Appendix A

In vivo complementation studies of Mycobacterium tuberculosis Chaperonin60s

A1.1 INTRODUCTION

Chaperonin60s are a class of molecular chaperones that mediate the correct folding of newly synthesized polypeptides in an ATP-dependent manner (Hartl, 1996). The GroE complex of *Escherichia coli*, encoding the chaperonins GroEL and GroES, has been shown to be essential for bacterial growth at all temperatures (Fayet *et al.*, 1989). The chaperonin GroEL is involved in the *de novo* folding of 10-30% of all the proteins in the bacterial cytosol. It is in the central cavity of GroEL that the unfolded polypeptides undergo productive folding. This central cavity has been shown to be essential for its role as a molecular chaperone (Weber *et al.*, 1998).

GroEL was first identified by the isolation of temperature sensitive *E. coli* mutants that were defective for the assembly of the head proteins of bacteriophage A. and T4 and the tail proteins of bacteriophage T5 and were also affected for cellular growth at elevated temperatures (Sternberg, 1973a; 1973b). One such mutant *groELA4* represents the E191G substitution in GroEL (Zielstra-Ryalls *et al.*, 1993). The mutant exhibits normal growth at 30°C however survival at 43°C requires the external contribution of a functional *groEL*-homologue. The strain thus becomes suitable for testing the ability of various genes to complement the Ts phenotype of the mutant.

Biochemical analysis of the *M. tuberculosis* Cpn60s showed that the proteins do not exist as tetradecamers, the canonical form of Cpn60s, but rather as dimers (Described in Chapter 4). Moreover, the activity of the *M. tuberculosis* Cpn60s when compared with *E. coli* GroEL showed that these proteins do not exhibit ATPase activity unlike GroEL. The proteins, however, were capable of complete prevention of substrate aggregation suggesting their role as ATP-independent molecular chaperones. Since the *M. tuberculosis* Cpn60s behaved in a manner much different from the usual chaperonins their role as chaperones was studied by *in vivo* complementation using the *groEL44* mutant strain.

A1.2 EXPERIMENTAL PROCEDURES

A1.2.1 Cloning of *E. coli and M. tuberculosis cpn60s*

The genes coding for the *E. coli groEL* and *M. tuberculosis cpn60.1* and *cpn60.2*, were PCR amplified and cloned in the low-copy-number plasmid vector, pCL1920, through an intermediate sub-cloning step in pBluescript (SK+). The previously described expression plasmid, pKKGLl, cosmid A10 and vector, pKY206 served as the templates for PCR amplification of *cpn60.1*, *cpn60.2* and *groEL*, respectively. The amplified products were ligated at the SmaI site of pBluescript (SK+) and the transformants in *E. coli* DH5α selected on the basis of blue-white selection. From the positive clones obtained the three genes were subcloned in the spectinomycin resistant (Sp^r) plasmid, pCL1920 (Fig. Al.l). Table Al.l describes the different primers used for PCR amplification and the designation of the different clones.

Table Al.l Primers used for cloning of *cpn60s* in pCL1920 Restriction endonuclease (RE) sites are highlighted as underlined sequences.

Primer name	*RE site*	*Primer sequence*	*Designation*
ECGLPCL.FOR	SalI	5' TACTATAGTCGACTATGGCAGCTAAAGACG 3'	pCLECGL
ECGLA2S.REV	SacI	5' ATATACGAGCTCTCATGCCGCCCATGCC 3^1	
GL1PCL.FOR	PstI	5'ATAGATCTGCAGGATGAGTAAGCTGATCGAAT ACG 3'	pCLGLl
GL1PCL.REV	BamHI	5' TACTATAGGATCCTCAGTGCGCGTGCCC 3'	
GL2PCL.FOR	HindIII	5' TAGATGAAGCTTAATGGCCAAGACAATTGCG3'	pCLGL2
GL2PCL.REV	SmaI	5' AGCTAGCCCGGGTCAGAAATCCATGCCACC 3'	

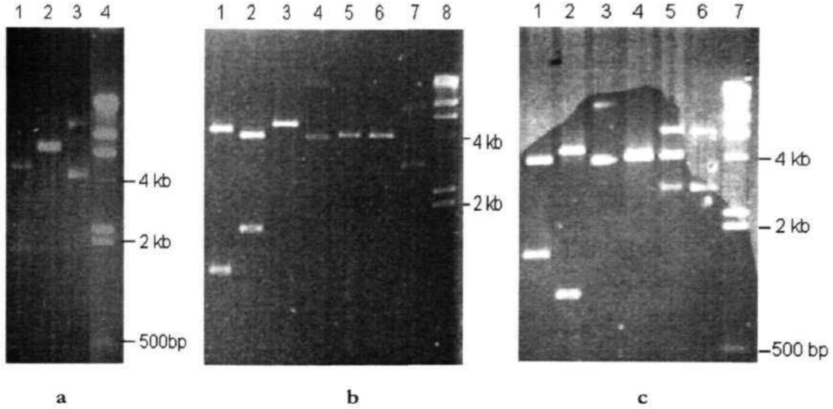

Figure Al.l Agarose gel scans showing cloning of *cpn60s* into pCL1920 (a) Cloning of *E. coli groEL* Lane 1: Sall/SacI digested clone showing the 1.6kb fall out of *groEL*. Lane 2: Positive clone linearized with BamHI. Lane 3: Undigested clone as a control. Lane4: λ-HindIII DNA digest as the marker. **(b) Cloning of *cpn60.1*** Lane 1: HindIII digestion of positive clone yielding the 1.1kb fall out. Lane *2:* BamHI/PstI digest giving the 1.6kb fall out confirming the positive clone. Lane 3: Positive clone linearized with SalI. Lanes 4 and 7 correspond to the undigested positive clone and pCL1920 as controls. Lanes 5 and 6 correspond to the linearized pCL1920 as control. Lane 8: X-HindIII DNA digest as marker. **(c) Cloning of *cpn60.2*** Lanes 1 and 4: SmaI/HindIII digested positive clone and pCL1920. The 1.6kb fall out confirms the clone. Lanes 2 and 5: SacI digested positive clone and the vector as a control. The positive clone releases the 1.1kb fall out. Lanes 3 and 6: Undigested positive clone and pCL1920 as controls. Lane 7: λ-HindIII DNA digest as marker.

A1.2.2 *In vivo* Complementation in *E. coli*

Complementation experiments were performed at 43°C as described by Chatellier *et al.* (1998). Plasmids pCLECGL, pCLGLl and pCLGL2 containing the *E. coli* and *M. tuberculosis cpn60* genes, along with the plasmid vector pCL1920 as negative control, were transformed into the temperature sensitive (Ts) *E. coli* strain, SV2 (kindly provided by Jean Chatellier and Alan Fersht). SV2 has the *groEL44* allele, representing an E191G substitution in GroEL (Zeilstra-Ryalls *et al.*, 1993). Sp^r transformants were selected at 30°C and the efficiencies of plating (EOP) for each determined at 43°C, relative to that at 30°C. 3ml cultures of appropriately transformed cells were grown overnight at 30°C in LB medium supplemented with spectinomycin. Cells were serially diluted 10-fold in LB medium and from each dilution 100µl aliquot was spread on LB plates containing spectinomycin. Similarly, 5µl from each 10-fold dilution was spotted onto two LB plates, one of which was incubated overnight at 30°C and the other at 43°C. The number of viable cells/ml of culture was deduced from the number of colonies obtained at the different dilutions.

A1.3 RESULTS

In order to investigate the functional properties of *M. tuberculosis* Cpn60.1 and Cpn60.2, we sought complementation of a Ts *grvEL E. coli* mutant, SV2. The results, presented in Table A 1.2, indicate that the SV2 derivative carrying the plasmid, pCLECGL (with *E. coli groEL*$^+$)exhibited an EOP of 0.05 at 43°C (relative to that at 30°C). The SV2 derivatives carrying pCLGLl and pCLGL2 (with *M. tuberculosis* genes for *cpn60.1* and *cpn60.2,* respectively) behaved like that carrying the vector pCL1920, with an EOP of $<10^{-5}$.

Table A1.2 Test for complementation of the Ts *E. coli groEL44*mutant Complementation was performed with plasmids expressing *E. coli groEL* or *M. tuberculosis cpn60.1* or *cpn60.2*

Plasmid (Description)	*Colony Forming Units*		*EOP (B/A)*
	30°C (A)	**43°C** (B)	
pCL1920 (Vector)	1×10^8	$< 10^3$	$< 10^{-5}$
pCLECGL **(*E. coli groEL*)**	1.3×10^8	6×10^6	0.05
pCLGLl **(*M.*** *tuberculosis* ***cpn60.1*)**	4×10^8	$< 10^3$	$< 10^{-5}$
pCLGL2 *(AT. tuberculosis* ***cpn60.2*)**	5×10^8	$< 10^3$	$< 10^{-5}$

Fig. A 1.2 clearly shows the ability of the *E. coli grvEL* to complement the mutant *grvEL* gene in the Ts mutant. The presence of neither *cpn60.1* nor *cpn60.2* could promote the growth of the Ts *groEL44 E. coli* mutant at 43°C.

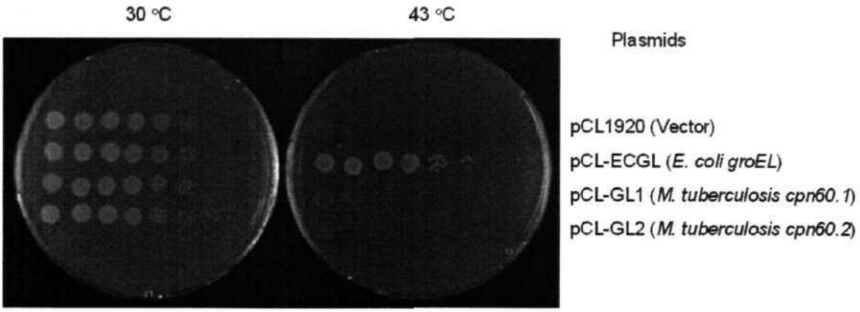
30 °C
43 °C
Plasmids
pCL1920 (Vector)
pCL-ECGL (*E. coli groEL*)
pCL-GL1 (*M. tuberculosis cpn60.1*)
pCL-GL2 (*M. tuberculosis cpn60.2*)

Figure A1.2 Test for complementation of the Ts *groEL44 E. coli* mutant Complementation was carried out with plasmids expressing *E. coli groEL* or *M. tuberculosis cpn60.1* or *cpn60.2.* Sp^r transformants of strain SV2 carrying plasmids as indicated at the right were grown overnight at 30°C in LB medium supplemented with spectinomycin, serially diluted and five microlitres from each 10-fold dilution spotted onto two LB-spectinomycin plates (from left to right in each row of the figure). One plate was incubated overnight at 30°C and the other at 43°C.

A1.4 CONCLUSIONS

The inability to achieve *in vivo* complementation of a Ts *groEL E. coli* mutant with the *M. tuberculosis cpn60* genes suggests that the mycobacterial counterparts of GroEL might have lost their function as molecular chaperones. Intriguingly, however, biochemical characterization of these proteins had suggested otherwise (Chapter 4). Thus, mere suppression of aggregation of substrate protein *in vitro* does not seem to impart the ability of GroEL-like function to the *M. tuberculosis* Cpn60s *in vivo.*

I propose that the anomaly between the *in vitro* and *in vivo* results can be explained by two possible means: (a) Cpn60s of *M. tuberculosis* require their cognate co-chaperonin for the role as molecular chaperones *in vivo* or (b) the two proteins, Cpn60.1 and Cpn60.2, are required simultaneously for effective *in vivo* complementation. In an attempt to resolve these questions, expression of the *M. tuberculosis cpn60.1* as a part of the *groESL* operon and co-expression of the *cpn60.1* and *cpn60.2* genes needs to be studied.

A1.5 REFERENCES

1. Chatellier, J., Hill, F., Lund, P.A., and Fersht, A. (1998). *In vivo* activities of GroEL minichaperones. *Proc. Natl. Acad. Sci. USA.* 95, 9861-9866.
2. Fayet, O., Ziegelhoffer, T., and Georgopoulos, C. (1989). The *groES* and *groEL* heat shock gene products of *Escherichia coli* are essential for bacterial growth at all temperatures./. *Bacteriol.* **171,** 1379-1385.
3. Hartl, F.U. (1996). Molecular chaperones in cellular protein folding. *Nature* **381,** 571-580.
4. Sternberg, N. (1973a). Properties of a mutant of *Escberichia coli* defective in bacteriophage *X* head formation (*groE*). I. Initial characterization. /. *Mol. Bio/.* 76, 1-23.
5. Sternberg, N. (1973b). Properties of a mutant of *Eschericbia coli* defective in bacteriophage *X* head formation *(groE).* II. The propagation of phage *X. J. Mol. Biol.* 76, 25-44.
6. Weber, F., Keppel, F., Georgopoulos, C, Hayer-Hartl, M.K., and Hartl, F.U. (1998). The oligomeric structure of GroEL/GroES is required for biologically significant chaperonin function in protein folding. *Nat. Struct. Biol.* 5, 977-985.
7. Zeilstra-Ryalls, J., Fayet, O., Baird, L., and Georgopolous, C. (1993). Sequence analysis and phenotypic characterization of *groEL* mutations that block λ and T4 bacteriophage growth./. *Bacteriol.* **175,** 1134-1143.

Appendix B

Identification of Conserved Residue Patterns in Small fi-barrel Proteins

This work has been published as:

Qamra, R., Taneja, B., and Mande, S. C. (2002). Identification of conserved residue patterns in small β–barrel proteins. *Protein Engg.* 15, 967-977

Protein Engineering vol.15 no. 12 pp.967-977, 2002

Identification of conserved residue patterns in small β-barrel proteins

Rohini Qamra, Bhupesh Taneja[1] and Shekhar C.Mande[2]

Centre for DNA Fingerprinting and Diagnostics, EC1L Road. Nacharam. Hyderabad 500 076, India

[1]Present address: Northwestern University. Chicago, IL. USA

[2]To whom correspondence should be addressed.
E-mail: shekhar@cdfd.org.in

Our abilities to predict three-dimensional conformation of a polypeptide, given its amino acid sequence, remain limited despite advances in structure analysis. Analysis of structures and sequences of protein families with similar secondary structural elements, but varying topologies, might help in addressing this problem. We have studied the small β-barrel class of proteins characterized by four strands *in* **= 4) and a shear number of 8** *(S* **= 8) to understand the principles of barrel formation. Multiple alignments of the various protein** sequences were **generated for the analysis. Positional entropy, as a measure of residue conservation, indicated conservation of non-polar** residues **at the core positions. The presence of a type II β-turn among the various barrel proteins considered was another strikingly invariant feature. A conserved glycyl-aspartyl dipeptide at the β-turn** appeared to be important in guiding **the** protein sequence **into the barrel fold.** Molecular **dynamics** simulations of **the** type II **β-turn peptide suggested that aspartate is a key residue in the folding of the protein sequence into the barrel. Our study** suggests **that the** conserved type **II P-turn and the non-polar** residues in **the barrel core are crucial for the folding of the protein's primary sequence into the β-barrel conformation.**

Keywords: β-barrel/molecular dynamics/protein folding/SH3/ type II β-turn

Introduction

The landmark work of Anfinsen indicated for the first time that the primary structure of a protein dictates its tertiary structure (Anfinsen, 1973). In a sequential protein folding model the primary structure of a protein initially yields α-helices, β-sheets and turns, the predominant secondary structural elements in proteins. By arranging these simple elements in precise patterns, complex protein structures assemble to achieve the diversity of protein functions. A major goal in understanding how the amino acid sequence of a protein specifies its structure is to understand how these elements of secondary structure are organized onto a tertiary scaffold. This requires learning how properties of individual amino acids are exploited in guiding an amino acid sequence into a particular fold. Much progress has been made in the last decade towards understanding the relationship between a protein's sequence and structure, yet the protein folding problem remains a captivating puzzle.

Researchers have learnt several rules governing the formation of helices and turns. However, the principles behind β-sheet formation are much less understood (Serrano, 2000). It is therefore especially intriguing to speculate how β-sheet proteins, having complex topologies and involving numerous contacts between residues distant in sequence, acquire their native structure. Several features of (5-sheet proteins have been suggested to be important for efficient folding and stability. The overall hydrophobic and polar pattern of amino acids may be a dominant driving force for defining a protein's topology (Eisenberg *et ai,* 1984; Bowie *et al.*, 1990; Kamtekar *et ai,* 1993). Recognition between amino acid side chains on neighboring β-strands may guide a correct strand register and hence stabilize the resulting β-sheets (Merkel *et al.*, 1999; Mandel-Gutfreund *et ai,* 2001). Another possibility is the formation of turns at critical locations in the protein structure. Turns may be particularly important for anti-parallel sheet formation and hence defining the protein topology. Supporting this hypothesis, recent studies indicate that the residues in the distal loop in the SH3 domain are important for nucleation of protein folding (Martinez and Serrano, 1999; Riddle *et al.,* 1999). Different combinations of these possible interactions are the most likely determinant of β-sheet topology and hence protein stability.

In the course of evolution, three-dimensional structures of proteins are conserved to a greater degree than their sequences, which determine their structure. Residue substitutions, which tend to destabilize a particular site, would probably be compensated by other substitutions that confer greater stability on the structure. For example, if volume conservation were important to structure and function, a substitution involving a reduction of volume in the protein core might result in a destabilizing pocket in the core. In this case, it might become necessary to substitute another residue at a position distant in the sequence but near in space. This second substitution should then have a larger side chain in order to conserve the overall volume of the core and therefore the overall folded structure. Thus, if structural compensation is a general phenomenon, neighbouring sites in the three-dimensional structure will tend to evolve in a correlated fashion owing to the compensation process. In the past decade there has been a great deal of progress in the development of methods for predicting interactions in protein structures by analysis of correlated changes in sequence evolution (Altschuh *et al.,* 1988; Shindyalov *et al,* 1994; Pollock and Taylor, 1997).

In this study, we have undertaken a comprehensive analysis of the sequence and structural variation seen in the small β-barrel proteins. A β-barrel is essentially identified by two geometric characteristics: the number of β-strands in the barrel (n) and the number of β-bridge staggers across the β-sheet (the shear number, *S)* (Murzin *et al.,* 1994). Within the all-(5 protein class in the Structural Classification of Proteins (SCOP) database there exist five folds which can be grouped together as small β-barrels (Murzin *et al.*, 1995). These barrels are characterized by the presence of four p-strands (n = 4) and a shear number of 8 *(S* = 8) (Murzin *et ai,* 1994). Although the five folds have similar secondary structure composition, each has a distinct topology. The goal of this study was to identify conserved features across these β-barrel folds, which may also be important in the initial steps of the folding pathway and in guiding the protein's primary sequence into a P-barrel with the specific topology. In the work described here, we constructed and analyzed multiple sequence alignments for protein sequences in each of these five barrel folds. We also aligned structures of the different proteins, within and across the folds. In order to determine certain structural features common to the barrel folds at both the sequence and structural level, we studied the conservation and covariation in the SH3-like barrel, GroES-like and the PDZ domain-like folds. Molecular dynamics (MD) simulations on a GroES peptide, derived from a conserved β-turn, were also carried out in order to address its role as a possible nucleation site in the folding pathway. By combining sequence and structural analysis it was possible to interpret the pattern of conservation seen in the three protein folds.

Materials and methods

The SCOP database classifies all-β proteins into 93 folds according to their topology and evolutionary relationships (Murzin *et al,* 1995). Each fold is divided into superfamilies which are further classified into different families, that is, a group consisting of proteins with residue identities of 30% and greater or those having similar structure and function. The five folds include the SH3-like barrel, GroES-like, PDZ domain-like, N-terminal domains of the minor coat protein g3p and the Sm motif of the small nuclear ribonucleoproteins, SNRNP. A total of 42 different protein structures are grouped together in the SH3-like barrel fold. In the GroES-like and PDZ domain-like folds a total of nine and eight different structures have been solved, respectively. Only one structure each has been solved for both the N-terminal domains of the minor coat protein, g3p fold and the Sm motif of small nuclear ribonucleoproteins, SNRNP fold. A representative structure from each family within a fold was considered for the sequence and structure comparisons carried out in this study. Table I lists the initial target sequences considered for initiating the analysis.

Structure comparison

A total of 20 structures were considered for structural comparison of proteins (Table I). Of these, 13 belonged to the SH3-like barrel fold, three to the GroES-like barrel and two to the PDZ domain-like barrel fold. One structure was selected for each of the N-terminal domains of the minor coat protein g3p and the Sm motif of small nuclear ribonucleoproteins, SNRNP fold. Coordinates for each of these proteins were retrieved from the PDB (Bernstein *et al,* 1977). Superimpositions were done among the structures within the same and also across the various β-barrel folds. For inter-fold superimpositions, one representative from each fold was taken. The representative structure corresponded to any one protein in a fold for which complete sequence analysis was done as per the criteria of more than 30 sequences in the multiple sequence alignment (addressed later). Hence, the structures chosen for the inter-fold comparison were the α-spectrin SH3 domain protein, 1shg; the *Escherichia coli* GroES, 1aon and the rat neuronal nitric oxide synthase, 1qav from the SH3-like barrel, GroES-like and the PDZ domain-like fold, respectively. The structures were superposed by visualization, followed by least-squares fitting using the lsq commands of O (Jones *et al.,* 1991).

Residue conservation

To measure the level of conservation at each position in the alignment, the frequency of occurrence of an amino acid at each position was determined. This was achieved by the calculation of the positional entropy at each position in the alignments obtained. A positional entropy of *n* is equivalent to the diversity of n residues occurring at the position with a frequency of I/n. A position that is completely conserved will thus have a positional entropy of 1. For position *ι*. with residues *r* = (A, C, D, ..., V, W, Y) occurring at frequencies *Pi(r),* the entropy *H(i)* is defined as

$$H(i) = -p_i(r)\ln[p_i(r)]$$

This entropy is known as the Shannon informational entropy (Shenkin *et al,* 1991).

The positional entropy is expressed as

$$N(i) = e^{H(i)}$$

Volume correlation

The correlation coefficient at each residue position in the alignment was calculated as a measure of covariation in the volumes of the side chains. The side-chain volumes were taken from Harpaz *et al.* (Harpaz *et al.* 1994). A pairwise correlation coefficient, $r(x,y)$ determined the correlation between two residue positions and was expressed as

$$r(x,y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{[(n\sum x^2) - (\sum x)^2]\ [((n\sum y^2) - (\sum y)^2]}}$$

where *r* = correlation coefficient, *n* = number of sequences, *x* = volume at residue position *i* and *y* = volume at residue position *j*.

Sequence alignment

A total of 20 initial target sequences corresponding to the representative protein in each family were considered for the analysis (Table I). The chosen target sequence was used for a BLAST search (*E* < 0.001) of the non-redundant database compilation (Altschul *et al.,* 1997). Homologous sequences were retrieved and the stretch of residues, aligned to the initial target domain, extracted from each protein sequence. Two or more domains within a protein sequence were considered as separate sequences. Thus, a sequence having two domains was split into two, each corresponding to a different domain, within the protein. These sequences were aligned using the ClustalW program (Thomson *et al.,* 1994). Once the initial alignment was constructed, sequences with ClustalW score of >90 were removed in order to remove any bias in the sequence analysis due to high degree of similarity. To avoid artifactual results arising out of inaccurate sequence alignments, sequences with a score of <25 were also removed. The remaining sequences were realigned such that, in the final alignment, no two sequences had a score of <25 or >90. Families where, after the editing, the number of sequences in the alignment was <30 were not considered for further analysis.

Only five of a total of 20 families in the *n* = 4, 5 = 8 β-barrel protein folds fulfilled the criterion of >30 sequences in the multiple alignment. These included the SH3 domain and the C-terminal domain of ribosomal protein L2 in the SH3-like barrel fold, GroES and alcohol dehydrogenase-like, N-terminal

Table I. β-Barrel proteins considered for sequence and structure comparisons

β-Barrel fold	Classification within the fold	Proteins used for sequence alignment construction	PDB ID
SH3-like barrel	C-Terminal domain of biotin and diphtheria toxin repressors		
	Biotin repressor (BirA)	Biotin repressor/biotin holoenzyme synthetase, C-terminal domain	1bia
	Diphtheria toxin repressor (DtxR)	Diphtheria toxin repressor (DtxR)	2dtr
	SH3 domain	a-Spectrin SH3 domain	1shg
	Myosin S1 fragment, N-terminal domain		
	Myosin S1 fragment, N-terminal domain	Mvosin S1 fragment	2mys
	Translation proteins - SH3 like domain		
	Ribosomal proteins L24p and L21e	Ribosomal proteins L24 (L24p)	1ffk
	N-Terminal domain of eukaryotic initiation translation factor 5a	N-Terminal domain of eukaryotic initiation translation factor Sa	2eif
	C-Terminal domain of ribosomal protein L2	C-Terminal domain of ribosomal protein L2	1rl2
	Electron transport accessory proteins		
	Photosystem I accessory protein E(PsaE)	Photosystem I accessory protein E (PsaE)	1psf
	Nitrile hydratase B-chain	Nitrile hydratase B-chain	2ahj
	Ferredoxin thioredoxin reductase (FTR), alpha (variable) chain	Ferredoxin thioredoxin reductase (FTR). alpha (variable) chain	1dj7
	R67 dihydrofolate reductase	R67 dihydrofolate reductase	1vie
	CcdB		
	CcdB	CcdB	2vub
	DNA binding domain of retroviral integrase		
	DNA binding domain of retroviral integrase	DNA binding domain of retroviral integrase	1ex4
GroES-like	GroES-like		
	GroES	Chaperonin-10 (GroES)	1aon
	Alcohol dehydrogenase-like, N-terminal domain	Alcohol dehydrogenase	3bto
	SacY-like RNA-binding domain		
	BglG-like antiterminator proteins	SacY	1auu
PDZ domain-like	PDZ domain-like		
	PDZ-domain	Neuronal nitric oxide synthase, NNOS	1qav
	Interleukin 16	Interleukin 16	Iill6
N-Terminal domains of the minor coat protein g3p	N-Terminal domains of the minor coat protein, g3p		
	N-Terminal domains of the minor coat protein, g3p	N-Terminal domains of the minor coat protein, g3p	2g3p
Sm motif of small nuclear ribonucleoproleins, SNRNP	Sm motif of small nuclear ribonucleoproteins, SNRNP		
	Sm motif of small nuclear ribonucleoprotcins. SNRNP	D1 core SNRNP protein	1b34

domain in the GroES-like fold and PDZ-domain in the PDZ domain-like fold. Sequence alignment data corresponding to these families were considered for statistical analysis.

Molecular dynamics simulations

MD simulations for a small peptide of the *E.coli* GroES were performed using the Discover module in the Insightll molecular modelling package (MSI/Biosys, San Diego, CA, 1997). The simulations were performed with a cubic periodic boundary condition (box dimensions 25x25X25) and consisted of the peptide solvated with water molecules. The effective water density in the solvation box was 0.96 g/cm^3. All atoms were considered explicitly and their interactions were computed using the CVFF force field. The time step in the MD simulations was 1 fs. All simulations began with 100 iterations of the energy minimizations of the peptide to relax the local forces. Subsequently, MD simulations were performed at 300 K for 500 ps. A seven-residue peptide with the original conformation as in the protein with an intact type II turn was the starting structure. Simulations were performed for the wild-type sequence of the peptide and also on two other peptides. In one of these, the aspartate was mutated to asparagine and in the second the aspartate was mutated to alanine. The native-like side chain-main chain hydrogen bond was retained in the aspartate to asparagine mutant.

Results

The study involved comparison of sequence and structure data for the different four-stranded β-barrel folds. According to the

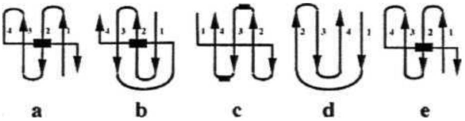

Fig. 1. Topologies of the β-barrel folds characterized by n = 4 and S = 8 according to the SCOP classification. The folds included in this category are (a) SH3-like barrel, (b) GroES-like. (c) PDZ domain-like, (d) N-terminal domains of the minor coat protein. g3p and (e) Sm motif of small nuclear ribonucleoproteins. SNRNP. The Sm motif of small nuclear ribonucleoproteins, SNRNP fold has the same topology as the SH3-like barrel. β-Strands are indicated by arrows. The boxes indicate the helices in the different folds.

number of strands forming a compact globular structure, these constitute the smallest barrels known. The difference among these different folds essentially lies in the manner in which the four β-strands are connected, thereby generating a unique topology (Figure 1). In each topological class multiple sequence alignments were generated for the sequence analysis and structures of proteins within a β-barrel fold superimposed on one another.

Comparisons within the *SH3-like* barrel fold

Superpositions were done among 13 structures in the SH3-like barrel fold (Table I). These structures superposed well on one another with a maximum r.m.s. deviation of 2.24 A for 29 atoms between the DNA binding domain of HIV-I integrase, 1ex4 and the diphtheria toxin repressor, 2dtr (Table IIa). The

Table II. Pair-wise r.m.s. deviations (A) when superimposing structurally equivalent Cα positions within a β-barrel fold. The numbers in parentheses represent the equivalent residues considered during the superpositions

(a) SH3-like barrel fold

	1vie	2dtr	1dj7	1psf	2vub	2mys	1shg	2ahj	1bia	1rl2	2eif	1ex4	1ffk
1vie	0.0	1.85(29)	1.49(46)	1.34 (40)	1.71 (47)	1.41 (37)	1.90(33)	1.98 (53)	1.90 (36)	1.74 (42)	1.70 (44)	1.71 (32)	1.53(42)
2dtr		0.0	1.99(36)	1.69(32)	1.87(34)	1.96(28)	1.81(35)	1.28(28)	1.73(37)	1.87(42)	1.66(21)	2.24(29)	1.47(28)
1dj7			0.0	1.45(53)	1.49(39)	1.83(46)	1.56(44)	0.91(46)	1.84(41)	1.72(42)	1.60(36)	1.56(37)	1.26 (45)
lpsf				0.0	1.51 (38)	1.87(36)	1.88(41)	1.31(48)	1.82 (37)	1.69(38)	1.41(33)	1.65(38)	2.03(54)
2vub					0.0	1.29(35)	1.23(38)	1.72(47)	1.74(37)	1.94(40)	1.97 (34)	1.39(33)	1.69(37)
2mys						0.0	1.94(40)	1.75(43)	1.87(32)	1.48(35)	1.41(34)	1.85(39)	1.72 (36)
lshg							0.0	1.61 (41)	1.89 (37)	2.21 (38)	1.94(30)	1.35(42)	1.81 (38)
2ahj								0.0	1.62(37)	1.46(41)	1.48(41)	1.54(38)	1.84(50)
lbia									0.0	1.89(39)	1.49(22)	1.99(35)	1.97(39)
1rl2										0.0	1.93(44)	1.96(42)	1.88(42)
2eif											0.0	1.50(33)	1.80(34)
1ex4												0.0	1.86 (37)
lffk													0.0

(b) GroES-like fold

	1aon	3bto	1auu
1aon	0.0	1.35(51)	1.69(37)
3bto		0.0	1.61 (34)
lauu			0.0

minimum r.m.s. deviation of 1.23 A for 38 atoms was seen between the α-spectrin SH3 domain protein, lshg and the CcdB protein, 2vub.

Of interest is the region at the type II β-turn of SH3-like barrel fold proteins. The turn, referred to as the diverging turn in the SH3 domain (Yi *et al.*, 1998), is present in the loop connecting strands 1 and 2 of the SH3-like barrel fold. Intriguingly, the occurrence of the type II turn in the SH3-Iike barrel fold proteins seems to be related to the length of the loop preceding the diverging turn. Of the 13 different SH3-like barrel fold structures studied (Table I), the type II turn was present in six. None of the remaining seven structures contained the type II turn. A common feature among these seven structures, lacking the type II turn, was the presence of a short loop between strands one and two of the barrel. The presence of a short intervening loop probably reduces the likelihood of the polypeptide chain from deviating from the folded barrel structure. Proteins containing the type II turn included the a-spectrin SH3 domain, lshg; photosystem I accessory protein, lpsf; diphtheria toxin repressor, 2dtr, nitrile hydratase β-chain, 2ahj; the ribosomal protein L24, 1ffk and ferredoxin thioredoxin reductase, 1dj7. A stretch of >11 residues in the loop connecting strands 1 and 2 of the β-barrel necessitated the presence of the type II turn, as observed in five of these structures. The presence of the turn, in these structures, appears to guide the polypeptide into the β-barrel helping in the formation of the folded barrel structure.

Multiple sequence alignments for each of the 13 proteins considered for structural comparisons were generated as described in Materials and methods. A BLAST search with the amino acid sequence of the a-spectrin SH3 domain (SH3 domain family) gave 219 hits with E < 0.001. Splitting of multi-domain sequences augmented this number to 302. Exclusion of sequences with ClustalW scores of <25 and >90 drastically reduced the number of sequences in the final alignment to 30. A BLAST search for the sequence of the C-terminal domain of ribosomal protein L2 (translation proteins - SH3-like domain family) resulted in an initial number of 132 hits with E < 0.001. A total of 65 sequences homologous to the ribosomal protein w -e finally obtained by editing the sequences in a manner similar to that described above. For all the remaining sequences subjected to BLAST search, the number of sequences after editing was <30. These sequence alignments were hence not considered for further analysis for reasons described in the Materials and methods section.

The degree of conservation at each position in the multiple alignment generated was determined using the Shannon Informational entropy calculation (Shenkin *et al.,* 1991). Tables IIIa and b give the positional entropy values at the core residue positions for the SH3 domain and the C-terminal domain of ribosomal protein L2 families, respectively. Core residue positions for the representative proteins in each family were identified by calculating the percentage accessibility of side chains using the NACCESS program (Hubbard *et al.,* 1991). Residues with side chain accessibilities of <7% were considered part of the core. A total of nine core positions were identified in the a-spectrin SH3 domain. The positional entropies were <3 at all nine core positions in the SH3 domain protein (Table IIIa). In the case of the C-terminal domain of ribosomal protein L2, 10 of the 12 core positions showed high residue conservation as indicated by a positional entropy of <3 at these positions (Table IIIb). Tables IIIa and b also indicate the prevalence of amino acid residues at the core residue positions for the SH3-like barrel fold families, SH3 domain and the C-terminal domain of ribosomal protein L2. These core positions, as seen from the data, are predominantly occupied by valine, leucine or isoleucine in both the families. The other residues occupying positions in the barrel core are the non-polar residues including phenylalanine. methionine, alanine and glycine. The presence of these residues contributes to the high hydrophobicity at the core of the barrel in this fold. High conservation of non-polar residues at the core residue positions suggests the importance of a hydrophobic interior in maintaining the integrity of the fold.

A covariance analysis of residue volumes indicated a significantly high correlation between residues at the core positions

Table III. **Position-specific** statistics for occurrence and conservation of residues at the core positions in the five β-barrel families. The first column indicates residue position numbers corresponding to sequence of (a) the α-spectrin SH3 domain protein, 1shg; (b) C-terminal domain of ribosomal protein L2. 1rl2; (c) the *E.coli* GroES protein, 1aon; (d) alcohol dehydrogenase-like, N-terminal domain protein, 3bto; and (e) the PDZ-domain protein, 1qav. Occurrence of the most prevalent residue is indicated in the last column

	A	C	Y	W	S	M	D	N	E	Q	V	L	I	H	G	P	K	R	F	T	Positional entropy	Most prevalent residue
(a) SH3 **domain:** a **total** of 30 sequences																						
9	7		4								18								1		2.79	V
11	28														2						1.27	A
23											6	22	2								2.07	L
25						8						1	1						20		2.33	F
31						1					3	8	18								2.72	]
33											18	8	4								2.52	V
44	2										6				22						**2.07**	G
53											9		1						20		2.10	F
58	1										27	1	1								1.54	V
(b) C-Terminal domain of ribosomal protein L2: a total of 65 sequences																						
131						3					2	54	6								1.86	L
134	1					7					3	1	53								1.97	I
140											38	2	25								2.20	V
143											26	1	38								2.10	I
151	13				1										51						1.78	G
160	1	1	20	1	21			3	1		9						1	7			**5.50**	S
161	49	2			2						8			1		3					2.44	A
163											16	22	27								2.93	1
171	17	4			2						31		3		1					7	4.17	V
173											23	31	11								2.78	L
187	2	58			3			1													1.61	C
189	65																				1.00	A
(c) GroES: a total of 86 sequences																						
10											71	2	13								1.70	V
12											54	5	27								2.27	V
38	15														71						1.57	G
40											70		16								1.60	V
59					1						78	5	2								1.47	V
65											72	!	13								1.61	V
67			26			1					1	5	3						50		2.88	F
73	5				5			5	1	3	7	1			5					54	3.99	T
84			1		1	1					13	53	9					1		4	3 3.63	L
86						31					4	42	4						5		3.27	L
91											15	6	65								1.99	I
(d) Alcohol dehydrogenase-like, N-terminal domain: a total of 52 sequences																						
36											49		2							1	1.29	V
38						1					16	6	28					1			2.99	I
69	30				1								1		20						2.31	A
73											45		7								1.48	V
83						I					20	27							4		2.67	L
89	2										50										1.18	V
91	1				1	1						6				41			1	1	2.62	P
150	5				1	2							1							43	1.93	T
152	3	1			1	1					33	8	3							2	3.52	V
157	4	3							1		33	11									2.87	V
(e) PDZ domain: a total of 35 sequences																						
84											4	22	9								2.43	L
106											4		31								**1.42**	1
108											9		26								1.76	I
116	18										4	1	2			9		1			3.69	A
117	35																				1.00	A
123											2	33									**1.24**	L
128	11	3			1					10	2						1	7			**5.05**	A
129											3	9	23								**2.30**	I
132											33		2								1.24	V
137											12	21	2								2.30	L
140	17	2			1				3	1	10									1	4.00	A
145	31										4										1.42	A
152	20				4					2					1	1				7	3.57	A
156											30	2	3								1.65	V
158												29	5						1		170	L
160	7										21	4	3								296	V

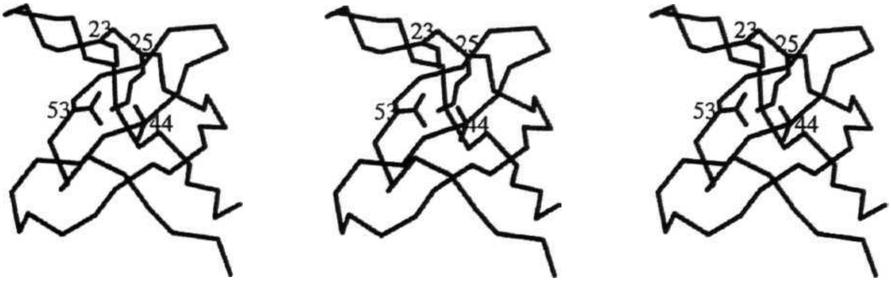

Fig. 2. Stereo-view of the SH3 domain protein, 1shg showing side chains of the correlated i sidues, 23, 25, 44 and 53. Figures 2 and 3 were generated using Motscript (Kraulis, 1991).

in the SH3-like barrel (Figure 2). A high negative correlation of –0.97 was seen between residue positions 23 and 44 (residue numbers corresponding to the α-spectrin SH3 domain). An increase in volume of the core due to a larger side chain at residue position 23 (mostly leucine or valine) is compensated by a reduction in the side chain volume at the correlated position 44 (mostly valine or glycine) (Table IIIa). Similarly, a negative correlation is seen between the volumes of residues 25 and 44 (a correlation value of –0.64). Since the side chain volumes of positions 23 and 25 are negatively correlated with respect to position 44, it was anticipated that the correlation coefficient between them would be positive. Indeed, the correlation coefficient of side chain volumes at positions 23 and 25 shows a high positive value of 0.68. Another compensatory pair of residues is present at positions 44 and 53. Residue 44 exists in the third strand of the α-spectrin SH3 domain and residue 53 in the fourth strand. A negative correlation of –0.84 is seen between these core residue positions. Positive correlations of 0.75 between residues 23 and 53 and 0.42 between residues 25 and 53 result in compensation of the overall core volume. This observation strongly supports the belief that maintenance of the total volume of the core would be important to keep the barrel structure intact.

Comparisons within the GroES-like fold

Three representative proteins in the GroES-like fold were considered for the comparative study (Table I). Superposition of proteins within the GroES-like fold was done as in the case of the SH3-like barrel fold. The representative proteins included for analysis superposed well with one another with a minimum r.m.s. deviation of 1.35 A for 51 atoms between the horse alcohol dehydrogenase, 3bto and the *E.coli* GroES, 1aon (Table IIb). Different proteins within the alcohol dehydrogenase-like, N-terminal domain family also superimposed very well on one another (data not shown). Comparisons revealed an overall conservation of the β-barrel core in the representative proteins.

A BLAST search yielded 170 hits for the *E.coli* GroES protein. Splitting of the multi-domain sequences increased this number to 174. Further editing as described earlier for the SH3-like barrel fold, however, reduced the number to 86. An initial number of 412 hits in a BLAST search for the alcohol dehydrogenase reduced to 52 sequences in the final alignment after appropriate editing. In case of the SacY protein, the number of sequences in the final alignment was <30. This protein and the corresponding family were thus omitted from further sequence and structural comparisons.

Core residue positions were identified in the two GroES-like fold proteins, the *E.coli* GroES and the horse alcohol dehydrogenase. Positional entropy values for the corresponding families at these core positions are shown in Tables IIIc and d. Of the 11 core residue positions in the *E.coli* GroES, eight were highly conserved with positional entropies of <3. Two of the three high-entropy positions, 84 and 86, were largely occupied by non-polar residues, the most predominant being leucine. Another variable position in the core, 73, was mostly occupied by threonine. Ten core positions were identified in the alcohol dehydrogenase. High residue conservation is seen at nine of the 10 core positions, indicated by a positional entropy value of <3 (Table IIId). These positions were largely occupied by small hydrophobic amino acid residues. The predominant residue at position 152, the only high-entropy position in the alcohol dehydrogenase, was valine. Tables IIIc and d indicate the prevalence of non-polar amino acid residues, including valine, leucine and isoleucine, at the core residue positions for the GroES-like fold proteins. The presence of polar, uncharged residues at the core positions, however, is not uncommon. As reported earlier, valines at the core positions in these proteins are seen to be mutable into isoleucines but not to leucines (Table III) (Taneja and Mande, 1999). Unlike the SH3-like barrel fold proteins, proteins in the GroES-like fold did not show a high correlation between residue volumes in the barrel core.

Comparisons within the PDZ domain-like fold

The two representative structures of the PDZ domain-like fold interleukin 16, 1il16 and the neuronal nitric oxide synthase, 1qav superposed well on one another with an r.m.s. deviation of 1.90 A for 77 atoms. Proteins within the PDZ domain family when compared among themselves superposed well on one another with an overall conservation of the β-barrel (data not shown).

A BLAST search with the amino acid sequence of the neuronal nitric oxide synthase (representative of the PDZ-domain family) gave 230 hits. This initial number first rose to 386 owing to splitting of multi-domain sequences, but a final number of 35 sequences was obtained after editing. The number of sequences in the final alignment obtained from the protein interleukin 16 was <30. This protein and the corresponding family were thus not considered for further analysis.

A total of 16 core positions were identified in the neuronal nitric oxide synthase. Of these, 12 positions show high residue

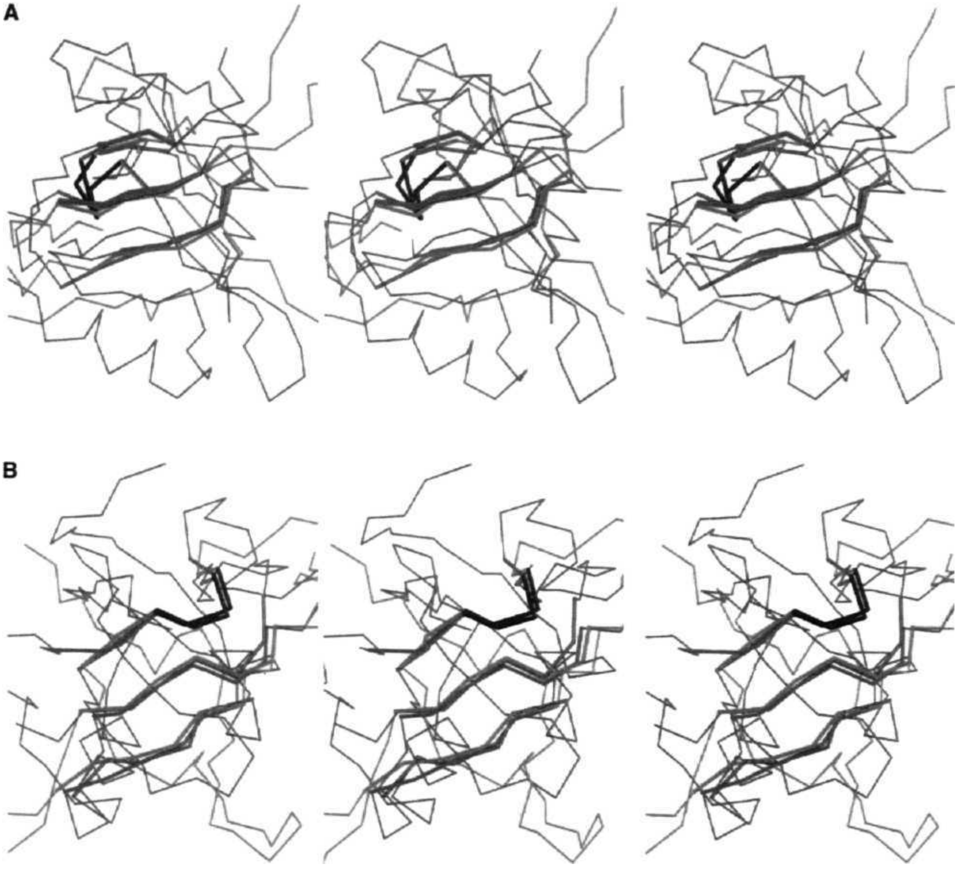

Fig. 3. (A) Superposed structures of the representative proteins of the SH3-like barrel. GroES-like and PDZ domain-like folds. The α-spectrin SH3 domain protein (1shg) is shown in pink, the *E.coli* GroES (laon) in blue and the neuronal nitric oxide synthase (lqav) in green. The superposed 3_{10} helix is highlighted in dark blue. (B) Superposed structures of the representative proteins of the SH3-like barrel. GroES-like and PDZ domain-like folds. The α-spectrin SH3 domain protein (lshg) is shown in pink, the *E.coli* GroES (laon) in blue and the neuronal nitric oxide synthase (lqav) in green. The superposed type II β-turn is highlighted in dark blue.

conservation with a positional entropy <3 at each of these positions (Table IIIe). These core positions are predominantly occupied by valine, leucine or isoleucine. Alanine seems to be the residue of choice for the remaining four positions. As in the GroES-like fold, the valines appear to be mutable to leucines rather than to isoleucines (Table IIIe). Core residue positions did not show a significant correlation among residues in proteins considered in this fold.

The final number of sequences in the multiple sequence alignments generated for the representative proteins in the N-terminal domains of the minor coat protein, g3p and Sm motif of small ribonucleoproteins, SNRNP families was <30. Since there were no sequence data for the two families owing to lack of fulfillment of the set criteria for sequence analysis, the two families and hence the corresponding folds were excluded from the study.

Comparisons across the β-barrel folds

One of the objectives of the study was to identify similarities and dissimilarities across the β-barrel folds. One representative structure from the three β-barrel folds, viz. a-spectrin SH3 domain (SH3-like barrel fold), *E.coli* GroES (GroES-like fold) and the neuronal nitric oxide synthase (PDZ domain-like fold) were therefore considered for the comparisons. The topologies of the three representative structures are different from one another as shown in Figure 1. Comparison of topologies of the SH3-like barrel and GroES-like fold shows the presence of a 3_{10} helix interrupting the fourth strand in both the β-

Table IV. Pair-wise r.m.s. deviations (A) when superimposing structurally equivalent Cot positions across the β-barrel folds. The numbers in parentheses represent the equivalent residues considered during the superpositions

	lshg (SH3 domain)	laon (*E.coli* GroES)	lqav (PDZ domain)
(a) Superimposition at the 3_{10} helix			
lshg (SH3 domain)	0.0	1.79(38)	1.48 (23)
laon (*E.coli* GroES)		0.0	1.24 (23)
lqav (PDZ domain)			0.0
(b) Superimposition at the type II β-turn			
lshg (SH3 domain)	0.0	1.62(34)	1.48(23)
laon (*E.coli* GroES)		0.0	1.50(31)
lqav (PDZ domain)			0.0

barrel folds. The first three strands of the barrel form a similar anti-parallel p-sheet in the two protein folds, yet the two have distinct topologies. The difference lies in the way in which the fourth β-strand hydrogen bonds with the other strands forming the barrel. In the case of the SH3-like barrel fold, the fourth strand runs anti-parallel to the third β-strand followed by the 3_{10} helix. This short helix juxtaposes the fourth strand to the first resulting in the formation of the barrel. The 3_{10} helix in GroES-like fold, however, juxtaposes the fourth strand to the third, for the formation of the complete barrel. Figure 1 also indicates the topology of the PDZ domain-like fold. This fold consists of two helices, one in the region connecting the p-strands two and three and the other between the third and the fourth strands.

The difference in topology of the three representative proteins thus makes it difficult to superpose the corresponding structures. The presence of a common β-barrel structural core, however, may allow comparison of the secondary structural elements forming the barrel in these proteins. Hence, ignoring the topology of the three folds, P-strands of the representative proteins were superimposed on one another. Structural comparisons yielded two alternative ways in which P-strands of the three proteins could be superimposed on one another with minimal r.m.s. deviation values. In one of the superimpositions, the 3_{10} helix of the α-spectrin domain superposes very well on that in the *E.coli* GroES (Figure 3a). The superposition is such that residues of strands 1, 2 and 3 of the α-spectrin domain align with those in strands 3, 2 and 1 of the *E.coli* GroES respectively. The r.m.s. deviation data are as shown in Table IVa. In the case of the PDZ domain-like fold, the structural alignment superposes strands 2, 3 and 4 of the neuronal nitric oxide synthase onto strands 1, 2 and 3 of the α-spectrin SH3 domain, respectively. In an alternative superposition, P-strands of the representative structures align such that strands 4, 1 and 2 of the a-spectrin SH3 domain align with strands 1, 2 and 3 of the *E.coli* GroES, respectively. The r.m.s. deviation data for this superimposition are given in Table IVb. Interestingly, this alternative superposition superimposes a type II (3-tum present in the three structures (Figure 3b).

The β-turn, referred to as the diverging turn in the SH3 domain, occurs at the intervening loop connecting strands 1 and 2. The turn has previously been reported to play a role in protein folding (Riddle *et al.*, 1999; Larson and Davidson, 2000). A similar P-tum is present just before the third strand in the PDZ domain-like fold. The ϕ–ψ values at this turn correctly place this turn in the type II category (ϕ_{i+1} = -63, $\psi_{i+,}$ = 140; ϕ_{i+2}= 98.6, ψ_{i+2}= –8.6).In the GroES-like fold the turn ($\phi_{i+,}$ = -58.8, $\psi_{i+,}$ = 134.6; ϕ_{i+2} = 96.3, ψ_{i+2} = -104) is present at the initiation of the third P-strand following the dome loop. The presence of the type II turn in the GroES-like and PDZ domain-like folds suggests that the turn in these protein folds may play a role similar to that observed in case of the SH3 domain. We considered this tum to be a crucial folding nucleus in the β-barrel folds treated in this study. Further analyses were therefore carried out in relation to the superimposition where the P-tum of all the three representative structures superposed on one another as indicated in Figure 3b.

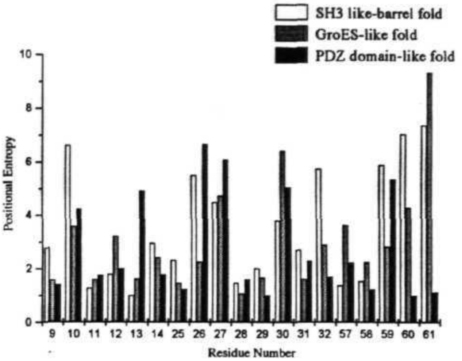

Fig. 4. Positional entropies at the structurally aligned positions of the three β-barrel folds: SH3-like barrel fold, PDZ domain-like fold and GroES-like fold. The position numbers correspond to that of the α-spectrin SH3 domain protein. Position numbers 9, 11, 25, 31 and 58 are the core positions in all the three protein barrels. Residues 28 and 29 correspond to the *i* + *2* and *i* + 3 positions in the type II turn in the three-dimensional structure.

Residue conservation

In order to assess the variability of ammo acid residues across the three P-barrel folds, positional entropies were compared at the structurally aligned positions of the three folds. Upon alignment of the representative structures in the SH3-like barrel, GroES-like and PDZ domain-like folds, three P-strands of each structure superimposed well on one another (Figure 3b). Sequences of the representative proteins were then aligned on the basis of the structural alignment. Residues spanning the first β-strand of the a-spectrin SH3 domain (9–14) aligned with positions 38–43 (strand 2) in the *E.coli* GroES and with residues 106–111 (strand 2) in the neuronal nitric oxide synthase. Residues 25-32, spanning the diverging type II turn (26–29) of the a-spectrin SH3 domain, aligned with positions 59–66 and 123–130 of the *E.coli* GroES and neuronal nitric oxide synthase, respectively. Residues 28 and 29 (numbers correspond to the a-spectrin SH3 domain) form the *ι* + 2 and *i* + 3 positions of the type II turn. Residues spanning strand 4 of the a-spectrin SH3 domain (57-61) structurally aligned with those in the first β-strand of *E.coli* GroES (residues 11–15) as also in the neuronal nitric oxide synthase (residues 95-98). Figure 4 shows the positional entropies at the structurally aligned residues. Among these 19 structurally aligned positions, high conservation in each of the three proteins is seen at eight positions. The positional entropies at these positions are <3 in all the three protein sequence alignments. Remarkably, five of these eight highly conserved positions form the core of the

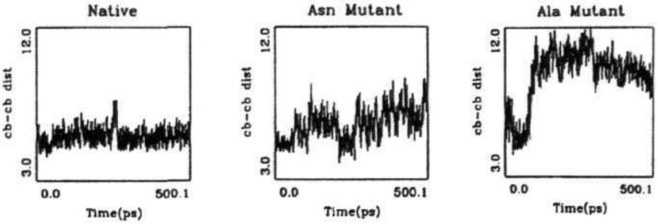

Fig. 5. Comparison of the Cβ–Cβ distance among the native and mutant peptides of *E.coli* GroES during molecular dynamics simulations. It can be seen clearly that the native peptide with aspartate at t + 3 position shows a stable Cβ–Cβ distance. The distance increases rapidly in the aspartate to alanine mutant peptide. See text for details.

Table V. Comparison of positional entropies of the β-barrel folds at the structurally aligned positions. Residue position numbers in the first column correspond to the α-spectrin SH3 domain protein, 1shg

Structurally aligned positions	Residue position in the structure	Positional entropy in SH3-like barrel fold	Positional entropy in GroES-like fold	Positional entropy in PDZ domain-like fold	Positional entropy <3 for all three folds
9	Core	2.8	1.6	1.4	Yes
10		6.6	3.6	4.2	
11	Core	1.3	1.6	1.8	Yes
12		1.8	3.2	2.0	
13		1.0	1.6	4.9	
14		**2.9**	2.4	1.8	**Yes**
25	Core	2.3	1.5	1.2	**Yes**
26		5.5	2.3	6.7	
27		4.5	4.7	6.1	
28	**β-Turn**	1.5	1.1	1.6	**Yes**
29	**β-Turn**	2.0	1.7	1.0	**Yes**
30		3.8	6.4	5.0	
31	Core	2.7	1.6	2.3	Yes
32		5.8	2.9	1.7	
57		1.4	3.6	2.2	
58	Core	1.5	2.3	1.2	Yes
59		6.0	2.9	5.4	
60		7.0	4.3	1.0	
61		7.4	9.4	1.1	

barrel in all the three protein structures (Table V). The core positions are largely occupied by small non-polar residues. High conservation at core positions in the protein barrel suggests the importance of core residues in the formation and maintenance of the barrel structure.

Of the remaining three highly conserved residue positions, two correspond to residues in the type II β-turn mentioned earlier. The *i* + 2 and *i* + 3 residue positions of the type II turn (corresponding to residues 28 and 29 in the α-spectrin SH3 domain) show a high residue conservation. The predominant residue at the *i* + 2 position is glycine and that at *i* + 3 is aspartate. A high residue conservation at the type II β-turn has earlier been reported for the GroES-like fold at the corresponding positions 62 and 63 of *E.coli* GroES. H-bonding between the side-chain carboxylate of aspartate and main chain amide of the first residue of the tum has been suggested to be important in juxtaposing the β-strands of the barrel, such that the barrel structure is maintained (Taneja and Mande, 1999). A similar side chain-main chain interaction in the SH3 domain has been suggested to stabilize the type II β-turn (Larson and Davidson, 2000). The high degree of conservation seen at the *i* + 2 and *i* + 3 residue positions of the type II turn in the three barrel folds suggests the importance of this turn in the formation or maintenance of the β-barrel structure.

Molecular dynamics simulations

Since the side chain–main chain interaction in the type II β-turn appears to be important for the barrel structure formation and maintenance, disruption of this interaction should result in the disintegration of the type II tum. This would ultimately result in the loss of the barrel structure. To investigate the stability of the type II P-tum upon alteration of the aspartate, we performed extensive MD simulations for the GroES peptide and its mutants. The sequence of the peptide taken for MD simulations was VKVGDIV (corresponding to residues 59—65 in the *E.coli* GroES). The starting conformation for the MD simulations was as observed in the crystal structure of the protein. Additional MD simulations were done with the aspartate mutated to asparagine in one case and to alanine in the other. Any alteration in the conformation in the type II turn would immediately be evident from changes in values of the dihedral angles in the type II turn.

A comparison of the ϕ and ψ values was performed for the different residue positions in the β-turn during the 500 ps simulation. The ϕ and ψ values of the *i* + 1 residue remain more or less similar in all the three peptides (data not shown). However, major deviations occur in the *fy-y* values of the *i* + 2 residue, glycine, when the *i* + 3 residue is mutated from aspartate to asparagine or alanine. While in the native peptide, the ϕ and ψ values at ι + 2 position fluctuate around the value of +100 and −40, respectively, ϕ_{i+2} changes to about +150 in the mutant peptides. The ψ_{i+2} value also drops from -2 to about -100 for both the mutant forms. This, however, occurs after an initial sudden rise of ψ_{i+2} from -2 to +70. A large deviation is seen at the fourth residue position in the P-turn in the mutant peptide. In the case of the aspartate to asparagine mutation, ϕ_{i+3} drops from about -60 to −141 while it remains stabilized in the native peptide. Comparison of the distance between Cβ atoms of the ι and 1 + 3 residue further corroborates the disintegration of the type II turn upon mutation of the fourth residue in the turn (Figure 5). While this distance is maintained at around 5.6 A in the native peptide, it increases to about 8 A in the aspartate to asparagine mutant and to about 10 A in the aspartate to alanine mutant peptide. These results indicate the importance of the side chain-main chain H-bond interaction in the type II turn maintenance. Alteration of aspartate to asparagine is hence sufficient to cause the disruption of the type II turn conformation. A high conservation of the turn, as also the residues in the turn, thus might be of evolutionary importance in maintaining the structure of the P-barrel.

Discussion

Within various protein families such as serine proteases, cysteine proteases and globins, the three-dimensional structure is remarkably similar despite considerable variations in the amino acid sequences. To a certain extent, conserved residues or conservative changes account for the structural conservation. In addition, correlated pairs of residues have an important role in stabilizing the protein structure. Determination of these conserved features along with the compensatory substitution patterns helps in increasing our understanding of features that may determine the three-dimensional structure of a protein.

In this study, we have attempted to identify folding determinants in the small p-barrel proteins. Conservation patterns across these β-barrel folds reveal interesting similarities of residues at the core of the protein barrels. Irrespective of the topologies, these proteins show a high conservation of small non-polar amino acid residues at the core positions. The core residue positions are predominantly occupied by valine, leucine and isoleucine. Interestingly, valines at the core positions are seen to be mutable into isoleucine and not leucine, an observation reported earlier (Taneja and Mande, 1999). The higher frequency of substitution of isoleucine by valine has been attributed to a higher P-sheet propensity of isoleucine and valine than leucine (Wilmot and Thornton, 1988). Branching of side chains at the Cβ positions in both valine and isoleucine, but not leucine has previously been suggested as a possible reason for such a mutation pattern (Taneja and Mande, 1999). The observed mutation pattern and a high conservation of non-polar side chains suggest that the overall hydrophobic pattern of amino acids may drive the protein sequence to collapse into the P-barrel conformation.

Correlation analysis of the SH3-like barrel fold suggests that maintenance of the total core volume occurs within the SH3 domain family of proteins. Amino acid substitutions resulting in an increase or a decrease in the volume of the core is compensated by replacement of another amino acid residue. This amino acid residue is present at a position that might be distant in sequence, but near in space to the mutated residue so as to conserve the total volume of the core and hence the overall folded structure. Interestingly, an earlier analysis of 266 SH3 sequences did not find evidence for correlated substitutions (Larson and Davidson, 2000). We suggest that our criteria of choosing sequence identities between 25 and 90 generates a more accurate multiple alignment for a robust statistical analysis. Accuracy of the alignment is reflected in observation of the covarying mutations.

Of the common features, the presence of a type II P-turn is the most intriguing. This turn seems to be important in the formation of the β-barrel. Earlier studies have reported the importance of the β-turn in SH3 domain (Riddle *et al.*, 1999; Larson and Davidson, 2000). Conservation of this turn, not only in proteins constituting one of the β-barrel folds but also across the various β-barrel folds considered in this study, suggests that this region might be an important nucleation site in the folding pathway of the β-barrel proteins (Riddle *et al.*, 1999; Larson and Davidson, 2000). This nucleation appears to be guided by the residues present within the turn. High residue conservation has been seen at the ι + 2 and ι + 3 residue positions. While the presence of glycine at $i + 2$ guides the protein sequence into a tum, aspartate at ι + 3 is important for a unique side chain–main chain interaction. Simulation studies corroborate similar conclusions of independent work carried on an SH3 peptide (Krueger and Kollman, 2001). Furthermore, our analysis suggests that alteration of aspartate to asparagine or alanine destabilizes the P-turn conformation. The glycyl-aspartyl dipeptide hence appears to be a major factor in helping maintain the integrity of the barrel.

Our study shows interesting similarities among proteins in the different P-barrel folds. Despite large differences in sequence and function, the occurrence of a conserved glycyl-aspartyl dipeptide, intriguingly at a conserved type II tum, suggests the importance of the turn and the residues forming it in the formation of the P-barrel. In addition, a conserved hydrophobic core suggests its role in maintenance of the barrel structure. Further studies such as site-directed mutagenesis should confirm the importance of these conserved features in the formation and maintenance of protein structure.

Acknowledgements

We thank Debasis Mohanty and Sharmila Mande for useful comments and suggestions. Coordinates for *Thermoplasma acidophilum* glucose dehydrogenase were kindly supplied by Garry Taylor. B.T. and R.Q. are CSIR Senior and Junior Research Fellows, respectively. Financial support for the work was provided by the Department of Biotechnology and by the Council of Scientific and Industrial Research.

References

Altschuh,D.. Vernet,T., Berti,P., Moras,D. and Nagai,K. (1988) *Protein Eng.*. 2, 193-199.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W and Lipman,D.J. (1997) *Nucleic Acids Res.*, 25, 3389-3402.

Anfinsen,C.B. (1973) *Science*, **181**, 223-230.

Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.E., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T and Tasumi,M. (1977) *J.Moi BioL*, **112**, 535-542.

Bowie,J.U., Reidhaar-OlsonJ.F. Lim,W.A. and Sauer,R.T. (1990) *Science*. 247. 1306–1310.

Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) *J. Mot. Biol.*, **179**, 125-142.

Harpaz,Y., Gerstein,M. and Chothia,C. (1994) *Structure*, 2. 641-649.

Hubbard,S.J., Campbell,S.F. and Thornton,J.M. (1991) / *Mol. Biol.*, **220**, 507-530.

Jones,T.A., Zou,J.Y., Cowan,S.W and Kjeldgaard,M. (1991) *Acta Crystallogr.*, **A47**, 110-119.

Kamtekar,S., Schiffer,J.M., Xiong,H., Babik,J.M. and Hecht,M.H. (1993) *Science*, 262, 1680–1685.

Kraulis,P.J. (1991) *JAppl. Crystallogr.*, **24**, 946–950.

Krueger,B.P. and Kollman,P.A. (2001) *Proteins: Struct. Funct. Genet.*, 45, 4–15.

Larson,S.M. and Davidson,A.R. (2000) *Protein Sci.*, 9, 2170-2180.

Mandel-Gutfreund,Y., Zaremba,S.M. and Gregoret,L.M (2001) *J. Mol. Biol.*, 305, 1145–1159.

Martinez,J C and Serrano.L. (1999) *Nature Struct. Biol., 6,* 1010-1016.

Merkel,J.S., Sturtevant,J.M. and ReganX. (1999) *Structure*, 7, 1333-1343.

Murzin,A.G., Lesk,A.M. and Chothia,C. (1994) J. *Mol. Biol.*, 236, 1382-1400.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C (1995) *J. Mol. Biol.*, 247, 536-540.

Pollock.D.D. and Taylor,W.R. (1997) *Protein Eng.*, 10, 647–657.

Riddle.D.S., Grantcharova,V.P., Santiago,J. V., Alm,E., Ruczinski,I. and Baker,D (1999) *Nature Struct. Biol.*, 6, 1016-1024.

Serrano,L. (2000) *Adv. Protein Chem.*, 53, 49–85

Shenkin,P.S., Erman,B. and Mastrendrea,L.D. (1991) *Proteins,* 11, 297-313.

Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) *Protein Eng.*, 7, 349–358.

Taneja,B. and Mande,S.C. (1999) *Protein Eng.*, 12, 815-818.

Thomson,J.D., Higgins,D.G. and Gibson,TJ. (1994) *Nucleic Acids Res.*, 22, 4673-80.

Wilmot,C.M and Thornton,J.M (1988) *J. Mol Biol.*, 203, 221-232.

Yi,Q., Bystroff,C., Rajagopal,P., Klevit,R.E. and Baker,D (1998) *J Mol. Biol.*. 2*3. 293–300.

Received April 26. 2002: revised October 3, 2002: accepted October 10. 2002