beta carbons were within 8 Angstroms of one another. Determining contacts via the presence of backbone hydrogen bonds produced similar results (data not included). To calculate expected contact frequencies, individual amino acid frequencies were derived using the relative occurrence of each amino acid across all contact pairs. Values for amino acids in a pairing were then multiplied together to establish an expected frequency for every possible pairing of amino acids.

Grouping of beta hairpin substructures. To characterize the amphipathic faces of each beta strand, solvent accessibility was averaged across odd and even numbered amino acid residues with the first amino acid being the residue closest to the turn region. Strands in which the odd amino acid residues have a higher mean accessibility were categorized as polar while strands with the opposite phenotype were categorized as hydrophobic. Solvent accessibility was chosen in lieu of

hydrophobicity or other metrics as PDB structures contain accessibility information and solvent accessibility is known to correlate with hydrophobicity [57].

Data and figures. All data was analyzed in R using the tidyverse family of packages [60] in combination with the data.table [61] and seqinr [62] packages. All figures were created using ggplot2 [63] and cowplot [64]. Supplementary Figure 3 additionally utilized the ggseqlogo package [65]. All processed data and analysis scripts are available at https://doi.org/10.5281/zenodo.4069580.

Acknowledgements

Funding. This research was supported by NIH awards R01 AI125337 and R01 AI148419 as well as Welch Foundation award F-1870. Funders had no role in study design. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Watkins AM, Arora PS (2014) Anatomy of β-strands at protein-protein interfaces. ACS Chem. Biol. 9:1747–1754.

2. Robinson JA (2008) β-Hairpin Peptidomimetics: Design, Structures and Biological Activities. Acc. Chem. Res. 41:1278–1288.

3. Chakraborty K, Shivakumar P, Raghothama S, Varadarajan R (2005) NMR structural analysis of a peptide mimic of the bridging sheet of HIV-1 gp120 in methanol and water. Biochem. J. 390:573–581.

4. Batalha IL, Lychko I, Branco RJF, Iranzo O, Roque ACA (2019) β-Hairpins as peptidomimetics of human phosphoprotein-binding domains. Org. Biomol. Chem. 17:3996–4004.

5. Remmert M, Biegert A, Linke D, Lupas AN, Söding J (2010) Evolution of Outer Membrane β-Barrels from an Ancestral ββ Hairpin. Mol. Biol. Evol. 27:1348–1358.

6. Chaturvedi D, Mahalakshmi R (2017) Transmembrane β-barrels: Evolution, folding and energetics. Biochim. Biophys. Acta - Biomembr. 1859:2467–2482.

7. Gupta K, Jang H, Harlen K, Puri A, Nussinov R, Schneider JP, Blumenthal R (2013) Mechanism of membrane permeation induced by synthetic β-hairpin peptides. Biophys. J. 105:2093–2103.

8. Marcelino AMC, Gierasch LM, Craik C (2008) Roles of beta-turns in protein folding: from peptide models to protein engineering. Biopolymers 89:380–391.

9. Dong N, Ma Q, Shan A, Cao Y (2012) Design and biological activity of beta-hairpin-like antimicrobial peptide. Sheng Wu Gong Cheng Xue Bao 28:243—250.

10. Edwards IA, Elliott AG, Kavanagh AM, Zuegg J, Blaskovich MAT, Cooper MA (2016) Contribution of amphipathicity and hydrophobicity to the antimicrobial activity and cytotoxicity

317 of β-hairpin peptides. ACS Infect. Dis. 2:442–450.

318 11. Mirecka EA, Shaykhalishahi H, Gauhar A, Akgül Ş, Lecher J, Willbold D, Stoldt M, Hoyer W

319 (2014) Sequestration of a β-Hairpin for Control of α-Synuclein Aggregation. Angew. Chemie Int.

320 Ed. 53:4227–4230.

321 12. Larini L, Shea JE (2012) Role of β-hairpin formation in aggregation: The self-assembly of the

322 amyloid-β(25-35) peptide. Biophys. J. 103:576–586.

323 13. Panteleev P V, Bolosov IA, Balandin S V, Ovchinnikova T V (2015) Structure and Biological

324 Functions of β-Hairpin Antimicrobial Peptides. Acta Naturae 7:37–47.

325 14. Worthington P, Langhans S, Pochan D (2017) β-hairpin peptide hydrogels for package

326 delivery. Adv. Drug Deliv. Rev. 110–111:127–136.

327 15. Lewandowska A, Ołdziej S, Liwo A, Scheraga HA (2010) β-hairpin-forming peptides; models

328 of early stages of protein folding. Biophys. Chem. 151:1–9.

329 16. FarzadFard F, Gharaei N, Pezeshk H, Marashi SA (2008) β-Sheet capping: Signals that initiate

330 and terminate β-sheet formation. J. Struct. Biol. 161:101–110.

331 17. Gunasekaran K, Ramakrishnan C, Balaram P (1997) β-Hairpins in proteins revisited: Lessons

332 for de novo design. Protein Eng. 10:1131–1141.

333 18. Santiveri CM, Jiménez MA (2010) Tryptophan residues: Scarce in proteins but strong

334 stabilizers of β-hairpin peptides. Pept. Sci. 94:779–790.

335 19. Mahalakshmi R (2019) Aromatic interactions in β-hairpin scaffold stability: A historical

336 perspective. Arch. Biochem. Biophys. 661:39–49.

337 20. Milner-White EJ, Poet R (1986) Four classes of beta-hairpins in proteins. Biochem. J.

338 240:289–292.

339 21. Popp A, Wu L, Keiderling TA, Hauser K (2014) Effect of Hydrophobic Interactions on the
340 Folding Mechanism of β-Hairpins. J. Phys. Chem. B:118–14234.
341 22. Rughani R V, Schneider JP (2008) Molecular Design of beta-Hairpin Peptides for Material
342 Construction. MRS Bull. 33:530–535.
343 23. Di Natale C, La Manna S, Avitabile C, Florio D, Morelli G, Netti PA, Marasco D (2020)
344 Engineered β-hairpin scaffolds from human prion protein regions: Structural and functional
345 investigations of aggregates. Bioorg. Chem. 96.
346 24. Cochran AG, Tong RT, Starovasnik MA, Park EJ, McDowell RS, Theaker JE, Skelton NJ (2001)
347 A minimal peptide scaffold for β-turn display: Optimizing a strand position in disulfide-cyclized
348 β-hairpins. J. Am. Chem. Soc. 123:625–632.
349 25. Wu C-H, Liu I-J, Lu R-M, Wu H-C (2016) Advancement and applications of peptide phage
350 display technology in biomedical science. J. Biomed. Sci. 23:8.
351 26. Tucker AT, Leonard SP, DuBois CD, Knauf GA, Cunningham AL, Wilke CO, Trent MS, Davies
352 BW (2018) Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of
353 Surface-Displayed Peptide Libraries. Cell 172:618-621.e13.
354 27. Wójcik M, Telzerow A, Quax WJ, Boersma YL (2015) High-Throughput Screening in Protein
355 Engineering: Recent Advances and Future Perspectives. Int. J. Mol. Sci. 16:24918–24945.
356 28. Bhattacharjee N, Biswas P (2010) Position-specific propensities of amino acids in the -
357 strand. BMC Struct. Biol. 10:1–10.
358 29. Fujiwara K, Toda H, Ikeguchi M (2012) Dependence of α-helical and β-sheet amino acid
359 propensities on the overall protein fold type. BMC Struct. Biol. 12:18.
360 30. Otaki JM, Tsutsumi M, Gotoh T, Yamamoto H (2010) Secondary Structure Characterization

361 Based on Amino Acid Composition and Availability in Proteins. J. Chem. Inf. Model. 50:690–700.

362 31. Sibanda BL, Blundell TL, Thornton JM (1989) Conformation of β-hairpins in protein

363 structures. A systematic classification with applications to modelling by homology, electron

364 density fitting and protein engineering. J. Mol. Biol. 206:759–777.

365 32. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank.

366 Nat. Struct. Biol. 10:980.

367 33. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of

368 hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637.

369 34. Koehl P, Levitt M (1999) Structure-based conformational preferences of amino acids. Proc.

370 Natl. Acad. Sci. 96:12524 LP – 12529.

371 35. Geisow MJ, Roberts RDB (1980) Amino acid preferences for secondary structure vary with

372 protein class. Int. J. Biol. Macromol. 2:387–389.

373 36. Trevino SR, Schaefer S, Scholtz JM, Pace CN (2007) Increasing protein conformational

374 stability by optimizing beta-turn sequence. J. Mol. Biol. 373:211–218.

375 37. Melnikov S, Mailliot J, Rigger L, Neuner S, Shin B-S, Yusupova G, Dever TE, Micura R,

376 Yusupov M (2016) Molecular insights into protein synthesis with proline residues. EMBO Rep.

377 17:1776–1784.

378 38. Tsutsumi M, Otaki JM (2011) Parallel and antiparallel beta-strands differ in amino acid

379 composition and availability of short constituent sequences. J Chem Inf Model 51.

380 39. Bowerman CJ, Nilsson BL (2012) Self-assembly of amphipathic β-sheet peptides: insights

381 and applications. Biopolymers 98:169–184.

382 40. Sivanesam K, Kier BL, Whedon SD, Chatterjee C, Andersen NH (2016) Hairpin structure

383 stability plays a role in the activity of two antimicrobial peptides. FEBS Lett. 590:4480–4488.

384 41. Zhang XC, Han L (2016) How does a β-barrel integral membrane protein insert into the

385 membrane? Protein Cell 7:471–477.

386 42. Blandl T, Cochran AG, Skelton NJ (2003) Turn stability in beta-hairpin peptides: Investigation

387 of peptides containing 3:5 type I G1 bulge turns. Protein Sci. 12:237–247.

388 43. Wang H, Varady J, Ng L, Sung S-S (1999) Molecular dynamics simulations of β-hairpin

389 folding. Proteins Struct. Funct. Bioinforma. 37:325–333.

390 44. Shapovalov MI, Vucetic S, Dunbrack RL (2019) A new clustering and nomenclature for beta

391 turns derived from high-resolution protein structures.

392 45. Pylaeva S, Brehm M, Sebastiani D (2018) Salt Bridge in Aqueous Solution: Strong Structural

393 Motifs but Weak Enthalpic Effect. Sci. Rep. 8:13626.

394 46. Ciani B, Jourdan M, Searle MS (2003) Stabilization of β-Hairpin Peptides by Salt Bridges:

395 Role of Preorganization in the Energetic Contribution of Weak Interactions. J. Am. Chem. Soc.

396 125:9038–9047.

397 47. Donald JE, Kulp DW, DeGrado WF (2011) Salt bridges: geometrically specific, designable

398 interactions. Proteins 79:898–915.

399 48. Bosshard HR, Marti DN, Jelesarov I (2004) Protein stabilization by salt bridges: concepts,

400 experimental approaches and clarification of some misunderstandings. J. Mol. Recognit. 17:1–

401 16.

402 49. Sussman D, Westendorf L, Meyer DW, Leiske CI, Anderson M, Okeley NM, Alley SC, Lyon R,

403 Sanderson RJ, Carter PJ, et al. (2018) Engineered cysteine antibodies: an improved antibody-

404 drug conjugate platform with a novel mechanism of drug-linker stability. Protein Eng. Des. Sel.

405 31:47–54.

406 50. Dombkowski AA, Sultana KZ, Craig DB (2014) Protein disulfide engineering. FEBS Lett.

407 588:206–212.

408 51. Dutton RJ, Boyd D, Berkmen M, Beckwith J (2008) Bacterial species exhibit diversity in their

409 mechanisms and capacity for protein disulfide bond formation. Proc. Natl. Acad. Sci.

410 105:11933–11938.

411 52. Brooks DJ, Fresco JR (2002) Increased Frequency of Cysteine, Tyrosine, and Phenylalanine

412 Residues Since the Last Universal Ancestor. Mol. & Cell. Proteomics 1:125 LP – 131.

413 53. Wiedemann C, Kumar A, Lang A, Ohlenschläger O (2020) Cysteines and Disulfide Bonds as

414 Structure-Forming Units: Insights From Different Domains of Life and the Potential for

415 Characterization by NMR . Front. Chem. 8:280.

416 54. Miseta A, Csutora P (2000) Relationship Between the Occurrence of Cysteine in Proteins and

417 the Complexity of Organisms. Mol. Biol. Evol. 17:1232–1239.

418 55. Tsuji J, Nydza R, Wolcott E, Mannor E, Moran B, Hesson G, Arvidson T, Howe K, Hayes R,

419 Ramirez M, et al. (2010) The frequencies of amino acids encoded by genomes that utilize

420 standard and nonstandard genetic codes. Bios 81:22–31.

421 56. Reed CJ, Lewis H, Trejo E, Winston V, Evilia C (2013) Protein adaptations in archaeal

422 extremophiles. Archaea 2013.

423 57. Lins L, Thomas A, Brasseur R (2003) Analysis of accessible surface of residues in proteins.

424 Protein Sci. 12:1406–1417.

425 58. Pastor MT, López de la Paz M, Lacroix E, Serrano L, Pérez-Payá E (2002) Combinatorial

426 approaches: A new tool to search for highly structured β-hairpin peptides. Proc. Natl. Acad. Sci.

427 99:614 LP – 619.

428 59. Tamm LK, Hong H, Liang B (2004) Folding and assembly of β-barrel membrane proteins.

429 Biochim. Biophys. Acta - Biomembr. 1666:250–263.

430 60. Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L, François R, Grolemund G,

431 Hayes A, Henry L, Hester J, et al. (2019) Welcome to the Tidyverse. J. Open Source Softw.

432 4:1686.

433 61. Dowle M, Srinivasan A (2020) data.table: Extension of `data.frame`.

434 62. Charif D, Lobry JR SeqinR 1.0-2: A Contributed Package to the R Project for Statistical

435 Computing Devoted to Biological Sequences Retrieval and Analysis BT - Structural Approaches

436 to Sequence Evolution: Molecules, Networks, Populations. In: Bastolla U, Porto M, Roman HE,

437 Vendruscolo M, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. pp. 207–232.

438 63. Wilkinson L (2011) ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. Biometrics

439 67:678–679.

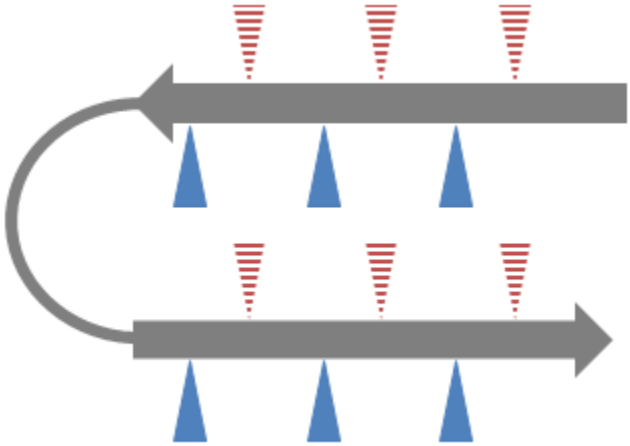
440 64. Wilke CO (2019) cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."

441 65. Wagih O (2017) ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics

442 33:3645–3647.

443

444 **Figures**



445

446 **Figure 1: General beta hairpin structure.**

447 Beta hairpins consist of two anti-parallel beta strands (grey arrows) linked with a flexible turn

448 region (grey line). Beta strands typically have amphipathic characteristics conferred by alternating

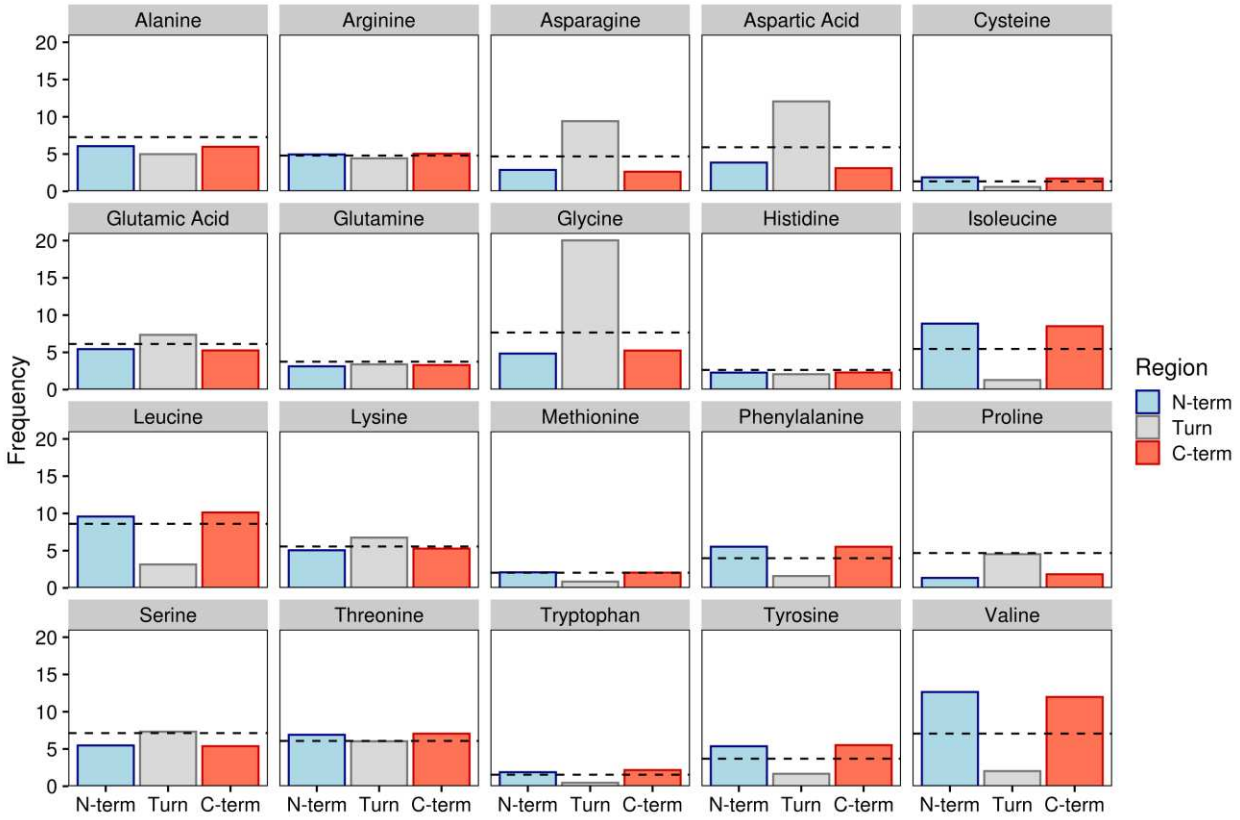
449 hydrophobic and hydrophilic residues. Triangles represent beta strand side amino acid side chains,

450 with red indicating hydrophobic and blue indicating hydrophilic residues. Dashed triangles indicate

451 side chains oriented away from the viewer while solid triangles indicate side chains oriented

452 towards the viewer.

453



454

455 **Figure 2: Amino acid frequencies by beta hairpin secondary structure region.**

456 Bars indicate average amino acid frequencies for each amino acid within a given region of all beta

457 hairpins. The black dashed line indicates background amino acid frequencies for all sites in all

458 proteins containing the beta hairpin motif. N-term and C-term refer to the N- and C-terminal beta

459 strands while turn denotes the turn region.

460

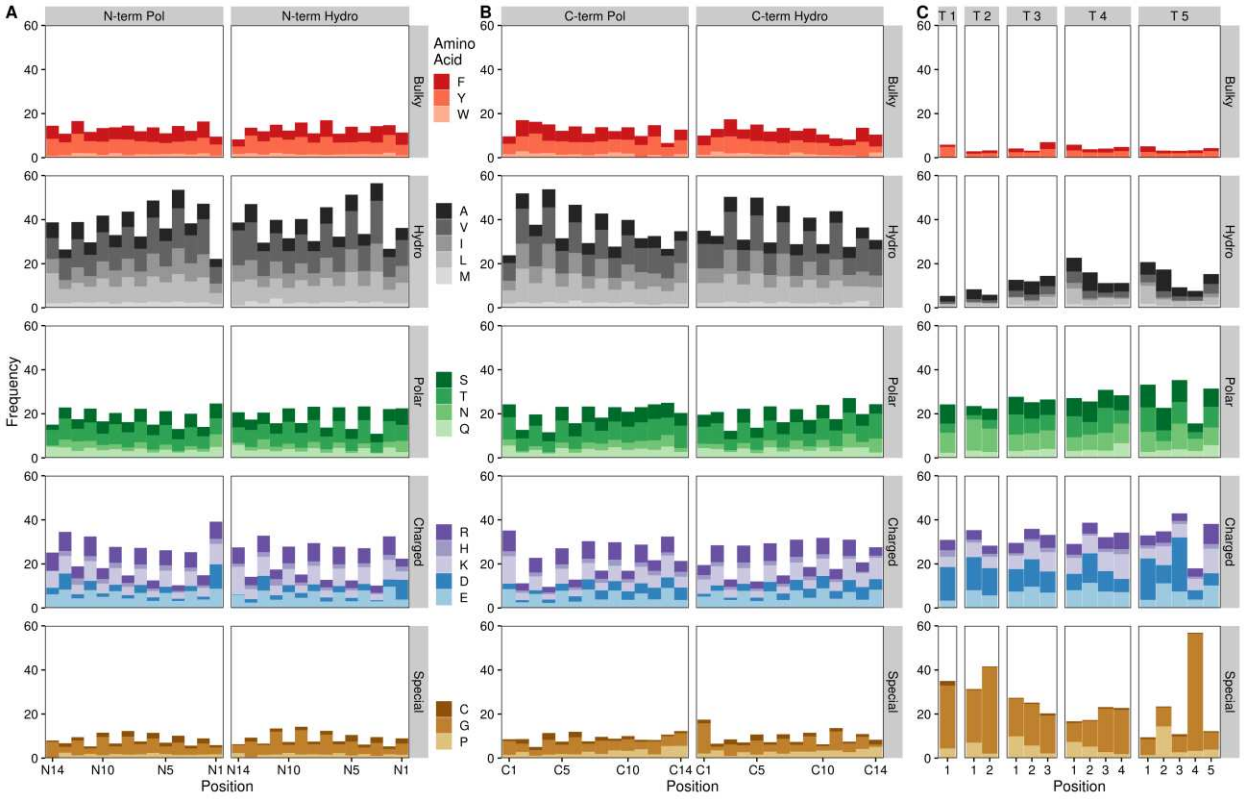
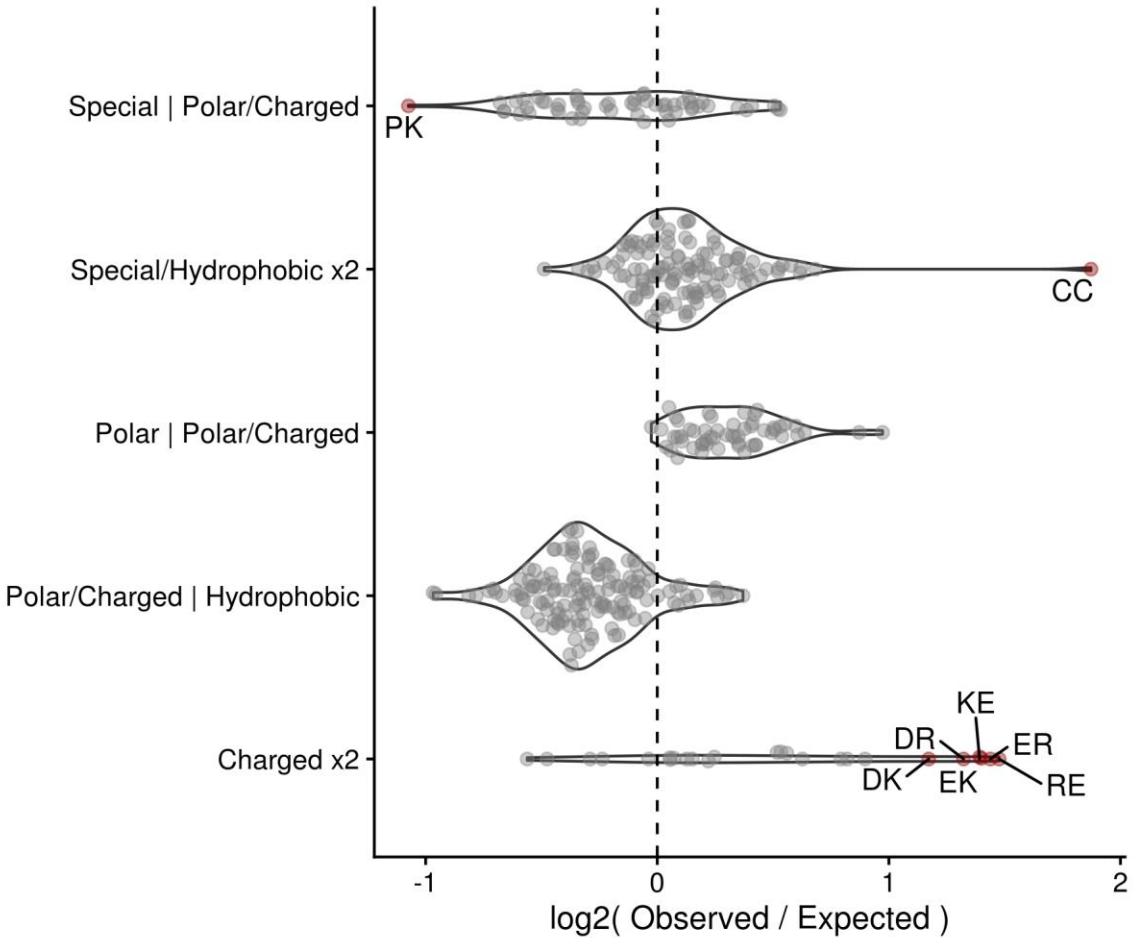


Figure 3: Amino acid frequencies by beta hairpin residue position

Bars indicate average amino acid frequencies for each amino acid at a given position across all beta hairpin structures. N-term and C-term refer to the N- and C-terminal beta strands while T # denotes a turn region of a given length (e.g. T 3 indicates a three residue turn region). Pol refers to beta strands containing a polar face adjacent to the turn region, Hydro denotes a hydrophobic face at this position. Beta strand residues are numbered from the turn region, with residue 1 representing the residue closest to the turn. Turn residues are numbered from N-terminal (residue 1) to C-terminal.



471

472 **Figure 4: Grouped differences in observed vs. expected residue contacts**

473 Dots represent individual contacting pairs with red, labelled dots indicating contacts that are

474 enriched or depleted at least two-fold vs. expected values. Residues are grouped as follows: Special

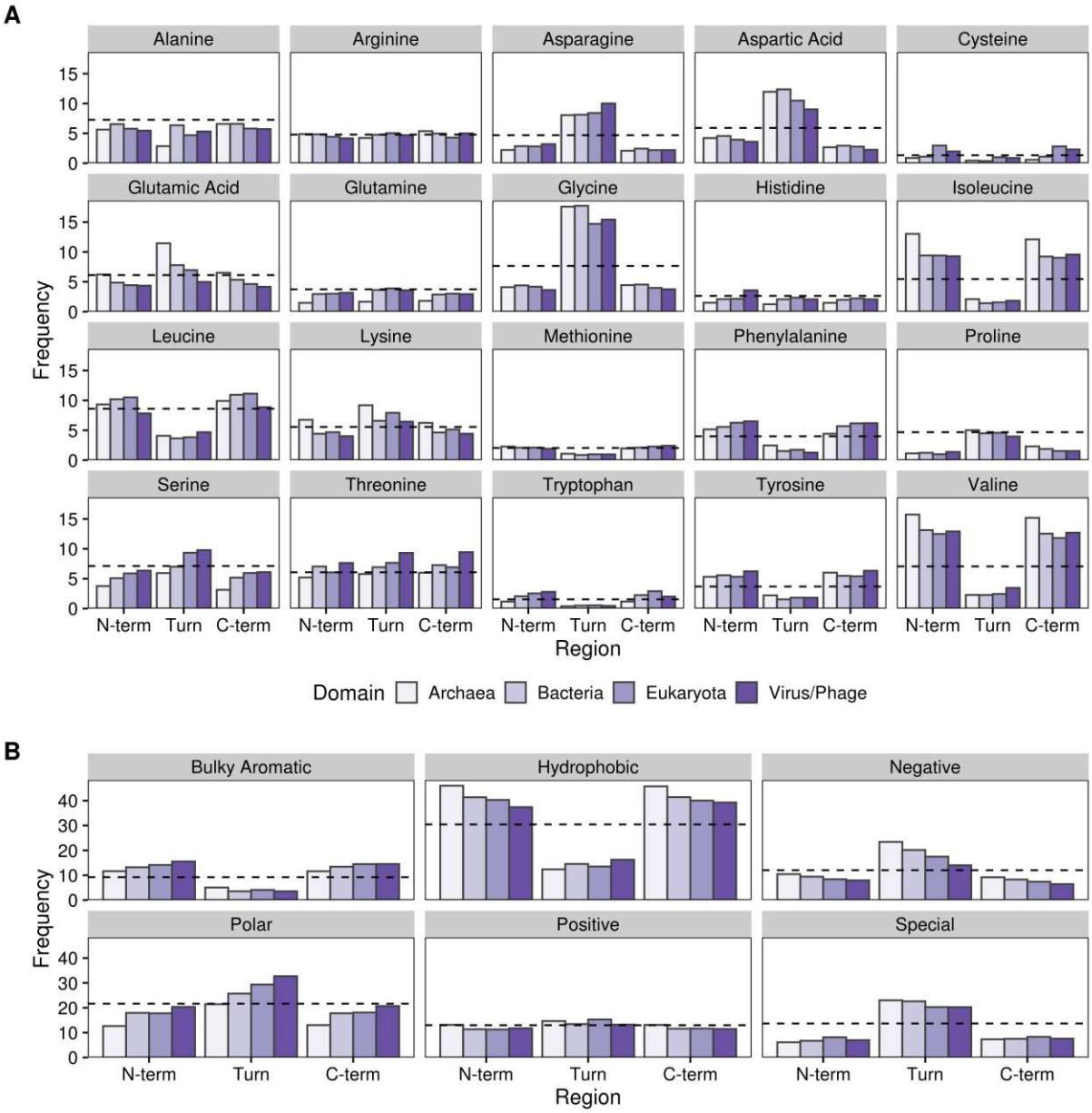
475 refers to cysteine, proline, glycine; Hydrophobic refers to valine, leucine, isoleucine, methionine,

476 alanine, tryptophan, tyrosine, phenylalanine; Polar refers to glutamine, threonine, serine,

477 asparagine; Charged refers to arginine, histidine, lysine, aspartic acid, and glutamic acid.

478

479 **Supplementary Material**



480

481 **Supplementary Figure 1: Amino acid frequencies by beta hairpin secondary structure region and**

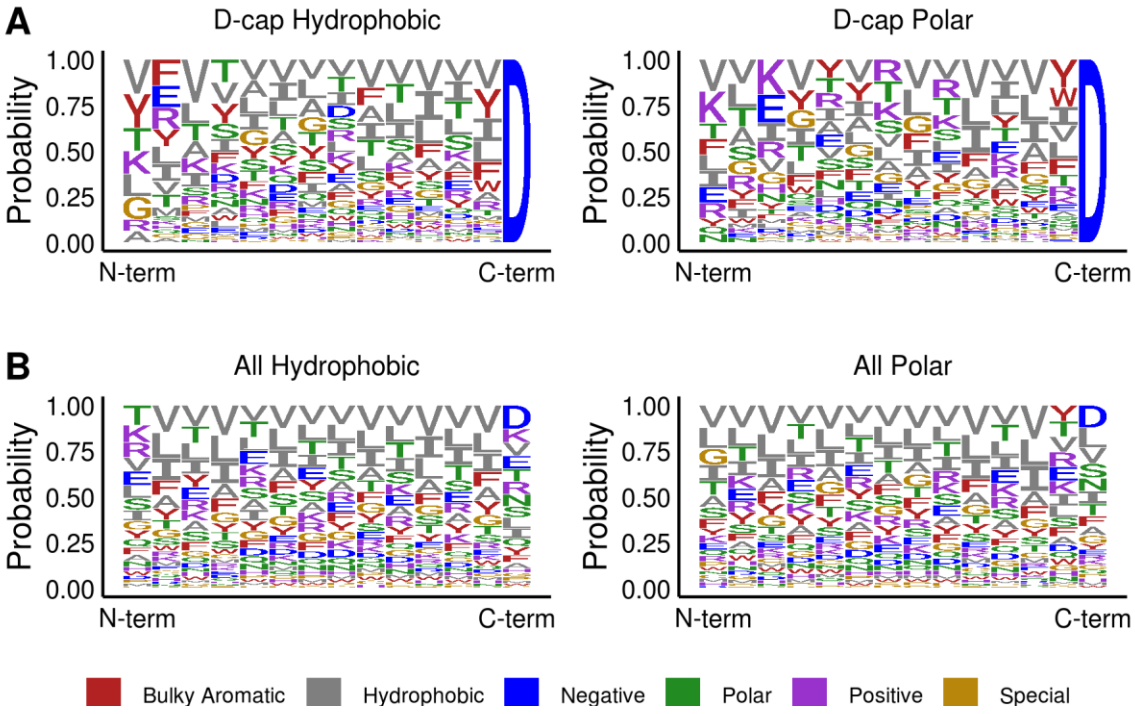
482 **organism domain.**

483 A) Bars indicate average amino acid frequencies for each amino acid within a given region of all

484 beta hairpins broken down by Domain of origin. The black dashed line indicates background amino

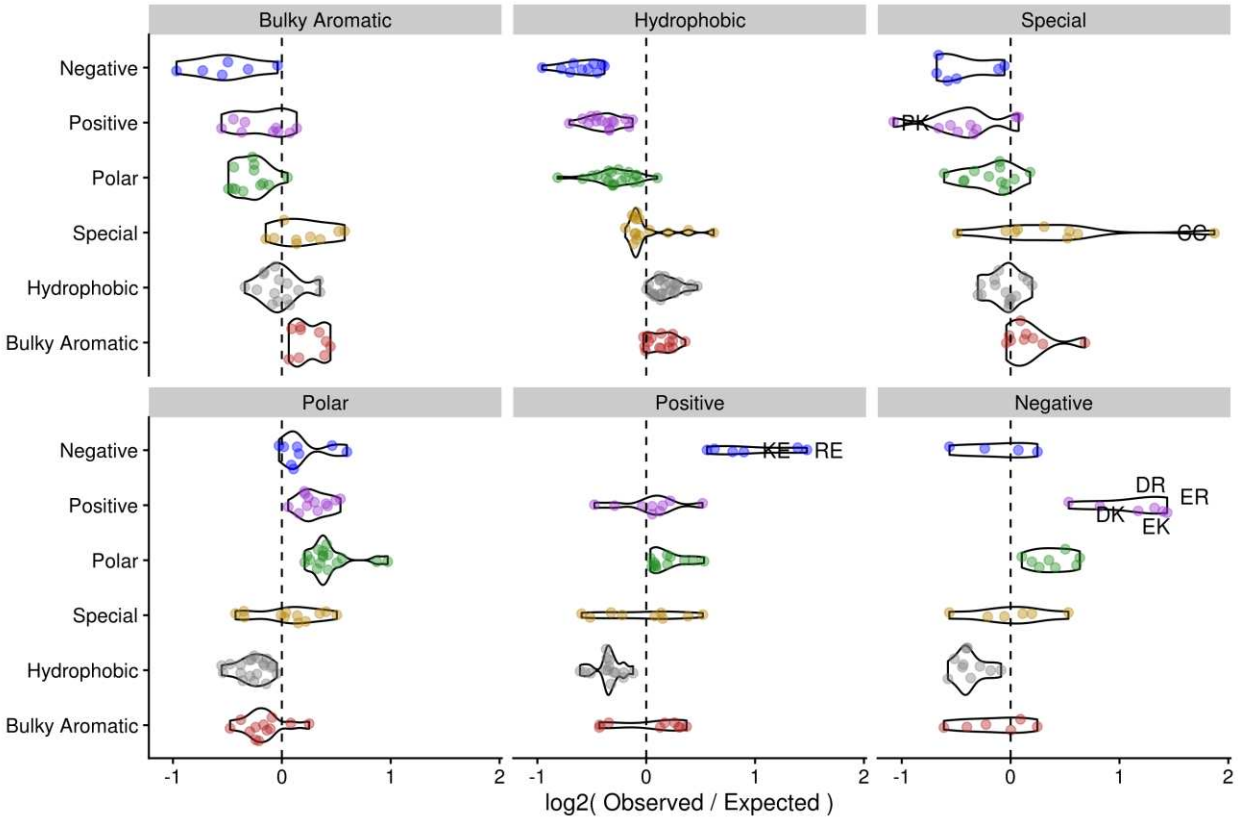
485 acid frequencies for all sites in all proteins containing the beta hairpin motif. N-term and C-term

refer to the N- and C-terminal beta strands while turn denotes the turn region. B) Same as in A) with amino acid frequencies grouped into specific classes. Amino acids are grouped as follows: Bulky Aromatic – phenylalanine, tryptophan, tyrosine; Hydrophobic – alanine, isoleucine, leucine, methionine, valine; Negative – aspartic acid, glutamic acid; Polar – asparagine, glutamine, serine, threonine; Positive – arginine, histidine, lysine; Special – cysteine, glycine, proline.



Supplementary Figure 2: Motif logo of amino acid probabilities in hairpin beta strands.

Amino acid letters are scaled to represent relative frequency at a given position in both beta strands, with more common amino acids listed first vertically. Color indicates the type of amino acid. A). Logo motif for hairpin beta strands with a C-terminal aspartic acid that start with either a hydrophobic (left) or polar (right) face relative to the turn region. B). As in A) but showing frequencies across all hairpin beta strands, not just those terminated with an aspartic acid residue. Amino acids are grouped as in Supp. Fig. 1.



501

502 **Supplementary Figure 3: Differences in observed vs. expected residue contacts**

503 Dots represent individual contacting pairs with labelled dots indicating contacts that are enriched

504 or depleted at least two-fold vs. expected values. Y-axis labels denote the class of the first residue

505 in the pair while facet headings denote the class of the second amino acid (e.g. the dot for DP

506 would be in the Negative row of the Special column). Amino acids are grouped as in Supp. Fig. 1.

507