Analysis of the sequence and structural features of the left-handed β-helical fold

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · OCTOBER 2008

Impact Factor: 2.63 · DOI: 10.1002/prot.22051 · Source: PubMed

4 AUTHORS, INCLUDING:

Cedric Govaerts
Université Libre de Bruxelles
57 PUBLICATIONS **3,036** CITATIONS
SEE PROFILE

Fred E Cohen
University of California, San Francisco
247 PUBLICATIONS **29,360** CITATIONS
SEE PROFILE

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Analysis of the sequence and structural features of the left-handed β-helical fold

Jay H. Choi,[1] Cedric Govaerts,[2] Barnaby C. H. May,[3,4] and Fred E. Cohen[1,4,5]*

[1] Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158
[2] Structure et Fonction des Membranes Biologiques, Universite Libre de Bruxelles, Bruxelles, Belgium
[3] Department of Neurology, University of California, San Francisco, California 94158
[4] Institute for Neurodegenerative Diseases, University of California, San Francisco, California 94158
[5] Department of Biochemistry and Biophysics, University of California, San Francisco, California 94158

ABSTRACT

The left-handed parallel β-helix (LβH) is a structurally repetitive, highly regular, and symmetrical fold formed by coiling of elongated β-sheets into helical "rungs." This canonical fold has recently received interest as a possible solution to the fibril structure of amyloid and as a building block of self-assembled nanotubular structures. In light of this interest, we aimed to understand the structural requirements of the LβH fold. We first sought to determine the sequence characteristics of the repeats by analyzing known structures to identify positional preferences of specific residues types. We then used molecular dynamics simulations to demonstrate the stabilizing effect of successive rungs and the hydrophobic core of the LβH. We show that a two-rung structure is the minimally stable LβH structure. In addition, we defined the structure-based sequence preference of the LβH and undertook a genome-wide sequence search to determine the prevalence of this unique protein fold. This profile-based LβH search algorithm predicted a large fraction of LβH proteins from microbial origins. However, the relative number of predicted LβH proteins per specie was approximately equal across the genomes from prokaryotes to eukaryotes.

Proteins 2008; 73:150–160.

Key words: parallel β-helix; left-handed parallel β-helix; hexapeptide motif; antifreeze protein; amyloid; fibrils; nanostructure design.

INTRODUCTION

The parallel β-helix is a repetitive fold where the repeating unit is a β-helical coil formed by segments of β-strand.[1–3] With few exceptions, both the right-handed β-helix (RβH) and left-handed parallel β-helix (LβH) share common structural features. Each rung of the canonical β-helix consists of two to three β-strands interrupted by turn or loop regions.[4] The β-helical rungs are aligned to form a cross-β structure such that elongated β-strands connected by hydrogen bonds lie parallel to the helical axis.[4,5] Structural repetition of coils creates a cylindrical hydrophobic core. The hydrophobic core of the β-helical proteins is characterized by buried stacks of similar side chains.[1] While RβH is generally characterized by β-strands connected by variable length of turns and loops, LβH is more rigid and repetitive than the RβH variant.[2]

Since the first crystal structure of a LβH protein was determined: UDP-*N*-acetylglucosamine acyltransferase, LpxA, from *E. coli*,[6] the structures of nine different proteins and their homologs have been reported to contain the LβH fold. To date, all known LβH structures are bacterial in origin and share a similar transferase activity.[1,3] One exception is the antifreeze protein from the spruce budworm.[1,7–9] All known bacterial LβH folds (referred to as a type-I LβH) have six residues per strand (18 residues per rung), described as an imperfect repeating hexapeptide motif, [LIV]-[GAED]-X_2-[STAV]-X. The smaller LβH structure of spruce budworm antifreeze protein (referred to as a type-II LβH) consists of five residues per strand (15 residues per rung). Type-I LβH and type-II LβH folds share a similar basic architecture and structural pattern. Each rung (or coil) of the canonical β-helix consists of three flat and untwisted parallel β-strands connected by either a one- or two-residue turn or a long external loop region.[1–3,10] In this study, loops are defined as a stretch of sequence containing more than two residues whose backbone alignment deviates from the normal LβH turn. Loop regions are a unique feature of the type-I LβH fold and occur uniformly across all known type-I LβH structures. The largest LβH domain identified to date is from the bifunctional *N*-acetylglucosamine 1-phosphate uridyltransferase of *E. coli* and

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: National Institutes of Health; Grant number: AG21601.
*Correspondence to: Fred E. Cohen, UCSF MC 2240, Genentech Hall, Room N472J, 600 16th Street, San Francisco, CA 94158-2517. E-mail: cohen@cmpharm.ucsf.edu
Received 5 December 2007; Revised 14 February 2008; Accepted 19 February 2008
Published online 8 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22051

Streptococcus pneumoniae, and it contains approximately nine rungs with one external loop between seventh and eighth rung.[11–13] No external loops have been observed in the two isoforms of the type-II LβH antifreeze proteins.[1,7–9] Each β-stand contains small, uncharged residues (V, A, S, T, C) and conserved larger hydrophobic residues (L, I, V) that face the interior of the LβH to create a hydrophobic core.[1–3,5] These constraining requirements at interior positions of the β-strand are presumed to have limited sequence divergence of LβH proteins throughout evolution.[14]

Known type-I LβH proteins are composed of a single LβH domain that is capped at either the N- or C-terminus with α-helical domains. All known type-I LβH proteins exist as native trimers, with adjacent monomers interacting at the surface of the LβH domains. In contrast, known type-II LβH proteins exist as a single LβH domain protein where the likely functional unit is a monomer.[8,9] Although oligomerization is thought to increase the overall stability of LβH proteins, no experimental evidence has been provided.

Recent modeling studies have proposed β-helical type folds as possible solutions for the structure of misfolded proteins associated with prion and Huntington's diseases.[5,15–18] These studies have converged on the β-helical architecture because of the structural features that these folds share with the unresolved structure of the longer chain amyloids. For example, biophysical studies have indicated that the longer chain amyloids are β-rich with extensive cross-β structure, a feature that could be accounted for by end-to-end polymerization of β-helical-like subunits. The LβH fold is highly regular and symmetrical with little variability in shape and size over the length of the domain.[19] These features also have led to the suggestion that LβH fold may be used as a building block for nanotubular structures with application in nanotechnology.[19–21] However, the accuracy of modeling studies employing LβH folds in the fields of amyloid and nanotechnology research has been limited by the relative absence of information pertaining to the sequence and structural features of this relatively rare protein fold. In this study, we have reexamined nine available LβH structures with the aim of more rigorously defining the structural features of the LβH. We show that there are strict residue preferences at β-helix turn regions, in addition to the highly conserved hydrophobic residues at the β-helix core. Molecular dynamics (MD) simulations using simple LβH models confirm that backbone hydrogen bonds and the hydrophobic core provide structural stability to the LβH fold. The stability of the type-I LβH structure relative to the smaller type-II LβH structure provides a possible explanation for the abundance of type-I LβH over type-II LβH in the genomes. A survey on the currently available proteomic database shows the existence of the LβH fold in all genomes.

MATERIALS AND METHODS

Construction of LβH position-specific sequence profile

Eight of 14 known type-I LβH structures [PDB ID codes: 1LXA, 1TDT, 1KRR, 1OCX, 1XAT, 1HV9, 1QRE, 1SSQ] and one type-II LβH structure (1M8N) along with sequences from 12 homolog proteins were used to construct LβH sequence alignments and build a position-specific sequence logo. The three-dimensional structures of the LβH domains were superimposed using the InsightII software,[22] and the sequences that comprise only a complete LβH rung were extracted to build a multiple alignment of the LβH sequences. The rungs that contained external loop regions and rungs that are located at the top and bottom of the LβH domain were excluded since they contained abnormal sequence composition. For amino acid propensities of LβH turn regions, the sequences of type-II LβH turns observed in a subset of type-I LβH proteins were extracted and included in the sequence alignment of type-II LβH turns. The positions of the LβH were defined using the following nomenclature: T_1 (1st turn residue), T_2 (2nd turn residue), B_1^i (1st β-strand residue, facing inside), B_2^o (2nd β-strand residue, facing outside), B_3^i (3rd β-strand residue, facing inside), and B_4^o (4th β-strand residue, facing outside).

Model construction of simple LβH

The simple LβH models were built on the scaffold of UDP *N*-acetylglucosamine acyltransferase (1LXA) and spruce budworm antifreeze protein (1M8N). The amino acids and coordinates from residues 120 to 135, 120 to 155 of 1LXA (type-I LβH), and 35 to 64, 45 to 74 of 1M8N (type-II LβH) were used as templates to build simple models of LβH with different numbers of rungs and residues using the InsightII software.[22] While interior residues were kept as the original template, the outer residues were replaced with the commonly occurring residues in native structures in order to build a consistent set of models. Models of LβH-GLY, LβH-ALA, and LβH-VAL [see Fig. 3(A)] were built with residues at all B_3^i position of type-I LβH model substituted with glycine, alanine, and valine, respectively. The side chain positions of the LβH models were subsequently optimized by using SCWRL 3.1.[23] The models were optimized by energy minimization using the GROMACS 3.1.3 package.[24] The final models of simple LβHs contained only complete rungs of β-helix with no abnormal turns or external loops.

MD simulation of LβH models

All simulations were performed with the GROMACS software package,[24] using the GROMMOS 43a3 force

field.[25] Simple LβH models were solvated individually in cubic boxes filled with water molecules.[26] A Single Point Charge water model was used for the solvent molecules in the simulation.[27] Sodium ions and chlorine ions were used to electroneutralize the system. Solutes, solvent, and counterions were coupled independently to reference temperature baths at 300 K.[28] and the pressure was maintained by coupling the system weakly to an external pressure bath at one atmosphere.[16] Bond lengths were constrained by the LINCS procedure[29] and nonbonded interactions were evaluated using twin-range cutoff of 0.8 and 1.4 nm for Lennard–Jones and Coulomb potentials. The long-range electrostatic interactions beyond the cutoff were treated with the generalized reaction field model, using a dielectric constant of 54.[16] The integration time step was set to 0.002 ps and the trajectory coordinates and energies were stored at 0.5-ps intervals. The analysis was performed using the built-in programs of GROMACS software package.

Proteomic database and data preparation

The complete nonredundant proteomic sequences for all of the organisms examined in this study were obtained from the Universal Protein Knowledgebase (UniProt)[30] consortium database, and the corresponding taxonomic data were obtained from the National Center for Biotechnology Information (NCBI).[31] Although, the UniProt database is considered as a nonredundant database, it still contains redundancies because of the sequences of subspecies or strains. To reduce the redundancies caused by subspecies, all the sequences were reorganized based on taxonomic categories, and all the subspecie sequences within the same species were grouped together. Using the FASTA program,[32,33] the identical (>98% sequence identities) sequences within the same taxonomic group were removed.

Genome-wide search for LβH folds

The HMMER software package[34] was initially used to predict LβH folds across the genomes. From the multiple sequence alignments of the LβH domain, hidden Markov profiles were created using the hmmbuild program available as part of the HMMER package. The hmmsearch of HMMER was performed iteratively against the prepared data set (above) with E value < 0.1. The predicted domain sequences of each LβH candidate protein were grouped together and aligned against the LβH position-specific sequence profile to identify LβH features that included (1) the number of rungs, (2) residues that face toward the inner core of the LβH, (3) the external loop regions, and (4) residues at the β-helix turn regions. Each candidate protein was scored and filtered based on the number of rungs that it contains, inner core volume estimated from the interior residue side chain van der Waals volume, length, and the number of external loops, and the occurrence of prolines at β-helical turns. The volume of interior residues was calculated as the sum of side-chain volumes of the interior residues for each complete rung.

RESULTS

Position-dependent amino acid propensities of the LβH

To evaluate the amino acid propensities of the repeating feature of the LβH fold, sequence fragments of type-I LβH and type-II LβH rungs were extracted from the nine known crystal structures of LβH proteins. Incomplete rungs located at the top and bottom of the LβH domains were excluded from the analysis since these capping sequences often contained sequences that do not follow the usual sequence pattern associated with LβHs. The residues that were part of the external loop regions and/or deviated from the structural alignments were also excluded from the analysis. Figure 1 shows the resulting amino acid propensities at each position of type-I LβH and type-II LβH. As reported elsewhere, the residues at the B_1^i and B_3^i positions are oriented toward the core of the LβH, and as such are limited to uncharged and hydrophobic residues in the proteins characterized to date at atomic level. Additionally, given the limited available volume of the interior of the LβH turn, B_1^i residues are limited to smaller side chains (V, A, S, T, C). Residues at positions B_2^o and B_4^o face the outside of the β-helix domain and represent a wide range of residues, including charged to aromatic side chains.

Glycine and proline in LβH turns

The amino acid residue propensities of the β-helix turn regions of the LβH are summarized in Table I. The type-I LβH turn consists of four residues (B_4^o, T_1, T_2, and B_1^i) and often resembles the classical type-II β-turn while the type-II LβH turn consists of three residues (B_4^o, T, and B_1^i). Hydrogen bonds are often observed in type-I LβH turns between the backbone carbonyl oxygen of the B_4^o residue and amide hydrogen of the B_1^i residue in type-I LβH. This same hydrogen bonding was not observed in type-II LβH turns. According to the current analysis, both type-I LβH and type-II LβH turn regions contain a high incidence of glycine. Glycine is highly conserved at the B_4^o position in the type-I LβH turn, most likely because of steric constraints as the B_4^o side-chain orients toward the β-strand of the following rung.[2] Glycine also has a relatively high occurrence at type-I LβH T_2 position and at type-II LβH T position while almost no glycines were observed at any other positions.

The occurrence of proline was limited to position T_1 of the type-I LβH. Some prolines were observed at T_2

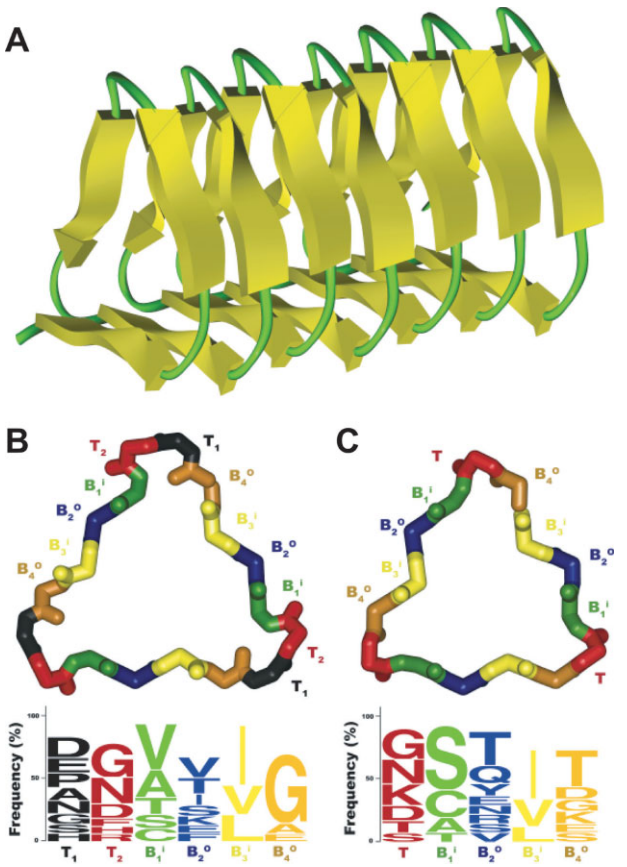


Figure 1

*Side and cross-sectional view of LβH structures: (**A**) The left-handed parallel β-helix (LβH) of 1HM9 (segment 260–379). (**B**) Type-I LβH, and (**C**) Type-II LβH, labeled T (turn) and B (β-sheet) with superscript "i" for a residue facing inside and "o" for a residue facing outside. The position dependent residue propensities are depicted as sequence logos. Percent amino acid residue frequencies >5% are shown.*

and B_4^o positions but only in cases where an external loop preceded the T_2 residue, thus allowing the necessary flexibility to accommodate proline or an incomplete rung located at the edge of the LβH domain. The location of proline in the LβH domain is also conserved with ~56% of prolines located at the top or bottom rungs of the LβH domain.

Structural stability of the LβH fold

While MD simulations are unable to calculate the absolute stability of a folded protein structure, they can be useful in studying the relative stability of related structures. We performed 10-ns MD simulations at 300 K on simple type-I LβH models, ranging from a 1-rung model (18 residues) to a 5-rung model (90 residues) [see Fig. 2(A)] in explicit solvent (for details, see Materials and Methods section). An analysis of positional root-mean-squared deviations (RMSD) relative to the starting models indicated that all the model systems had reached equilibrium after ~2 ns [Fig. 2(B)]. The average RMSD relative to the starting models were plotted against the number of rungs [Fig. 2(C)]. The relative stability of the LβH structures estimated by RMSD calculation showed a large stability difference between the 1- and 2-rung models. However, the subsequent addition of rungs to the 2-rung model to generate the 3-, 4-, and 5-rung models did not show significant additional stability contribution in the system. Table II summarizes the average RMSD calculated for C_α and all atoms, showing that the C_α RMSD measure was the sufficient measure for the structural deviation of LβH MD simulations. Secondary structure content was determined by the DSSP algorithm.[35]

Table I

Sequence Statistics of LβH Turn Regions[a]

A. Position-dependent amino acid residue propensities of type-I LβH and type-II LβH folds[b]

	Position in LβH rung			
Type I	B_4^o	T_1	T_2	B_1^i
	GLY 55%	ASP 16%	GLY 27%	VAL 36%
		PRO 12%	ASN 21%	ALA 22%
		GLU 12%	ASP 12%	THR 14%
		ASN 10%		SER 11%
		ALA 10%		CYS 10%
Type II	B_4^o	T	B_1^i	
	THR 29%	GLY 24%	SER 51%	
	GLY 10%	ASN 20%	CYS 22%	
	LYS 10%	LYS 17%	ALA 12%	
	ASP 10%	ASP 12%		
		THR 10%		

B. Distribution of proline and glycine[c] (%)

	Position in LβH rung					
Type I	T_1	T_2	B_1^i	B_2^o	B_3^i	B_4^o
Proline	100	0	0	0	0	0
Glycine	9	30	0	1	0	60
Type II	T	B_1^i	B_2^o	B_3^i	B_4^o	
Proline	0	0	0	0	0	
Glycine	71	0	0	0	29	

C. Location of prolines in β-helix domain[d]

β-helix rungs	72% (39/54)
Top and bottom rungs	56% (30/54)
Middle rungs	17% (9/54)
Loop region	28% (15/54)

[a]A and B were calculated based on 87 type-I LβH and 23 type-II LβH turn regions.
[b]Amino acid residue frequencies with >10% are listed.
[c]Percentage of proline and glycine residues observed in each position.
[d]Percentage calculated from 54 prolines that were found in β-helix domain of 8 type-I LβH proteins, including the incomplete rungs at top and bottom of type-I LβH domain.

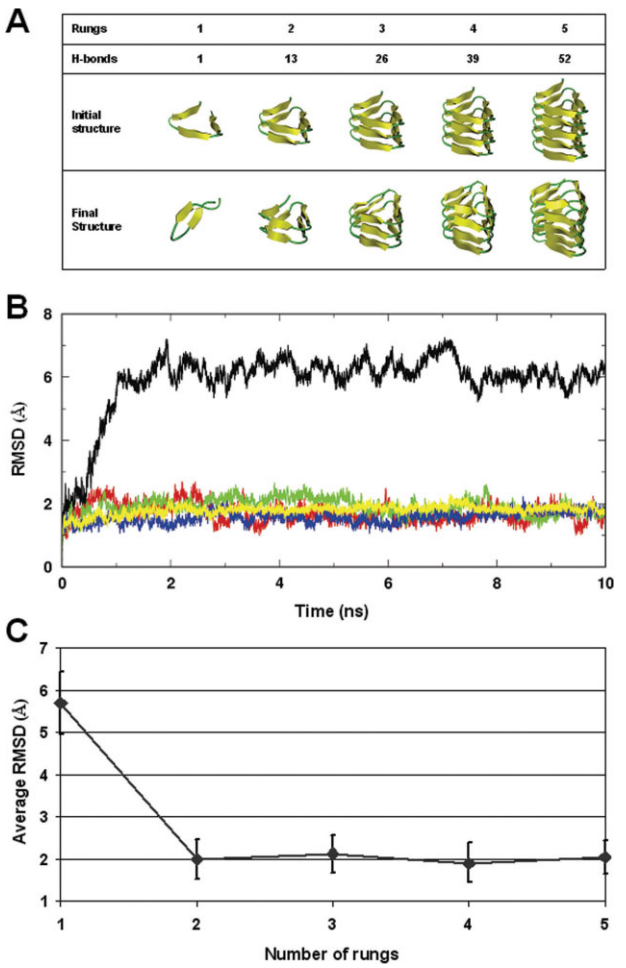


Figure 2

*MD simulation of LβH models with increasing numbers of rungs. (**A**) Initial structures and final structures after a 10-ns MD simulation. (**B**) Backbone RMSD of LβH models relative to their initial structures as a function of simulation time at 300 K. 1-, 2-, 3-, 4-, and 5-rung LβHs are shown in black, red, green, blue, and yellow, respectively. (**C**) Average backbone RMSD of the last 2-ns interval. Error bars are calculated from five independent simulations.*

The secondary structure content of a 1-rung LβH model after 10 ns showed an increase in random coil and a decrease in β-sheet content, providing an estimate of how structural elements changed over time. However, the DSSP algorithm was not sufficiently accurate to determine the secondary structure changes to the 3-, 4-, 5-rung LβH models. According to the DSSP algorithm, these models gained β-sheet content with a corresponding loss of turn content, while the actual structures were shown to be relatively well maintained by visual inspection and Ramachandran plot analysis (data not shown). Changes to the interstrand backbone–backbone hydrogen bonding network between initial and final structures provided another measure of structural deviation after the 10-ns MD simulations. While the 2-, 3-, 4-, and 5-rung LβH models maintained their hydrogen bonding network within the standard deviation, the 1-rung LβH model gained hydrogen bonds during the MD simulation as result of collapse of LβH core to form a two-stranded antiparallel β-sheet [Fig. 2(A)].

We sought to understand any positional contribution to the stability of the LβH fold by starting from a 1-rung LβH model (18 residues) and "growing" the LβH fold to a 2-rung LβH model (38 residues) in two-residue increments. The relative stability of individual models was measured by RMSD after a 10-ns MD simulation at 300 K. The starting 18-residue model (1-rung LβH model) contained a single hydrogen bond at the β-helix turn region between two adjacent β-strands. The incremental addition of two residues to yield the 20-residue model formed the first hydrogen bond between backbones of parallel β-strands. As expected, there was a notable stability gain with the creation of a complete rung. The subsequent addition of residues showed a gradual increase in stability as the number of hydrogen bonds between rungs increased (Supplementary Fig. 1).

The relative stability contribution of the LβH hydrophobic core was also explored using MD simulations. In order to conduct this calculation, interior residues (at B_3^i positions) of type-I LβH (LβH-WT) were substituted with glycine, alanine, or valine in order to build LβH-

Table II

Structural Statistics From the MD Simulation of LβH

A. Average RMSD with respect to the starting LβH structures[a] (Å)

	Backbone (C_α)	Backbone (all atoms)
1-rung LβH	5.87 (0.31)	6.78 (0.58)
2-rung LβH	1.90 (0.30)	2.73 (0.30)
3-rung LβH	2.06 (0.18)	2.85 (0.27)
4-rung LβH	1.76 (0.17)	2.47 (0.21)
5-rung LβH	1.95 (0.16)	2.67 (0.13)

B. Secondary structure elements content[b] (%)

	β-Sheet	Turn/Bend	Coil
LβH-WT	33	33	33
1-rung LβH	15 (16)	36 (8)	49 (14)
2-rung LβH	59 (8)	21 (7)	21 (5)
3-rung LβH	55 (5)	23 (4)	20 (3)
4-rung LβH	65 (11)	18 (6)	17 (6)
5-rung LβH	81 (11)	8 (5)	11 (7)

C. Interstrand backbone–backbone hydrogen bonds[c]

	1 rung	2 rung	3 rung	4 rung	5 rung
LβH-WT	1	13	26	39	52
LβH-MD	4 (1)	13 (1)	23 (4)	36 (3)	49 (2)

The analysis was performed on results of five independent trials MD. Average values are reported with standard deviations in parentheses.

[a]The backbone RMSD values were calculated with respect to the initial structures, averaged over the 8–10 ns interval of the MD trajectories.

[b]The secondary structure content was calculated using the DSSP algorithm.

[c]Backbone–backbone hydrogen bonds between strands were calculated using InsightII software.

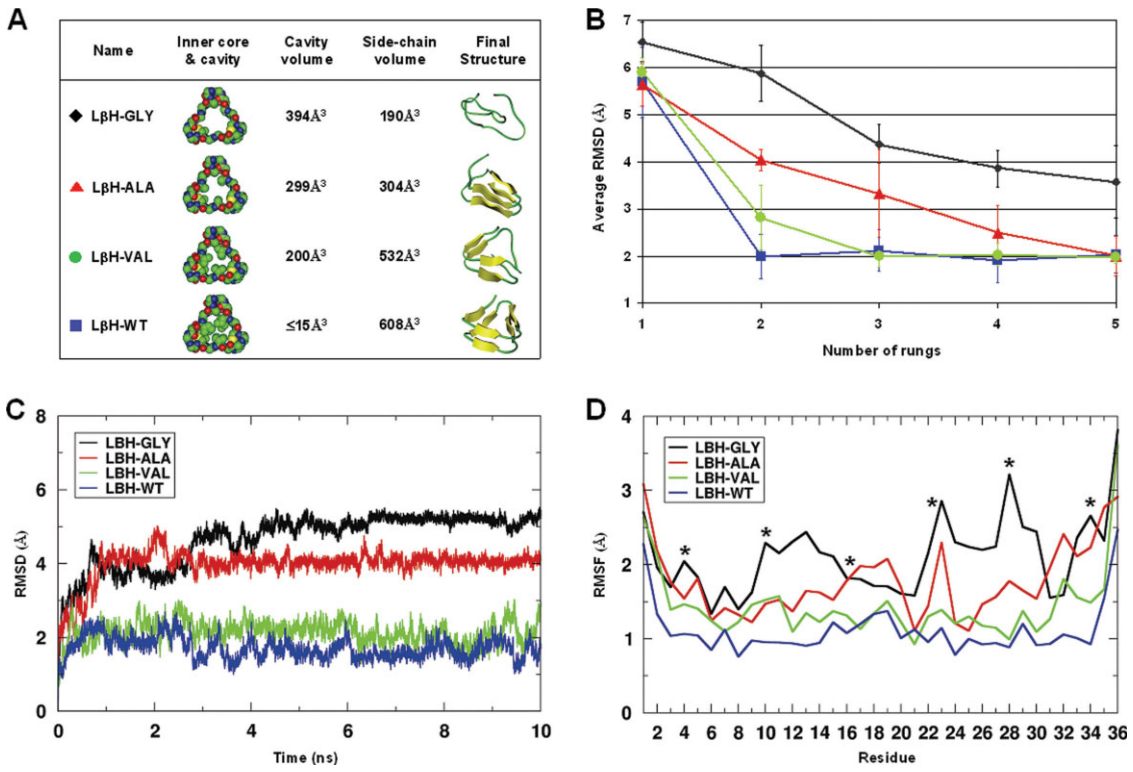


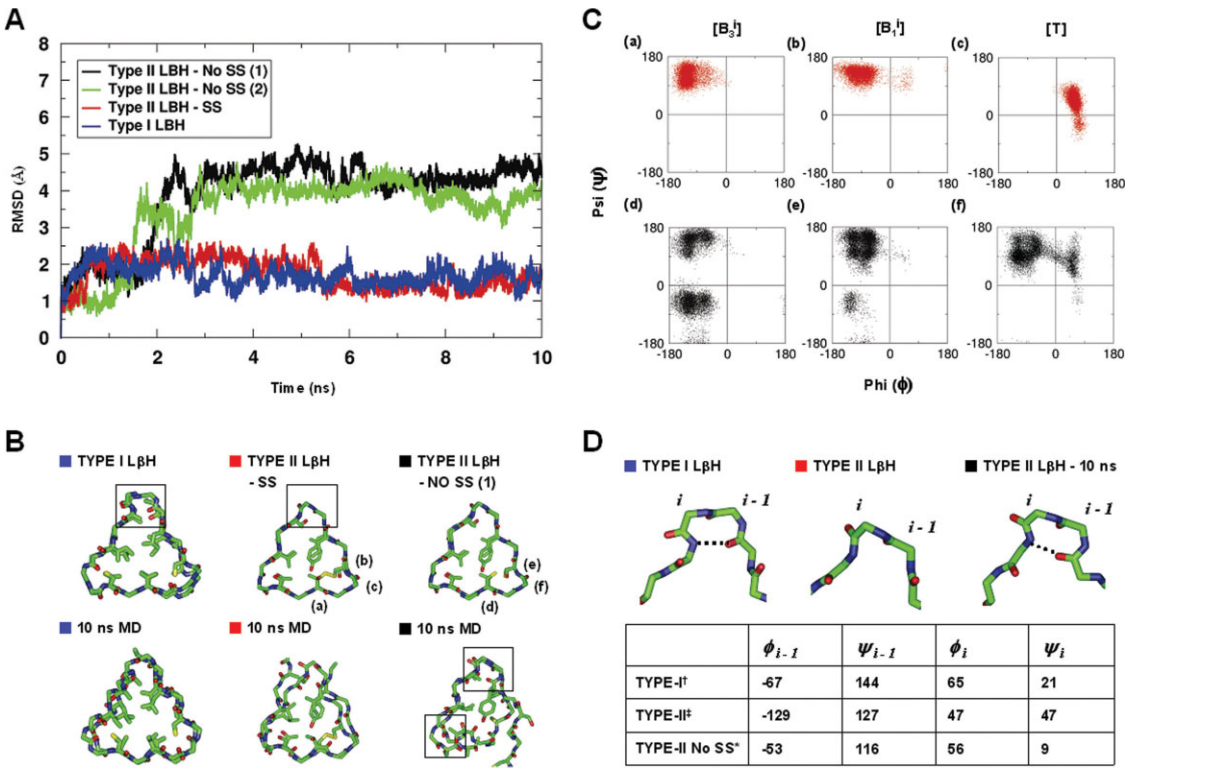
Figure 3

*MD simulations of LβH models with varying cavity sizes. (**A**) LβH-GLY, LβH-ALA, LβH-VAL models with residues at B_3^i positions substituted with glycine, alanine, and valine, respectively. LβH-WT is the wild-type LβH model based on 1LXA (residues 120–155). For 2-rung LβH models, the cavity volumes of the interiors were calculated by CASTp[36] and the side-chain volumes were calculated as the sum of side-chain volumes of the interior residues. Side-chain volumes were defined as the van der Waals volume of individual amino acids minus the volume of glycine.[37] Final structures of LβH models after 10-ns MD simulation at 300 K are shown. (**B**) Average backbone RMSDs of the last 2 ns interval from five independent trials of MD simulation for 10 ns at 300 K are plotted for each model. (**C**) Backbone RMSD and (**D**) residue-based Cα RMSF of 2-rung LβH models for LβH-GLY, LβH-ALA, LβH-VAL, and LβH-WT are shown in black, red, green, and blue, respectively. (*) indicates the interior residues at position B_3^i.*

GLY, LβH-ALA, and LβH-VAL models, respectively [Fig. 3(A)]. The cavity volumes of LβH-GLY, LβH-ALA, LβH-VAL and LβH-WT were estimated by CASTp.[36] Conversely, the side chain volume of the interior residues for each model was calculated as the sum of the van der Waals volume of individual amino acid side chains, minus the volume of glycine.[37] MD simulations of these models were performed at 300 K for 10 ns, and average RMSD were calculated based on the starting models [Fig. 3(B)]. LβH-VAL showed a significant decrease in RMSD values, indicative of a comparable stability relative to LβH-WT. The relatively high RMSD value of LβH-ALA indicated that the defective packing of the hydrophobic core significantly decreased the stability of the LβH structure. As shown in the final structure of LβH-ALA [Fig. 3(A)], the architecture of the β-helix fold transformed into a two-strand per rung structure resulting from the collapse of the β-helix core. However, the destabilizing effect of the hydrophobic core could be overcome by elongation of β-helix rungs such that the stability of a 5-rung LβH-ALA model was comparable to that of a 5-rung LβH-WT structure [Fig. 3(B)]. A 5-rung model of LβH-GLY showed relatively high RMSDs, indicating that glycine substitutions significantly decreased the structural stability of the LβH because of the hydrophobic core defect and excessive backbone flexibility. An examination of root-mean-square fluctuation (RMSF) of Cα before and after MD simulations indicated significant movement at core residues of the LβH-GLY model [Fig. 3(D)].

Stability of type-I LβH vs. type-II LβH

The relative stability of type-I LβH and type-II LβH was examined by MD simulations using 2-rung models. Unlike type-I LβH, known type-II LβH proteins contain disulfide bonds between cysteines at positions B_3^i and B_1^i of successive rungs. The hydrophobic core of type-II LβH is less regularly packed with aromatic residues and disulphides inside the β-helix than that of type-I LβH. The stability of a 2-rung model of a type-II LβH with a B_3^i–B_1^i disulfide bridge was compared with a 2-rung



	ϕ_{i-1}	ψ_{i-1}	ϕ_i	ψ_i
TYPE-I†	-67	144	65	21
TYPE-II‡	-129	127	47	47
TYPE-II No SS*	-53	116	56	9

Figure 4

*Stability analysis of Type-I LβH vs. Type-II LβH. (**A**) 10-ns MD simulations of 2-rung models of type-I LβH and type-II LβH with and without the disulfide bridge at 300 K. Two types of type-II LβH without the disulfide bridge were examined. Type-II LβH - No SS (1) model used two rungs of 1M8N lacking the disulfide bridge. Type-II LβH - No SS (2) model used the same template as the Type-II LβH - SS model, but the disulfide bridge was removed by substituting a cysteine residue with a serine residue at position B_3^i. (**B**) Initial and final structure of type-I LβH, type-II LβH with and without the disulfide bridge. (**C**) Ramachandran plot analysis showing the distribution of the main chain dihedral angles of three-residue positions: residues at position B_3^i (a and d) and B_1^i (b and e), and residue at position T (c and f). Type-II LβH with and without the disulfide bridge was shown in red and black, respectively. (**D**) Turn regions of Type-I LβH, type-II LβH, and type-II LβH without the disulfide bridge after 10-ns MD simulation. †Values were taken from Ref. 2. ‡The mean values of dihedral angles were calculated from 15 type-II LβH turn regions of 1M8N. *The mean values were calculated from the final structures of MD from five independent trials.*

model of type-I LβH using MD simulation (see Fig. 4). The disulfide bridged type-II LβH model showed comparable stability to the type-I LβH model in 10-ns MD simulation. To understand the stability contribution made by the disulfide bridge to the type-II LβH, two rungs of type-II LβH that did not contain disulfide bonds were compared with a 2-rung type-I LβH model and a 2-rung disulfide bridged type-II LβH model using MD simulation. Figure 4(A) shows the RMSD analysis of the three model systems. The Cα atoms of the type-I LβH rungs and disulfide bridged type-II LβH rungs after 10-ns MD simulation remained within 2 Å of the starting structure, while the Cα atoms of the type-II LβH rungs with no disulfide bridges deviated from the initial position by more than 4 Å and lost β-helical architecture. Figure 4(B,C) show the difference in movements of backbone and backbone torsion angles for residues located at the corner involved with the disulfide bridge. Large movements of the backbone Cαs were observed in the type-II LβH lacking the disulphide bridge. In addition, an examination of the 10-ns structure of the type-II LβH lacking the disulfide bridge showed that some of type-II LβH turns transformed into a type-I-LβH-like turn, with hydrogen bonds observed between the amide hydrogen of the B_1^i residue and backbone carbonyl oxygen of a residue located two residues upstream from the B_1^i residue [Fig. 4(D)]. Our findings suggest that: (1) the disulfide bonding seen in type-II LβH is a major contributor to the stability of this fold; (2) in the absence of a disulfide bridge, the type-II LβH is significantly less stable than the type-I LβH. These findings provide a reasonable explanation for the prevalence of type-I LβH over type-II LβH proteins.

Distribution and prevalence of the LβH fold

To date, there are only nine LβH proteins of known structure, thus limiting the amount of available proteomic data with which to assess the prevalence and distribution of residues in this fold. A couple of human pro-

teins have been predicted to adopt the LβH fold,[38] and the Pfam database[39] categorizes a number of putative LβH proteins, based on the ubiquitous hexapeptide motif. Nonetheless, it remains unclear how frequently the LβH fold might occur in the genomes. In order to accurately predict the prevalence of the LβH fold in the genomes, the proteomic data from the UniProt database was reorganized based on NCBI taxonomic categories[31] and processed to remove all redundant sequence data using a pairwise comparison within the same specie group. A total of 144,549 redundant sequences were eliminated from 4,135,679 protein sequences using this approach.

The prediction of LβH folds was performed using amino acid sequence patterns as well as structural constraints. LβHs contain highly conserved residues that are constrained by alternating residue positions oriented toward the LβH core or the outside (see discussion above). This repeating motif intrinsically exists in all known type-I LβH structures. The sequence pattern of the type-II LβH is less obvious and dominated by the functional TXT motif of the spruce budworm antifreeze proteins, therefore making this fold more difficult to detect. The LβH sequence profiles were constructed by a multiple sequence alignment of sequence fragments of known LβH proteins, as described earlier. Using these LβH profiles, the revised UniProt proteomic data was examined using the hmmsearch program[34] to identify proteins with matching sequence patterns. At the end of each search round, proteins with E value < 0.1 were selected to build a new profile for a subsequent search round. The process was repeated iteratively until no new sequences were identified. For the type-I LβH, this iterative search converged after the fifth search round.

The candidate LβH proteins were subsequently examined using structural constraints based on the features of known LβHs. These criteria included: (1) the predicted LβH sequence should contain at least two predicted LβH rungs, the minimal stable unit; (2) the side chain volume of interior residues should not exceed 420 and 400 Å^3 per rung for type-I LβH and type-II LβH, respectively (maximum van der Waals side-chain volume calculated from known LβH protein and maximum core volume estimated by CASTp was ~382 and 420 Å^3 for LβH-I LβH and ~360 and 400 Å^3 for LβH-II LβH, respectively); (3) the location and length of the external loop should be consistent with known structures, such that no more than one external loop should occur in a single rung and the loop should range from 1 to 50 residues in length; (4) proline is only allowed at β-helix turn positions at T_1 and T_2. From the 5539 initial candidates identified by sequence pattern, 662 proteins were eliminated using these structural constraints, as shown in Figure 5(A). The remaining 4877 predicted LβH proteins were categorized by NCBI taxonomy data as shown in Figure 5(B), and represented as the percentage of LβH protein occurrences in the genomes. The crude estimate of LβH distribution in the genomes was determined by calculating the ratio of predicted LβH proteins to the total number of proteins in the proteomic data. These findings were consistent with the taxonomic distribution of known LβH structures in the Protein Data Bank (PDB),[40] and Pfam database records. However, estimating the average occurrence of predicted LβH proteins per species in each taxonomic category suggested that the actual taxonomic distribution of LβH might be largely equivalent across the genomes. In the case of type-II LβH protein prediction, no novel proteins were found

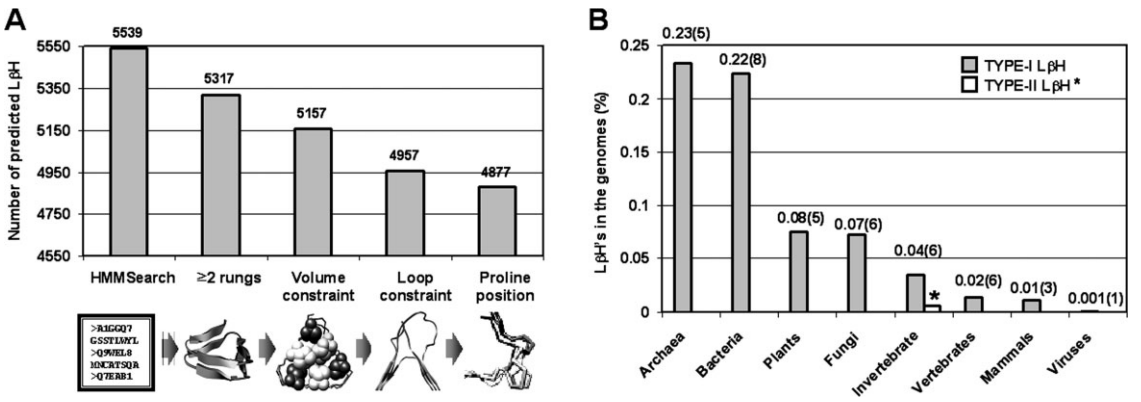


Figure 5

*Genome-wide search for LβH domains using sequence patterns and structural constraints. (**A**) Histogram of the total number of predicted LβHs based on sequence profile and the use of structural criteria to filter the predictions. (**B**) Percentage of predicted LβH over the total number of proteins in each division. Average numbers of LβH predicted per species in each division are shown in parentheses. *LβH type II was predicted in 38 antifreeze proteins of the insect spruce budworm and closely related species.*

other than the 38 homologs of spruce budworm antifreeze proteins.

LβH loop regions and domain size

In an attempt to further characterize structural components of the LβH fold, the external loop regions and LβH domain size were examined from known and predicted LβH proteins. For the predicted LβH proteins, the loop regions were established based on sequence content and location relative to the predicted β-helix domains. The predicted loops were categorized based on length. The occurrence of loops of various lengths was calculated, and plotted along with observed data from known LβH proteins [Fig. 6(A)]. The observed and predicted loop distribution was comparable with respect to loop length and occurrence. The domain size of the predicted LβH proteins was computed by calculating the number of rungs that each predicted LβH protein contains, as shown in Figure 6(B). Based on this distribution, the most abundant size of LβH domain was estimated to be four to six rungs. The analysis of the occurrence and distribution of loops and domain size, suggests that most LβH domains are composed of four to six consecutive LβH rungs, interrupted by external loops of variable length (1–40 residues).

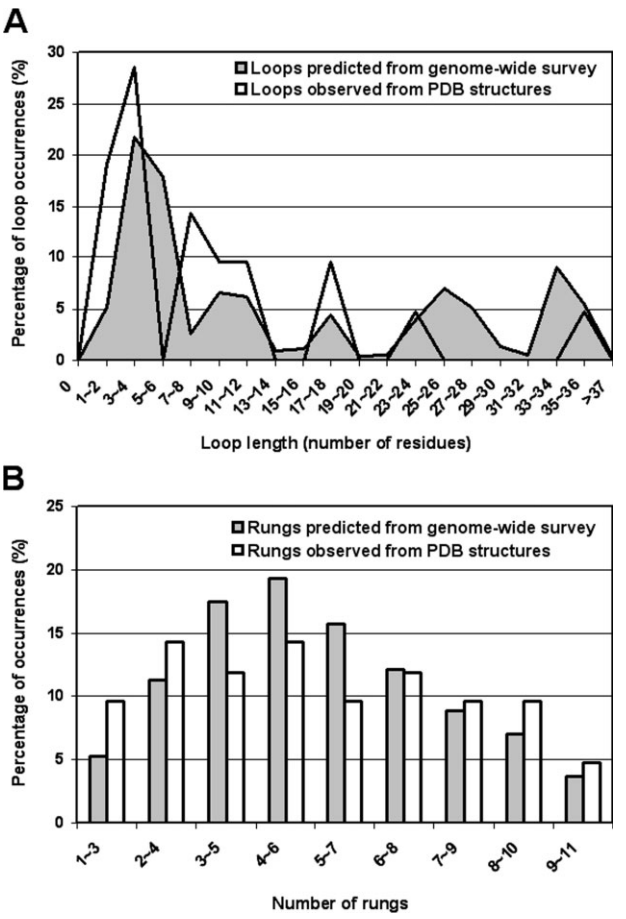


Figure 6

*The distribution of LβH loop regions and LβH domain. (**A**) Occurrences of LβH external loops and their sizes. Loop sizes and occurrences predicted from the genome-wide search (unfilled line) and observed from 14 PDB structures (filled line) were overlaid. (**B**) The size of predicted (shaded bars) and observed (open bars) LβH domains based on the number of rungs. Both graphs (A) and (B) were plotted on the x-axis of a sliding window interval because of difficulties of defining the start and end of the LβH domain.*

DISCUSSION

Residue propensities in LβH turns: glycine and proline

We have focused on the occurrence of glycine and proline residues to account for their possible roles in the LβH structure. β-Helix turn regions include a high propensity for glycine at position T_2 of type-I LβH and position T of type-II LβH. This feature can be explained by the left-handed α-helical conformation (α_L) adopted at these positions, allowing the β-strand to propagate in a new direction and hence facilitate folding of the β-helix. Proline also has a unique propensity at the β-helix turn regions in type-I LβH of being constrained to only the T_1 position. Since β-strand positions (B_1^i, B_2^o, B_3^i, and B_4^i) require backbone–backbone hydrogen bond formation with the β-strand of the following rungs, proline would undermine the structural stability of the fold. The extremely low occurrence of proline and glycine at other positions suggests that these substitutions are structurally prohibited. These unique residue propensities of LβH turns may be used to identify or distinguish the rung structures of type-I LβH and type-II LβH.

In addition to the positional propensity of proline in the LβH rung structure, we also observed a high occurrence of proline at either the top and bottom regions of β-helix domain. This feature may explain a possible role for prolines in folding of the LβH. The introduction of prolines at the edge of the β-helix domain may decrease conformational entropy, thus facilitating folding of the LβH domain or serve to terminate or cap the repetitive β-helical structure. In the case of type-II LβH, disulfide bridges formed by cysteines located inside of the β-helix were observed on the top and bottom regions of LβH domain. Their role in the stability and folding of LβH may be similar to that of prolines in type-I LβHs. It can be speculated that the folding process of an LβH may not be thermodynamically favorable without these structural features to restrict conformational flexibility.

Structural stability of the LβH

Key features of the LβH fold that have received interest in modeling studies of amyloid fibril and nanotubular

structures are the inherent compactness of the fold and the availability of accessible β-faces to initiate and sustain self-assembly. Previous modeling studies have suggested that the 2-rung structure of the LβH (~36 residues) was the minimal unit required to create a stable nuclei for seeded fibrilization of β-helices,[5,16] and it has also been used as a building block in the design of the nanotubular fibers.[19–21]

The structural stability of the LβH and the role of hydrogen bonds and the hydrophobic core have been previously reviewed.[1,2] This report systematically examined the changes in stability as a result of elongation of the β-helix rungs and defects to the hydrophobic core. We have shown that the 2-rung LβH is likely to be the minimal stable unit. An essential determinant of global stability was shown to be the hydrophobic core, such that incorrect packing, or packing defects rendered the 2-rung LβH unstable. Based on the relative stability comparison between two-rung models of type-I LβH and type-II LβH, it is conceivable that type-I LβH is a more sensible fit as a building block of self-assembling structures.

Genomic prevalence of the LβH fold

To date, there are no known human or mammalian proteins that have been documented to incorporate a LβH fold. The largest proportion of predicted type-I LβH proteins are of bacterial or microbial origins. It is an intriguing question as to whether LβH-like protein folds exist in other genomes given that LβH-like structures have been proposed as models of misfolded human proteins associated with disease. Initial speculation was that since all known type-I LβHs were microbial in origins, and the misfolded form of disease related proteins is toxic to mammalian cellular environments, the LβH fold may only exist in a narrow-range of species that can tolerate its unique structural feature. Our genome-wide survey of LβH suggests that the LβH fold probably exists evenly across the genomes, rather than occurring in a narrow range of species or taxonomic categories. Our analysis predicted the human proteins, dynactins p25 and p27 subunits, eukaryotic translation initiation factor 2B epsilon and gamma subunits, and GDP-mannose pyrophosphorylase A and B, as containing an LβH domain. These results are consistent with other predictions of human LβH proteins.[38] However, to date, these structural predictions have not been confirmed experimentally. The notion that LβH proteins exist in the human proteome may support the idea that the LβH fold could serve as a building block for the fibrilization of misfolded proteins associated with human disease (e.g. prion and Huntington's disease). However, experimental evidence supporting the fibrilization of the β-helix fold is limited.[41] If it is assumed that the fibrilization of LβH domains is possible, then native soluble LβH proteins may avoid this outcome by the incorporation of structural elements to restrict self-assembly or their ability to self-assemble is an essential part of their function. For example, all known type-I LβH proteins contain α-helical domains at either the N- or C-terminal of the LβH domains. If self-assembly of the LβH fold were possible via exposed β-faces, these α-helical domains would cap propagation and hence inhibit fibrilization. Further experimental studies are required to elucidate the self-assembly of the LβH fold and define a possible role in human misfolding diseases and nanotechnology.

CONCLUSIONS

Recent modeling studies have suggested that LβH-like folds may be possible structural solutions to the misfolded isoforms of proteins involved in neurodegenerative diseases. However, it is not clear whether disease-related proteins such as prion proteins or polyglutamine rich sequences can adopt LβH-like folds in a relatively dehydrated state since they do not include the distinctive patterns that are found in known LβH proteins. Since the LβH hydrophobic core is critical to the structural stability of the fold, the inclusion of alanine- and glycine-rich sequences that are a feature of some amyloidogenic proteins does not seem compatible with the LβH architecture. While some theoretical studies proposed the possibility of hydrophilic or charged residue inclusions in the LβH core, such as models of polyglutamines, further studies that incorporate the impact of the dehydrated state will be necessary to probe the structural compatibility of LβH with those residues that are unprecedented in documented structures.

ACKNOWLEDGMENTS

The authors thank Dr. Roland Dunbrack, Jr., for assistance with SCWRL, Jerome Nilmeier for MD simulation, and Dr. Elaine Meng for graphic assistance, and Sarit Helman and Dr. Holger Wille for critically reading the manuscript.

REFERENCES

1. Jenkins J, Pickersgill R. The architecture of parallel beta-helices and related folds. Prog Biophys Mol Biol 2001;77:111–175.
2. Iengar P, Joshi NV, Balaram P. Conformational and sequence signatures in beta helix proteins. Structure 2006;14:529–542.
3. Kajava AV, Steven AC. Beta-rolls, beta-helices, and other beta-solenoid proteins. Adv Protein Chem 2006;73:55–96.
4. Simkovsky R, King J. An elongated spine of buried core residues necessary for invivo folding of the parallel beta-helix of P22 tailspike adhesin. Proc Natl Acad Sci USA 2006;103:3575–3580.
5. Govaerts C, Wille H, Prusiner SB, Cohen FE. Evidence for assembly of prions with left-handed beta-helices into trimers. Proc Natl Acad Sci USA 2004;101:8342–8347.
6. Raetz CR, Roderick SL. A left-handed parallel beta helix in the structure of UDP-*N*-acetylglucosamine acyltransferase. Science 1995; 270:997–1000.

7. Graether SP, Kuiper MJ, Gagne SM, Walker VK, Jia Z, Sykes BD, Davies PL. Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. Nature 2000;406:325–328.
8. Leinala EK, Davies PL, Jia Z. Crystal structure of beta-helical antifreeze protein points to a general ice binding model. Structure 2002;10:619–627.
9. Leinala EK, Davies PL, Doucet D, Tyshenko MG, Walker VK, Jia Z. A beta-helical antifreeze protein isoform with increased activity. Structural and functional insights. J Biol Chem 2002;277:33349–33352.
10. Parisi G, Echave J. The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. BMC Evol Biol 2004;4:41.
11. Olsen LR, Roderick SL. Structure of the *Escherichia coli* GlmU pyrophosphorylase and acetyltransferase active sites. Biochemistry 2001; 40:1913–1921.
12. Brown K, Pompeo F, Dixon S, Mengin-Lecreulx D, Cambillau C, Bourne Y. Crystal structure of the bifunctional N-acetylglucosamine 1-phosphate uridyltransferase from *Escherichia coli*: a paradigm for the related pyrophosphorylase superfamily. EMBO J 1999;18:4096–4107.
13. Sulzenbacher G, Gal L, Peneff C, Fassy F, Bourne Y. Crystal structure of *Streptococcus pneumoniae N*-acetylglucosamine-1-phosphate uridyltransferase bound to acetyl-coenzyme A reveals a novel active site architecture. J Biol Chem 2001;276:11844–11851.
14. Parisi G, Echave J. Structural constraints and emergence of sequence patterns in protein evolution. Mol Biol Evol 2001;18:750–756.
15. Yang S, Levine H, Onuchic JN, Cox DL. Structure of infectious prions: stabilization by domain swapping. Faseb J 2005;19:1778–1782.
16. Langedijk JP, Fuentes G, Boshuizen R, Bonvin AM. Two-rung model of a left-handed beta-helix for prions explains species barrier and strain variation in transmissible spongiform encephalopathies. J Mol Biol 2006;360:907–920.
17. Merlino A, Esposito L, Vitagliano L. Polyglutamine repeats and beta-helix structure: molecular dynamics study. Proteins 2006;63: 918–927.
18. Stork M, Giese A, Kretzschmar HA, Tavan P. Molecular dynamics simulations indicate a possible role of parallel beta-helices in seeded aggregation of poly-Gln. Biophys J 2005;88:2442–2451.
19. Zheng J, Zanuy D, Haspel N, Tsai CJ, Aleman C, Nussinov R. Nanostructure design using protein building blocks enhanced by conformationally constrained synthetic residues. Biochemistry 2007;46:1205–1218.
20. Haspel N, Zanuy D, Zheng J, Aleman C, Wolfson H, Nussinov R. Changing the charge distribution of beta-helical-based nanostructures can provide the conditions for charge transfer. Biophys J 2007;93:245–253.
21. Haspel N, Zanuy D, Aleman C, Wolfson H, Nussinov R. De novo tubular nanostructure design based on self-assembly of beta-helical protein motifs. Structure 2006;14:1137–1148.
22. InsightII software. San Diego, CA: Accelrys, Inc. 2000.
23. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003; 12:2001–2014.
24. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem 2005;26:1701–1718.
25. Daura X, Mark AE, van Gunsteren WF. Parametrization of aliphatic CHn united atoms of GROMOS96 force field. J Comput Chem 1998;19:535–547.
26. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
27. Berendsen HJ, Postma JP, van Gunsteren WF, Hermans J. Interaction models for water in relation to protein hydration. In: Pullman B, editor. Intermolecular Forces. Dordrecht: Reidel Publishing Company; 1981. pp 331–342.
28. Berendsen HJ, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
29. Hess B, Bekker H, Hermans J, Berendsen HJ, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463–1472.
30. The Uniprot Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res 2007;35 (database issue):D193–D197.
31. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2000; 28:10–14.
32. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Meth Mol Biol 1994;24:307–331.
33. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 1994;25:365–389.
34. Eddy SR. Profile hidden Markov models. Bioinformatics 1998; 14:755–763.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
36. Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 2003;31:3352–3355.
37. Creighton TE. Proteins: structures and molecular properties. New York: Freeman; 1993.
38. Parisi G, Fornasari MS, Echave J. Dynactins p25 and p27 are predicted to adopt the LbetaH fold. FEBS Lett 2004;562:1–4.
39. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34 (database issue):D247–D251.
40. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
41. Schuler B, Rachel R, Seckler R. Formation of fibrous aggregates from a non-native intermediate: the isolated P22 tailspike beta-helix domain. J Biol Chem 1999;274:18589–18596.