Protein Engineering vol.12 no.10 pp.815–818, 1999

SHORT COMMUNICATION

Conserved structural features and sequence patterns in the GroES fold family

Bhupesh Taneja and Shekhar C.Mande[1]

Institute of Microbial Technology, Sector 39-A, Chandigarh 160 036, India

[1]To whom correspondence should be addressed. Email: shekhar@bragg.imtech.ernet.in

An irregular, all β-class of proteins, comprising members of the chaperonin-10, quinone oxidoreductase, glucose dehydrogenase and alcohol dehydrogenase families has earlier been classified as the GroES fold. In this communication, we present an extensive analysis of sequences and three dimensional structures of proteins belonging to this family. The individual protein structures can be superposed within 1.6 Å for more than 60 structurally equivalent residues. The comparisons show a highly conserved hydrophobic core and conservation of a few key residues. A glycyl-aspartate dipeptide is suggested as being critical for the maintenance of the GroES fold. One of the surprising findings of the study is the non-conservative nature of Ile to Leu mutations in the protein core, although Ile to Val mutations are found to occur frequently.

Keywords: alcohol dehydrogenase/β-barrel/chaperonin-10/GroES-fold

Introduction

The GroES fold has been described as an irregular β-barrel and has been found to occur in at least four different functional classes of proteins: the quinone oxidoreductases (QOR), the alcohol dehydrogenases (ADH), the glucose dehydrogenases (GDH) and the chaperonin-10 (cpn10) (Murzin, 1996). The fold is characterized essentially by four β-strands ($n = 4$) and a shear number of 8 ($S = 8$). It falls in one of the theoretically deduced β-barrel classes (Murzin *et al.*, 1994a). The shear number (S) and the number of strands (n) match closely with the SH3 fold family. However, the two are distinct from each other topologically (Murzin *et al.*, 1994b). In addition to the four β-strands, there is an insertion of a short 3_{10} helix before the third strand of the β-barrel. The insertion of this 3_{10} helix helps in marginally widening the β-barrel. In order to characterize the sequence determinants of the GroES fold and study the conserved patterns of the three-dimensional structures, we have carried out extensive comparisons of all the known amino acid sequences and the three-dimensional structures belonging to the GroES fold proteins. The analysis shows interesting features, such as a highly conserved hydrophobic core and conservation of a few key residues across the various protein families. Conservation of these residues is shown to be important for the maintenance and integrity of the three-dimensional fold.

Materials and methods

The analysis is based on comparison of 157 protein sequences and eight three-dimensional structures. All the sequences were retrieved from the Swiss-Prot Release 36.0 (Bairoch and Apweiler, 1999). The 157 sequences retrieved comprise 86 of the various alcohol dehydrogenases, 58 of the chaperonin-10, 12 of the quinone oxidoreductases and the *Thermoplasma acidophilum* glucose dehydrogenase. The sequences were aligned using the ClustalW program (Thomson *et al.*, 1994) and the resulting alignments were improved by manual fitting.

The three-dimensional structures for the proteins belonging to the GroES fold family were retrieved from the Protein Data Bank (PDB) (Sussman *et al.*, 1998). Of the eight structures compared, three belonged to the chaperonin-10 family, namely, GroES of *Escherichia coli* (1AON; Xu *et al.*, 1997), chaperonin-10 of *Mycobacterium leprae* (1LEP; Mande *et al.*, 1996) and the gp31 protein of the T4 phage (1G31; Hunt *et al.*, 1997), and three to the alcohol dehydrogenase family. The proteins belonging to the alcohol dehydrogenase family were horse alcohol dehydrogenase (2OHX; Eklund *et al.*, 1976), human beta1 alcohol dehydrogenase (1DEH; Hurley *et al.*, 1994) and human sigma alcohol dehydrogenase (1AGN; Xie *et al.*, 1997). The other two structures considered in these comparisons were of *E.coli* quinone oxidoreductase (1QOR; Thorn *et al.*, 1995) and *T.acidophilum* glucose dehydrogenase (John *et al.*, 1994). The structures were superposed by visualization followed by least-squares fitting using the lsq commands of O (Jones *et al.*, 1991).

The secondary structure assignments and the accessible surface areas for all the protein structures were calculated using the dssp program (Kabsch and Sander, 1983). The amino acid residues with an accessible surface area of less than 5% for at least one representative of each class were classified as residues in the core. These residues were further confirmed as core residues by manually inspecting their positions in the respective three-dimensional structures.

Results and discussion

The overall topology of all the structures that were compared is very similar, as shown by Murzin (1996). The three-dimensional structures of the four families can be superimposed very well with one another (Table I). Approximately 60 residues of each protein superimpose within 1.6 Å r.m.s. deviation of another protein. Only the glucose dehydrogenase structure shows somewhat larger r.m.s deviation, upto 2.1 Å, for the structurally equivalent positions when compared with the other structures. About 40 structurally conserved residues form the characteristic β-barrel of the GroES fold and contribute to its structural core.

Figure 1 shows the overall folding pattern of *E.coli* GroES. Major deviations in pairwise structural superpositions of the different structures occur in various loops connecting the β-strands. Two of these loops characterize the variation of the chaperonin-10 structures from quinone oxidoreductase and alcohol dehydrogenase families. The first, designated as the mobile loop (Figure 1), is a large insertion connecting the first and second β-strands of the barrel in cpn10 and gp31 structures but is absent in the quinone oxidoreductase, glucose dehydro-

Table I. Pairwise r.m.s. deviations (Å) when superimposing structurally equivalent Cα positions

	E.coli QOR	Horse ADH[a]	*Thermoplasma* GDH	*M.leprae* cpn10	*E.coli* GroES	T4 gp31
E.coli QOR	0.0	1.23 (87)[b]	2.07(72)	1.41 (49)	1.51 (51)	1.37 (50)
Horse ADH		0.0	1.79 (53)	1.37 (48)	1.32 (49)	1.52 (53)
Thermoplasma GDH			0.0	1.47 (43)	1.45 (46)	1.65 (48)
M. leprae cpn10				0.0	1.6 (71)	1.53 (61)
E.coli GroES					0.0	1.5 (63)
T4 gp31						0.0

[a]The catalytic subunit of horse ADH superposed very well with the human beta1 ADH and the human sigma ADH catalytic subunits, along their entire lengths, with an r.m.s. deviation of only 0.35 and 0.52 Å, respectively. Hence only the horse ADH was considered as a representative of the ADH class for comparison with the other structures.
[b]The numbers in parentheses represent the equivalent residues considered during the superpositions.

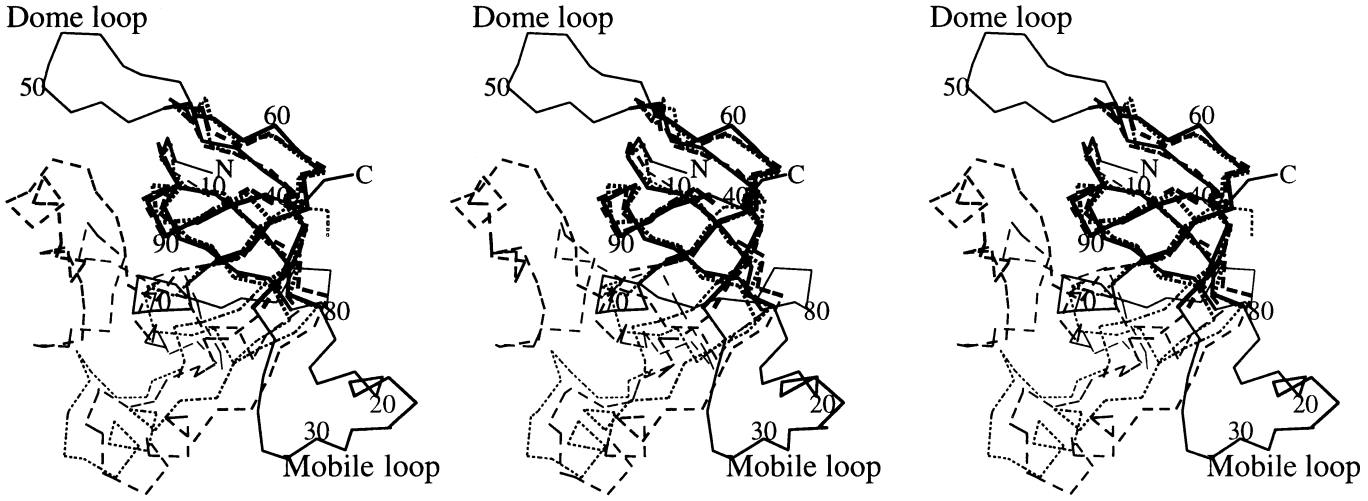

Fig. 1. Structural superposition of the GroES fold proteins. The overall fold of the catalytic domain of ADH (residues 31–160) is shown by the dashed line, QOR (residues 27–112) by the dotted line and the *E.coli* GroES (residues 5–97) by the solid line. Although gp31, GDH and *M.leprae* cpn10 are structurally similar, they are not shown here for clarity. The regions involved in the formation of the β-barrel core are shown in bold. The mobile loop and the dome loop, which are characteristic of the cpn10 family but are absent in the other protein families, are indicated. Every tenth residue along with the N-terminal (N) and the C-terminal (C) of *E.coli* GroES are also labeled. The first four residues of *E.coli* GroES are not shown for clarity. The figure was drawn using MOLSCRIPT (Kraulis, 1991).

genase and alcohol dehydrogenase families. This loop in cpn10 is known to be important for recognition of chaperonin-60, whereby it loses its flexibility during the complex formation between chaperonin-60 and chaperonin-10 (Xu *et al.*, 1997). In the quinone oxidoreductase, glucose dehydrogenase and alcohol dehydrogenase families, the topologically equivalent region forms part of the active site and contacts the nucleotide-binding domain (Murzin, 1996).

Another major difference among the various structures is the insertion of a second loop, the dome loop, connecting the second and the third strands of the barrel in the chaperonin-10 structures (Figure 1). This loop forms a lid of the GroES dome (Hunt *et al.*, 1996; Mande *et al.*, 1996), essentially closing the 'Anfinsen cage' of the GroEL cavity, where unfolded proteins are believed to bind to GroEL. The dome loop is, however, absent in the glucose dehydrogenase, quinone oxidoreductase and alcohol dehydrogenase families. Although gp31 is known functionally to substitute GroES in *E.coli* (Hunt *et al.*, 1997), this loop is interestingly considerably shortened in gp31, resembling the quinone oxidoreductase and alcohol dehydrogenase structures.

All the members of the cpn10 family are known to be homoheptamers. Structure determination of the *E.coli* cpn10 (GroES) and its homologue in *M.leprae* has revealed that the heptamers are arranged in the shape of a dome-like structure and with an approximate sevenfold symmetry (Hunt *et al.*, 1996; Mande *et al.*, 1996). There are two distinct clusters of hydrophobic residues in the cpn10 family, one at the structural core of the monomer and the other at the interface of the monomers in the heptameric assembly. We analyzed sequence conservation at these positions, to check if any definite patterns emerge as fold determinants. As expected, our findings suggest that conservation of residues at the core of the monomer is far more stringent than at the interface.

Among the conserved structural features in all the protein families considered in our sequence analysis is the occurrence of a glycine–aspartate sequence at the end of the second β-strand (positions 62 and 63 of *E.coli* GroES). Absolute conservation of the glycyl–aspartyl dipeptide across these different protein families as divergent as a viral sequence, an archaeon sequence and mammalian sequences is as interesting as the conservation of the fold among these groups. This conserved sequence forms a part of a type II β-turn (also referred to as a glycine turn; Richardson, 1981) positioned at the initiation of the third β-strand. In type II turns, the second residue is normally in the poly-Pro conformation, while the third residue is in the left-handed 3_{10} conformation. In all the structures examined in this study, we find that the glycine

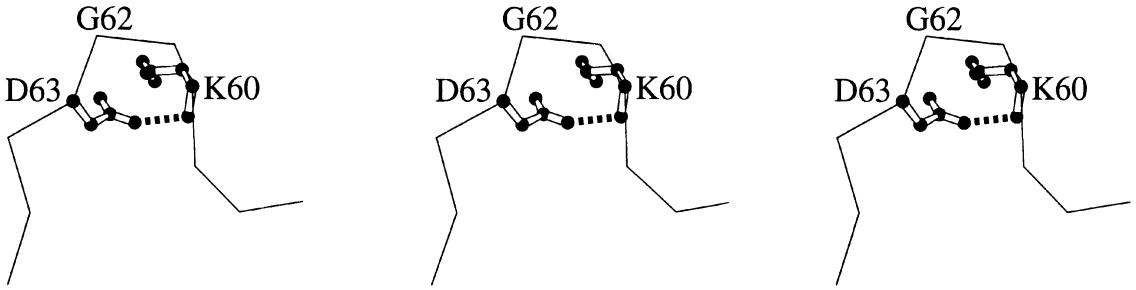

Fig. 2. Stereoview of the conserved Gly–Asp and its role in maintaining the structural integrity of the GroES fold proteins. The side chain carboxylates of the conserved aspartate (Asp63 of *E.coli* GroES in the figure) are involved in hydrogen bonding to the main chain nitrogen of the first residue (Lys60) of a type II β-turn or the glycine turn. This interaction correctly juxtaposes the second and the third β-strands for the formation of the β-barrel core of these proteins. Also shown is the conserved Gly62 of *E.coli* GroES. The corresponding Gly–Asp residues in ADH, QOR, *M.leprae* cpn10, T4 gp31 and GDH occur at positions 86–87, 82–83, 65–66, 66–67 and 85–86, respectively (see text for further details).

Table II. Position specific statistics for occurrence of residues at the eight core positions (the core positions are predominantly occupied by small hydrophobic residues in all the three protein families)[a]

(A) Chaperonin-10: a total of 58 sequences

	G	A	P	V	L	I	M	H	F	Y	W	S	T	C	D	N	E	Q	K	R
10				47		11														
12				40	3	15														
40				44		14														
59				52	4	1						1								
65				49		9														
67					4	4	1		32	17										
84				12	33	6		1	3			1	2							
86				3	24	6	20		5											

(B) Alcohol dehydrogenases: a total of 86 sequences

	G	A	P	V	L	I	M	H	F	Y	W	S	T	C	D	N	E	Q	K	R
10		7		65	6	8														
12				27	9	48	1								1					
40				78		8														
59				29	29	1	1		13	4	9									
65	19	4		57	1	3						2								
67	2	4	61			1							1			2			15	
84		11		4		2			4			1	56	6						2
86	4	15		50	4	6	4						1	1		1				

(C) Quinone oxidoreductases: a total of 12 sequences

	G	A	P	V	L	I	M	H	F	Y	W	S	T	C	D	N	E	Q	K	R
10				8	2	2														
12				4	1	7														
40				7		5														
59		1	1	1		2			6		1									
65				12																
67		1			1					5		1	3	1						
84		4		1	1			3				1		1						1
86		6		2	1	3														

[a]The position numbers correspond to that of the *E.coli* GroES sequence.

occupies the third position of the turn and is in the canonical left-handed 3_{10} conformation. As expected for type II turns, all the four α-carbons appear to be nearly in a plane. The type II turns generally connect two consecutive antiparallel β-strands in protein structures or help the polypeptide reverse its direction (Richardson, 1981). In the GroES fold family, neither of the two cases is observed.

The type II turn seems to be important for maintaining the integrity of the fold, by involving a unique side chain–main chain interaction The side chain carboxylates of the aspartate are involved in hydrogen bonding to the main chain nitrogen of the first residue of the turn (Figure 2), thereby restricting the polypeptide on either side of the β-turn from approaching each other. The aspartate thus correctly juxtaposes the second and third β-strands of the barrel with respect to the core of the protein. In the absence of the aspartate, we hypothesize that the second and third β-strands would form an antiparallel β-sheet, as commonly observed in other protein structures. This hypothesis can easily be tested by site-directed mutagenesis of these two residues.

Occurrence of the 3_{10} helix inserted between the second and the third strands of the β-barrel appears to be a conserved feature of the GroES fold family. The reasons for the conservation of the 3_{10} helix among all the protein structures appear to be intriguing. A detailed sequence analysis and site-directed mutagenesis of the residues involved can shed more light on the role of the 3_{10} helix in the integrity of the fold.

On comparing the three-dimensional structures, we identified eight residues that are shielded from the solvent and form the hydrophobic core of the proteins. These eight residues are seen to be highly conserved across the sequences with very little variation (Table II). Considering the volume of the core to be that of the contributing side chains (Harpaz *et al.*, 1994), average core volumes of the chaperonin-10, quinone oxidoreductase and alcohol dehydrogenase families are 1171.3, 1102.7 and 1043.8 $Å^3$, respectively. Hence all the three families have similar core volumes. The small difference in the volumes of the cpn10 and alcohol dehydrogenase families is due to the predominant occupation by aromatic residues at site 67 of *E.coli* and the corresponding positions of other cpn10 sequences.

Out of the eight identified positions in the core of the GroES fold proteins (Table II), positions 10, 12 and 40 are found to be most conserved, while positions 67, 84 and 86 are seen to be more variable (position numbers corresponding to *E.coli* GroES). Position 67 of *E.coli* GroES has mostly aromatic side chains in the chaperonin sequences, but is largely occupied by Pro residues in the alcohol dehydrogenase sequences.

A majority of the side chains in the hydrophobic core are small, non-polar side chains. The predominantly occurring residue is valine, which is highlighted by the mutation patterns at each of the individual sites. Interestingly, valines at the core positions are seen to be mutable into isoleucines, but not to leucines (Table II). Considering that Ile and Leu have similar side chain volumes, the higher frequency of substitution of Ile by Val was rather unexpected. A probable reason could be the higher β-sheet propensity of Ile and Val than Leu (Wilmot and Thornton, 1988). Another possible reason could be the branching of side chains at the Cβ position in both Val and Ile, while it is at the Cγ position in Leu. Therefore, in the event of compensatory mutations, Val to Leu or Ile to Leu mutation would appear to be non-conservative. A more plausible explanation can be sought from the genetic code. A single base mutation to convert Ile to Val requires a transition mutation at the first position of the triplet codon, whereas for Ile to Leu it would be a transversion mutation. Since transition mutation rates have a bias over transversion rates (Huelsenbeck and Rannala, 1997), the substitution of Ile/Val by Leu would be less probable. A similar observation is also noted from the amino acid substitution matrices of Dayhoff (1978).

The interesting similarities between the different GroES fold proteins, therefore, suggest a possible evolutionary relatedness among them. Occurrence of ligand binding at the topologically equivalent site may seem to suggest a common evolutionary origin of the four protein families (Murzin, 1996), such as that commonly found in TIM barrel proteins (Farber, 1993). The quinone oxidoreductase and alcohol dehydrogenase proteins do indeed show high sequence similarities, reinforcing the conclusions regarding evolutionary divergence. The divergence of sequences may have preceded divergence of different kingdoms and therefore losing trace of sequence similarities between the chaperonin-10 and other families. However, the evolutionary pressure seems to have preserved the amino acids responsible for core formation, and also the glycyl–aspartyl dipeptide sequence for maintaining the integrity of the fold. Further detailed comparison of other β-barrel classes of proteins can help in the identification of such fold determinants, the importance of which can be confirmed beyond doubt by various tools including site-directed mutagenesis. Nevertheless, the identification and importance of such fold determinants should provide the necessary impetus in the prediction of tertiary structures from first principles.

Acknowledgements

We thank Alexey Murzin for useful comments and suggestions on the manuscript, Garry L.Taylor for providing the coordinates of glucose dehydrogenase and the Bioinformatics facility of Institute of Microbial Technology for access to computers. B.T. is a CSIR Junior Research Fellow.

References

Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54.
Dayhoff,M. (1978) *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345–358.
Eklund,H., Nordstrom,B., Soderlund,E.Z.G., Ohlsson,I., Soderberg,T.B.B.-O., Tapia,O. and Branden,C.-I. (1976) *J. Mol. Biol.*, **102**, 27–59.
Farber,G.K. (1993) *Curr. Opin. Struct. Biol.*, **3**, 409–412.
Harpaz,Y., Gerstein,M. and Chothia,C. (1994) *Structure*, **2**, 641–649.
Huelsenbeck,J.P. and Rannala,B. (1997) *Science*, **276**, 227–232.
Hunt,J.F., Weaver,A.J., Landry,S.J., Gierasch,L. and Deisenhofer,J. (1996) *Nature*, **379**, 37–45.
Hunt,J.F., Saskia,M.V., Henry,L. and Deisenhofer,J. (1997) *Cell*, **90**, 361–371.
Hurley,T.D., Bosron,W.F. and Stone,C.L. (1994) *J. Mol. Biol.*, **239**, 415–420.
John, J., Crennell,S.J., Hough,D.W., Danson,M.J. and Taylor,G.L. (1994) *Structure*, **2**, 385–393.
Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard, M. (1991) *Acta Crystallogr.*, A**47**, 110–119.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **90**, 946–950.
Mande,S.C., Mehra,V., Bloom,B.R. and Hol,W.G.J. (1996) *Science*, **271**, 203–207.
Murzin,A.G. (1996) *Curr. Opin. Struct. Biol.* **6**, 386–394.
Murzin,A.G., Lesk,A.M. and Chothia, C. (1994a) *J. Mol. Biol.*, **236**, 1369–1381.
Murzin,A.G., Lesk,A.M. and Chothia, C. (1994b) *J. Mol. Biol.*, **236**, 1382–1400.
Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 167–339.
Sussman,J.L., Lin,D., Jiang,J., Manning,N.O., Prilusky,J., Ritter,O. and Abola,E.E. (1998) *Acta Crystallogr.*, D**54**, 1078–1084.
Thomson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
Thorn,J.M., Barton,J.D., Dixon,N.E., Ollis,D.L. and Edwards,K.J. (1995) *J. Mol. Biol.* **249**, 785–799.
Wilmot,C.M. and Thornton,J.M. (1988) *J. Mol. Biol.*, **203**, 221–232.
Xie,P., Parsons,S. H., Speckhard,D.C., Bosron,W.F. and Hurley,T.D. (1997) *J. Biol. Chem.*, **272**, 18558–18563.
Xu,Z., Horwich,A.L. and Sigler,P.B. (1997) *Nature*, **388**, 741–750.

Received March 19, 1999; revised June 11, 1999; accepted July 5, 1999