PROTEINS: Structure, Function, and Genetics 12:278–298 (1992)

Comparison of the Hemocyanin β-Barrel With Other Greek Key β-Barrels: Possible Importance of the "β-Zipper" in Protein Structure and Folding

Bart Hazes and Wim G. J. Hol
BIOSON Research Institute, Department of Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands

ABSTRACT **The Greek key β-barrel topology is a folding motif observed in many proteins of widespread evolutionary origin. The arthropodan hemocyanins also have such a Greek key β-barrel, which forms the core of the third domain of this protein. The hemocyanin β-barrel was found to be structurally very similar to the β-barrels of the immunoglobulin domains, Cu,Zn-superoxide dismutase and the chromophore carrying antitumor proteins. The structural similarity within this group of protein families is not accompanied by an evolutionary or functional relationship. It is therefore possible to study structure–sequence relations without bias from nonstructural constraints. The present study reports a conserved pattern of features in these Greek key β-barrels that is strongly suggestive of a folding nucleation site. This proposed nucleation site, which we call a "β-zipper," shows a pattern of well-conserved, large hydrophobic residues on two sequential β-strands joined by a short loop. Each β-zipper strand is near the center of one of the β-sheets, so that the two strands face each other from opposite sides of the barrel and interact through their hydrophobic side chains, rather than forming a hydrogen-bonded β-hairpin. Other protein families with Greek key β-barrels that do not as strongly resemble the immunoglobulin fold—such as the azurins, plastocyanins, crystallins, and prealbumins—also contain the β-zipper pattern, which might therefore be a universal feature of Greek key β-barrel proteins.**

Key words: folding nucleation, hydrophobic cluster, conserved loop length, structure-sequence relationship, sequence patterns

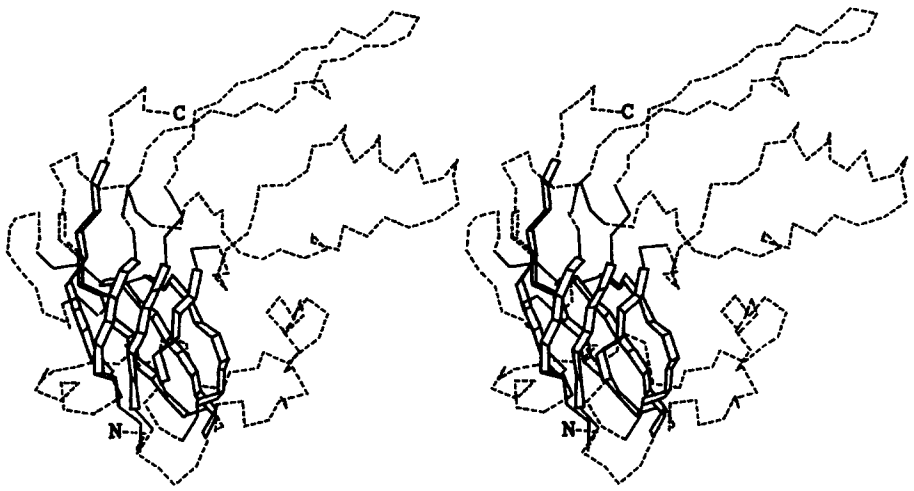
INTRODUCTION

Hemocyanins are large dioxygen-transporting molecules occurring in many species of arthropods and molluscs. The crystal structure of an arthropodan hemocyanin, that of the spiny lobster *Panulirus interruptus,* has been solved in our laboratory at 3.2 Å resolution.[1,2,3] Each subunit of arthropodan hemocyanins consists of three domains. In this article we focus on the third hemocyanin domain (hereafter also referred to as "hecy-3d"). The core of this domain is formed by a seven-stranded Greek key β-barrel, from which several long loops extend. The barrel can also be described as two β-sheets facing each other. The β-barrel of hecy-3d resembles that of the immunoglobulins (IgG) and of Cu,Zn-superoxide dismutase (SOD), as has been described briefly before.[1,3] There is, however, a fourth structurally related protein family, formed by neocarzinostatin, actinoxanthin, and auromomycin. These proteins form a class of chromophore carrying antitumor proteins (CCAPs), which damage DNA and are consequently highly toxic.[4]

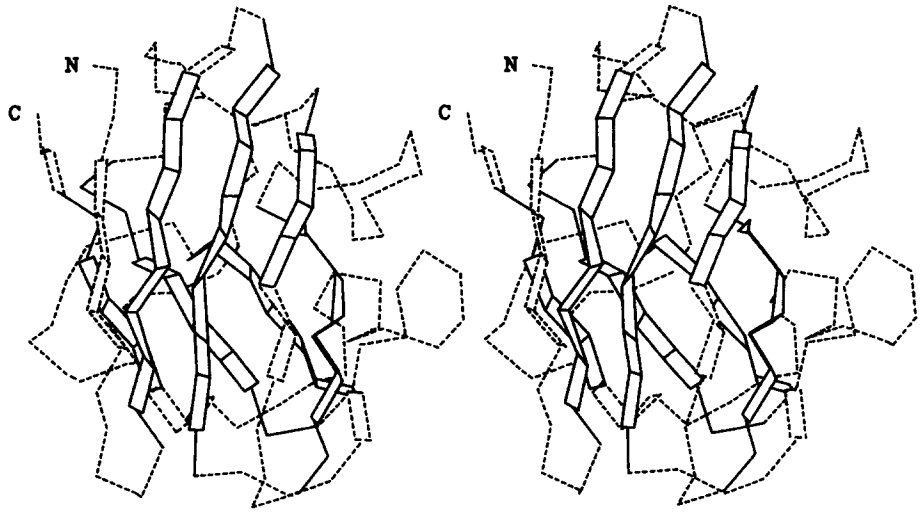
Various studies on the protein families mentioned above have been reported in the literature. The similarity between SOD and the IgG domains has been described by Richardson et al.[5] The variation of sequences within the SOD family has been discussed in detail in a paper by Getzoff et al.[6] The immunoglobulins have also been studied intensively and papers relating structures and sequences are for instance those of Lesk and Chothia[7] and Taylor.[8] There is, however, no comparative study including all four protein families, relating structure to sequence, which is the goal of the present study. Pictures of the β-barrels of all four protein families are given in Figure 1A–E. Topology diagrams are given in Figures 2A–D. Their close topological relationship is evident from these figures.

At the amino acid sequence level these four families of protein domains have little similarities, as will be discussed below, and hence no obvious relationship to a common ancestor is apparent. Also functionally the proteins are very different. SOD has catalytic activity, whereas the immunoglobulins can specifically bind to an antigen by means of their

Received September 6, 1990; revision accepted July 9, 1991.
Address reprint requests to Wim G.J. Hol, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

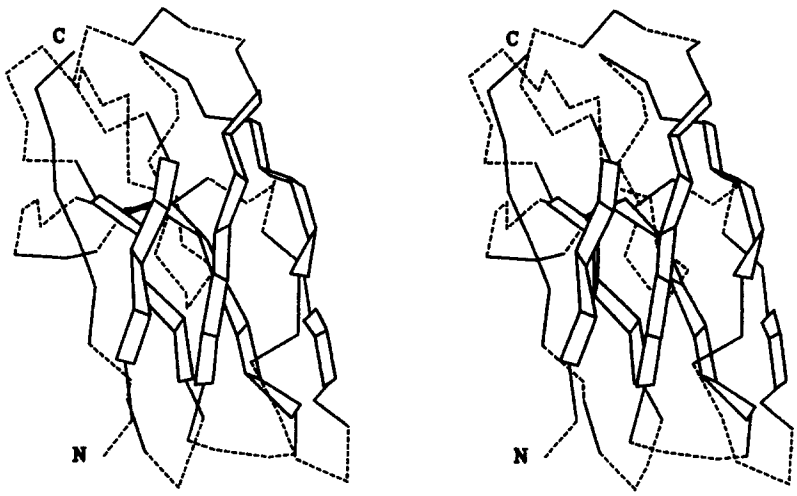


A HECY BARREL HECY BARREL

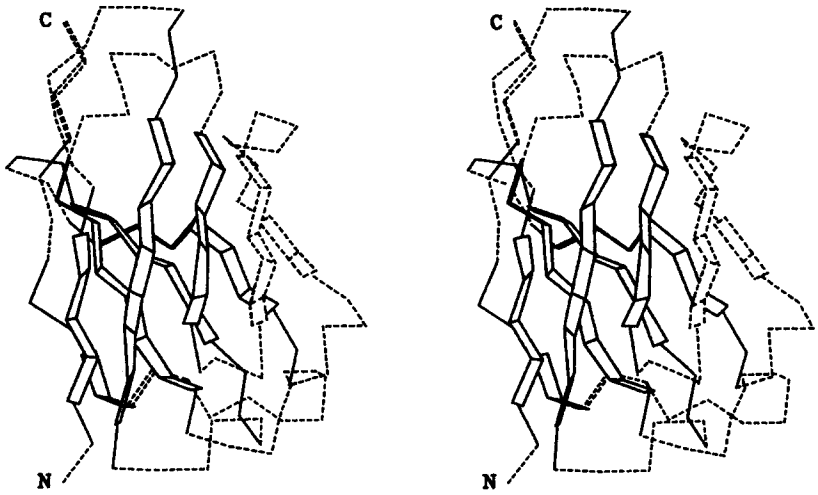


1B SOD BARREL SOD BARREL

Fig. 1. C_α tracing of the β-barrels with β-strands depicted as ribbons. Residues that are structurally equivalent with hecy-3d are drawn in continuous lines. For hecy-3d itself, continuous lines are used for all residues that have an equivalent in at least one of the other structures. All other residues are linked by dashed lines. **(A)** Hecy-3d, **(B)** SOD, **(C)** ACX, **(D)** IgG variable domains represented by $KOLV_H$, **(E)** IgG constant domains represented by $KOLC_L$. Figure continues through page 281.

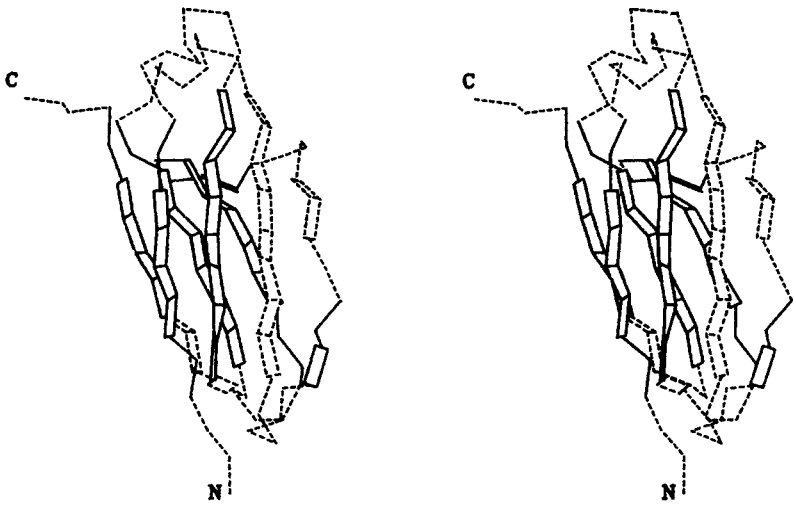


C ACX BARREL ACX BARREL



1D KOLVH BARREL KOLVH BARREL

Fig. 1C and D. Legend appears on page 279.



1E KOLCL BARREL KOLCL BARREL

Fig. 1E. Legend appears on page 279.

hypervariable loops. The CCAPs also have a binding function, but they bind chromophores in a different way at the side of the barrel.[9] For hecy-3d there is as yet no known function. This situation with four similar structures, but unrelated sequences and functions, provides an interesting possibility to study structure–sequence relations. Questions we would like to answer are, how similar are the β-barrels and which sequence properties are essential in determining the tertiary structure. In our comparisons we have also investigated the length of the loops connecting the β-strands. This led us to consider subsequently the folding of other Greek key β-barrel proteins with different topologies, such as the γ-crystallins, prealbumins, azurins, and the related plastocyanins.

MATERIALS AND METHODS

The protein structures used in this study are listed in Table I. The CCAPs were represented by actinoxanthin (ACX), since the recently determined high resolution structure of auromomycin[9] is not yet included in the protein data bank. From several IgG structures elucidated, the KOL Fab fragment was chosen, since it had the highest resolution and the lowest *R*-factor. This Fab fragment was divided into its four domains: variable light (V_L), constant light (C_L), variable heavy (V_H), and constant heavy (C_H). In the latter part of this paper the protein families, azurin, plastocyanin, γ-crystallin, and prealbumin, are also analyzed, in connection with a proposed general folding mechanism for Greek key β-barrel proteins. The members of the families used for this structural analysis are also included in Table I.

Amino acid sequence analysis was carried out with the sequences of the structures listed in Table I, extended with the sequences of other members of these protein families. For the hemocyanins the 7 aligned sequences as given by Linzen et al.[17] and the recently published *Panulirus interruptus* c chain sequence[18] were used. For the SOD family the sequence alignment of 12 species as given by Getzoff et al.[6] was taken. The IgG sequences considered comprise the list of 60 sequences of variable heavy domains as presented by Taylor[8] and the 24 complete sequences of constant domains presented in the Dayhoff atlas[19] as "alignment 30." For the CCAPs the alignment of 3 sequences as given by Samy et al.[4] was taken.

Sequences for azurin, plastocyanin, γ-crystallin, and prealbumin were obtained from the PIR database and aligned manually. The number of sequences used per family was 7, 16, 13, and 4 for azurins, plastocyanins, γ-crystallins, and prealbumins, respectively. Since γ-crystallin consists of two β-barrels per chain, the 13 sequences code for 26 barrels of which two barrels had an incomplete sequence.

Structural superpositions were first carried out visually on an Evans & Sutherland PS390 graphics system using the program FRODO.[20] At this stage each structure was aligned to the hecy-3d structure.

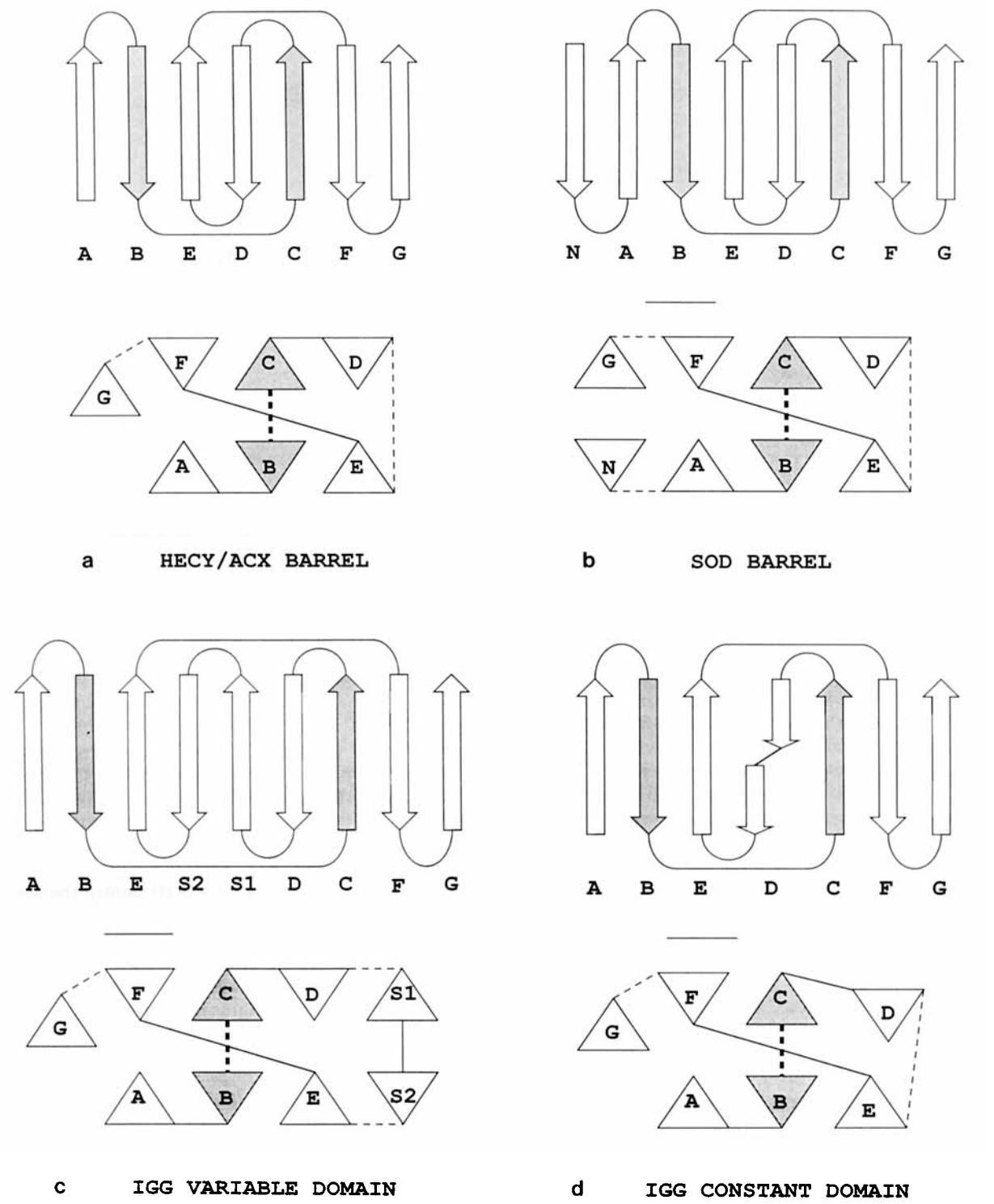


Fig. 2. Topology diagrams of **(a)** hecy-3d and ACX, **(b)** SOD, **(c)** IgG variable domains, **(d)** IgG constant domains. The arrow diagrams can be formed back into the Greek key β-barrel by folding the right half of the picture backward. The triangle diagram shows a top view of the barrel. Triangles with their apex at the top are running toward the reader. Triangles with their apex at the bottom run into the paper. Dashed lines run along the bottom of the barrel and solid lines run on top. The accentuation of the "B-strand-loop-C-strand" denotes its special function in the formation of the Greek key β-barrel as is proposed in this paper.

TABLE I. Structures Used

Protein	PDB code	Resolution (Å)	*R*-factor (%)	Reference
*P. interruptus** hemocyanin	—	3.2	24.5	Volbeda and Hol[3]
Bovine Cu,Zn-superoxide dismutase	2SOD	2.0	25.6	Tainer et al.[10]
Actinoxanthin	1ACX	2.0	—†	Pletnev et al.[11]
KOL Fab fragment	1FB4	1.9	18.9	Marquart et al.[12]
Calf eye lens γ-crystallin	1GCR	1.6	23.0	Summers et al.[13]
Human prealbumin	2PAB	1.8	29.0	Blake et al.[14]
P. aeruginosa azurin	1AZU	2.7	35.0	Adman and Jensen[15]
Poplar plastocyanin	1PCY	1.6	17.0	Guss and Freeman[16]

*Coordinates have been deposited at the protein databank.
†The resolution as given in the coordinate file is 2.0 Å. The *R*-factor is not specified.

Several superposition alternatives were tested manually to optimize both the fit and the number of superimposed residues. Subsequently a refined fit was obtained by the following sequence of steps:

1. Select stretches of three or more consecutive residues that each have C_α positions deviating less than 3.5 Å from three consecutive hecy-3d C_α positions. (Where two consecutive residues are equivalent to hecy-3d and are separated from another equivalenced stretch by only one residue, then these are also included.)
2. Perform a least-squares superposition of these C_α positions.[21]
3. Repeat these two steps until no new equivalences are made anymore.

Once the structural alignment of each protein studied with hecy-3d was obtained, this formed the starting point for the superpositioning of the non-hecy-3d structures onto each other.

Secondary structure assignments and surface accessibilities were obtained with the DSSP program.[22] Loop lengths were defined as the number of residues between β-strands as given by the DSSP program. Side chain contacts were calculated with a local program. Side chains were considered to be in contact with each other if they had atoms within a distance of 4.0 Å.

RESULTS

Structural Comparison and Sequence Alignment

The seven representatives of the four Greek key-β-barrel containing protein families superimpose with rms values varying from 1.28 to 2.11 Å for 34 to 90 equivalent C_α atoms (Table II). As expected, the group of related IgG domains shows the higher similarity. However, the worst superposition between pairs of IgG domains (C_H vs V_H) shows a similar rms deviation as found for the comparisons of hecy-3d with the other, unrelated, structures. This pair of IgG domains also has a very low sequence identity. In comparison, hecy-3d shows a rather high sequence identity to C_L and ACX (17.6 and 17.9%, respectively, for equivalenced residues). We assume however, that this is due to similar restrictions imposed by the similar structures, and is not a result of a common ancestor.

Based on the structural superposition, a sequence alignment was made, using hemocyanin as a template. This alignment is shown in Figure 3. Residues in extended conformation are underlined. Residues whose side chains contribute to the protein core are indicated by a "+" and the percentage of hydrophobic residues (Gly, Ala, Cys, Val, Ile, Leu, Pro, Met, Phe, Tyr, or Trp) in all 107 sequences is given for each position. This percentage was calculated by first taking the average per family, followed by averaging these averages. This procedure was used in order to obtain equal contributions per family. In addition the IgG variable and constant domains were treated as two separate groups, since otherwise the large number of variable domain sequences would dominate the percentage hydrophobicity for the IgG family. Figure 3 also introduces a new sequence numbering scheme, to simplify the description of comparisons below.

From Figure 3 it is clear that the IgG E-strands, with the exception of V_H, do not superimpose well on the other structures. The E-strands are, however, hydrogen bonded to the B-strand, and in this respect should be homologous to the E-strands in the other families. Nevertheless their position and orientation have changed such that their C_α atoms no longer superimpose within the 3.5 Å cut-off limit. Also the IgG D-strands behave somewhat abnormally. In the constant domains the N-terminal part of the D-strand is hydrogen bonded to the C-strand, whereas the C-terminal part is hydrogen bonded to the E-strand. In the non IgG structures the D-strand is

TABLE II. Results of Superpositioning of Greek Key β-Barrel Domains*

	HECY	SOD	ACX	V_L	C_L	V_H	C_H
HECY	—	5.2	17.9	14.3	17.6	4.6	15.9
SOD	1.78 (58)	—	8.2	8.8	9.3	12.0	4.9
ACX	2.02 (67)	1.86 (49)	—	14.6	9.1	9.2	5.7
V_L	2.01 (49)	2.10 (34)	1.99 (48)	—	19.4	27.8	18.2
C_L	1.76 (51)	2.10 (54)	1.58 (44)	1.33 (62)	—	22.2	33.7
V_H	2.11 (65)	2.02 (50)	1.86 (54)	1.49 (90)	1.56 (63)	—	9.8
C_H	2.00 (44)	2.02 (41)	1.92 (35)	1.42 (55)	1.28 (89)	1.90 (61)	—

*Below the diagonal the rms deviation of superimposed C_α atoms is given (Å). The number of superimposed residues is given in parentheses. Above the diagonal the sequence identity of the superimposed residues is given (%).

```
CODE        A1  A2   A3   A4   A5   A6   A7   A8   A9  A10  A11
                +         +                   +
PHOBIC      72  96   11  100   63   48   79   35   94   23   69
HECY 410 GLY MET  VAL  VAL  ASN  GLY  VAL  ALA  ILE  ASP  GLY 420
SOD   14      VAL  GLN  GLY       THR  ILE  HIS  PHE  GLU  ALA  22
ACX    2                      PRO  ALA  PHE  SER  VAL  SER  PRO   8
VL     4                 LEU       THR  GLN  PRO                 7
CL   114 ASN PRO  THR  VAL       THR  LEU  PHE  PRO  PRO       122
VH     2      VAL  GLN  LEU       VAL  GLN  SER  GLY             8
CH   123 GLY PRO  SER  VAL       PHE  PRO  LEU                  129

CODE        B1  B2   B3   B4   B5   B6   B7   B8   B9  B10  B11  B12
                +         +         +         +         +
PHOBIC      59  18   45   90   14   98   30  100   47   98    5   58
HECY 457 HIS ASN  GLU  PHE  THR  TYR  LYS  ILE  THR  MET  SER  ASN  468
SOD   25      ASP  THR  VAL  VAL  VAL  THR  GLY  SER  ILE  THR  GLY   35
ACX   15 GLY GLN  SER  VAL  SER  VAL  SER  VAL[1] ALA  ALA            24
VL    16      GLN  ARG  VAL  THR  ILE  SER  CYS  THR  GLY  THR  SER   26
CL   132           ALA  THR  LEU  VAL  CYS  LEU  ILE  SER  ASP       140
VH    15 GLY ARG       LEU  ARG  LEU  SER  CYS  SER  SER  SER  GLY   26
CH   139      GLY  THR  ALA  ALA  LEU  GLY  CYS  LEU  VAL  LYS  ASP  149

CODE        C1  C2   C3   C4   C5   C6   C7   C8   C9  C10  C11  C12  C13
                +         +         +         +
PHOBIC      22  29   69   42   99   13  100   66   58   52   93   66   82
HECY 474 ARG LEU  ALA  THR  PHE  ARG  ILE  PHE  LEU  CYS  PRO  ILE  GLU  486
SOD   40      ASP  HIS  GLY  PHE  HIS  VAL  HIS                          46
ACX   27           GLU  THR  TYR  TYR  ILE  ALA  GLN  CYS  ALA            35
VL    31      SER  ILE  THR  VAL  ASN  TRP  TYR  GLN  GLN  LEU  PRO      41
CL   144 GLY ALA  VAL  THR  VAL  ALA  TRP  LYS  ALA  ASP                 153
VH    31      SER  TYR  ALA  MET  TYR  TRP  VAL  ARG  GLN  ALA  PRO  GLY  42
CH   153 GLN PRO  VAL  THR  VAL  SER  TRP  ASN                          160

CODE        D1  D2   D3   D4   D5   D6   D7   D8   D9  D10  D11  D12  D13  D14
                                  +    +         +         +
PHOBIC      87  88   28   50   86   22   26   96   24   67   15   81   30   45
HECY 501 PHE CYS  ILE  GLU  LEU  ASP  LYS  PHE  PHE  GLN  LYS  VAL  PRO  SER  514
SOD   80                 GLY  ASP       GLY  ASN  VAL  THR  ALA  ASP  LYS   89
ACX   42 ALA CYS  ASN[2]      ALA  THR  SER  PHE  THR  THR  ASP                54
VL    44 ALA PRO  LYS  LEU  LEU  ILE  TYR  ARG                               51
CL   154      GLY  SER  PRO  VAL  157                 165  LYS  PRO  SER     167
VH    44 GLY LEU  GLU  TRP  VAL  ALA  ILE  ILE  TRP  ASP                      53
CH   163           ALA  LEU  THR                                            165
```

```
CODE        E1  E2   E3   E4   E5   E6   E7   E8
                +         +         +
PHOBIC       1  26   97   35   56   12   54   23
HECY 517 GLU THR  ILE  GLU  ARG  SER  SER  LYS  524
SOD   95      VAL  ASP  ILE  VAL  ASP  PRO  100
ACX   58 ALA ALA  SER  PHE  SER  PHE       63
VL
CL
VH    77 ASN THR  LEU  PHE       GLN  MET   83
CH

CODE        F1  F2   F3   F4   F5   F6   F7   F8   F9  F10
                +         +         +
PHOBIC      79  46   79   69  100   73   62   21   28   63
HECY 580 PHE ASN  LEU  TYR  VAL  ALA  VAL  THR  ASP  GLY  589
SOD  111 ILE GLY  ARG  THR  MET  VAL  VAL  HIS  GLU  LYS  120
ACX   89 CYS ASN  LEU  GLY  ALA  GLY  ASN                   95
VL    85 SER ASP  TYR  TYR  CYS  ALA  SER  TRP              92
CL   191 ARG SER  TYR  SER  CYS  GLN  VAL  THR  HIS        199
VH    92 GLY VAL  TYR  PHE  CYS  ALA  ARG  ASP  GLY       100
CH   197 GLN THR  TYR  ILE  CYS  ASN  VAL  ASN  HIS       205

CODE        G1  G2   G3   G4   G5   G6   G7   G8   G9  G10
                +         +         +         +
PHOBIC      66  71   63   53   65   61   85   38   66   46
HECY 641 ASN ILE  LYS  HIS  VAL  VAL  VAL  LYS  ILE  VAL  650
SOD  143 ALA CYS  GLY  VAL  ILE  GLY                      148
ACX   98 LEU ASN       GLY  HIS  VAL  ALA  LEU  THR  PHE  107
VL   100 VAL PHE  GLY  THR  [3]LYS  VAL  THR              109
CL   204 VAL GLU  LYS  THR  VAL  ALA  PRO                 210
VH   115 TYR TRP  GLY  GLN  GLY[3] PRO  VAL  THR  VAL  SER 125
CH   212 VAL ASP  LYS  ARG  VAL  GLU                      217

4
PHOBIC      94  38   0              57   55   33              100  50   0
HECY 422 LEU ILE  THR  424  497  GLU  ALA  ARG  499  570  LEU  PRO  LYS  572
ACX   12 ALA SER  ASP   14   39  GLY  GLN  ASP   41   65  VAL  ARG  LYS   67
```

Fig. 3. Sequence alignment based on structural superpositioning. Residues in extended conformation are underlined. Residues with side chains that contribute to the protein core are marked with a "+." The line PHOBIC gives the percentage of hydrophobic residues at each position (for the definition of this percentage, see text). [1]The sequence given by Samy et al. (1983) has a glycine at this position. However, in the PDB structure it is a valine. [2]Residues 42–44 are separated from residues 48–54 by a tripeptide. [3]IgG variable domains have an additional residue in the G-strand due to a β-bulge. [4] Hecy-3d and ACX have another nine residues in common, that do not belong to the main part of the β-barrel.

fully hydrogen bonded to the C-strand. In the variable domains a 'normal' D-strand is present. It should, however, be noted that the D-strand in this case is part of the hypervariable loop (see Fig. 2C).

The superposition of SOD and an IgG as described by Richardson et al.[5] is not identical to the alignment shown in Figure 3. In spite of these differences, the accuracy of both alignments is similar (in terms of rms values and number of superimposed residues) for the comparisons of SOD with three of the four IgG domains. Only for the superpositioning of the variable heavy domain, our alignment might be better, having both a lower rms value (2.02 vs 2.28 Å) and a larger number of superimposed residues (50 vs 45).

Amino Acid Contacts

After the sequence alignment, various properties of amino acids at equivalent positions were examined. An analysis of the contacts made between residue pairs in the β-barrel showed that not a single contact is conserved when considering all structures. It is therefore clear that a specific feature like, e.g., the conserved disulfide bridge in the IgG do-

TABLE III. Loop Lengths Between β-Strands*

	AB	BC	CD	DE	EF	FG
HECY	41–45	6–6	18–19	4–4	45–55	50–53
SOD	2–4	4–5	34–43	5–5	11–14	18–22
ACX	9–9	3–4	15–16	7–7	24–25	4–5†
IgG V	9–10‡	8–11	5–6	21–28	8–8	1–15
IgG C	10–13	4–7	4–9	4–6	7–12	3–4

*Loop length is defined as the number of residues between strands as given by the DSSP program.
†The last β-strand in ACX is not listed as such by the DSSP program. Strand boundaries were chosen in analogy to the hemocyanin structure.
‡The first β-strand in V_L is not listed as such by the DSSP program. Strand boundaries were chosen in analogy to the V_H structure.

mains is not a generally conserved feature of this type of Greek key β-barrels.

Amino Acid Properties

Next, the physical properties of single amino acids were analyzed. From studies of protein cores of families of Greek key β-barrel proteins[7,23] and mutants of λ repressor,[24] it appeared that the hydrophobic character of core residues is strongly conserved. Also the total volume of the side chains of core residues is conserved. The volume of individual core residues appeared to be, however, very variable. Volume and degree of hydrophobicity therefore seem to be important properties to study in more detail.

A quick glance at the aligned sequences showed that the property "charge" is not generally important, since none of the positions has a conserved charge, not even when the sign of the charge is ignored. Consequently we focused on the following three properties, using all 107 available sequences.

Hydrophobicity

Strongly conserved hydrophobicity (> 85% in Fig. 3) is found for the following positions: A2–A4–A9, B4–B6–B8–B10, C5–C7–C11, D1–D2–D5–D8, E3, F5, G7. Most of these residues contribute with their side chains to the hydrophobic core. Hydrophilic residues sometimes occupy positions denoted as core positions ("+" in Fig. 3). This occurs most often in strands A, D, E, and G, which form the edges of the sheets. At the edge of the sheet the polar side chain headgroups can still be pointing away from the hydrophobic core toward the solvent. Also at the beginning and end of β-strands hydrophilic residues can be allowed to occupy core positions, depending on the local structure. The expected pattern of hydrophobic residues in a β-strand at positions i, $i+2$, $i+4$ . . . resulting in a hydrophobic face[25] is found in most strands when looking on a per family basis. Between families the pattern can, however, differ in length or can be shifted by two residues. Comparing all families together therefore results largely in the loss of this feature. Only the B-strand is well conserved in this respect.

Hydrophilicity

The hydrophilicity is less well conserved in the sequences studied. Only positions A3, B5–B11, C6, D11, and E1–E6 have a predominant hydrophilic character (< 15% hydrophobic in Fig. 3).

Side chain volume of core residues

From an analysis of side chain volumes, we found that at positions B6, C5, and C7 not a single alanine or glycine was observed in all 107 sequences. Also position B8 is predominantly occupied by rather large hydrophobic residues. The only exception is SOD, which will be discussed later.

Comparison of Loop Lengths

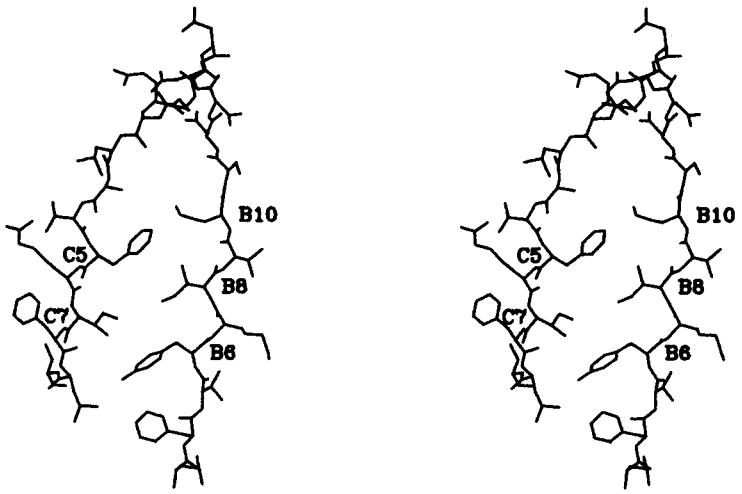
Since the similarity of amino acids appeared to be very limited indeed, we also looked at the lengths of loops. It appears that all but one of the loops greatly vary in length. Some loops can consist of over 50 residues in one protein, while in another the length is only 2 residues. Only one of the connecting loops, the BC-loop, is of a quite constant length. It is always between 3 and 11 residues in all protein families compared (Table III).

DISCUSSION

Common Features Observed

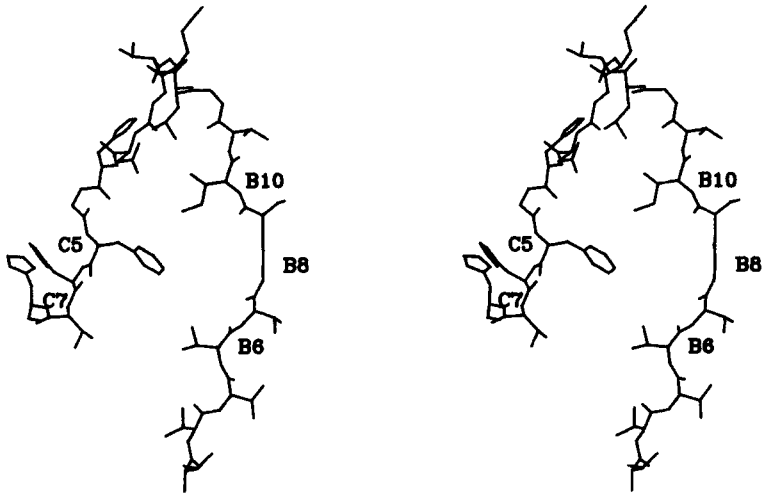
The extent of amino acid properties conserved by all families as described above appears to be rather low. Structurally strands B, C, and F are most homologous, but this could be expected since these strands occupy positions at the center of the sheets. Also the relatively more extensive conservation of sequence properties in strands B and C can be partly ascribed to the position of these strands within the barrel. Interestingly, the conserved short loop between strands B and C and the absence of small residues at several positions, as mentioned above, were not expected and could not be explained from structural constraints in the folded protein. The amino acid sequence has, however, to be compatible with both the thermodynamic (a stable native structure) and kinetic (protein folding) constraints.

In the next sections, the relation between the



A

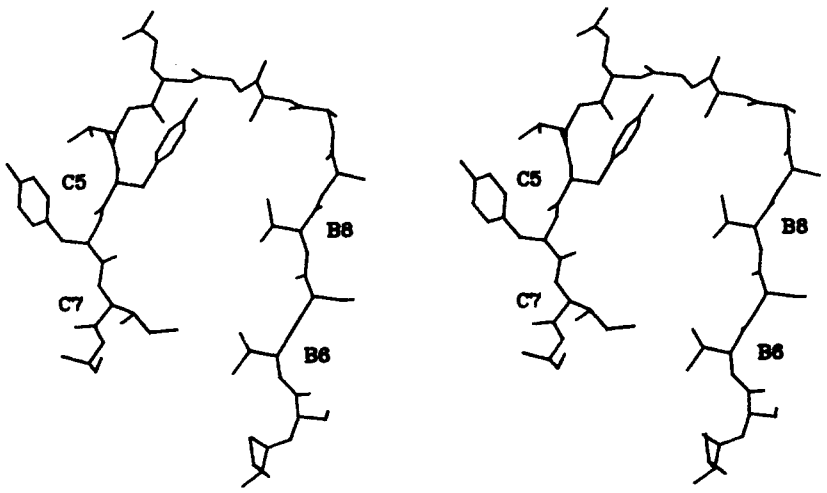
HECY NUCLEUS HECY NUCLEUS



B

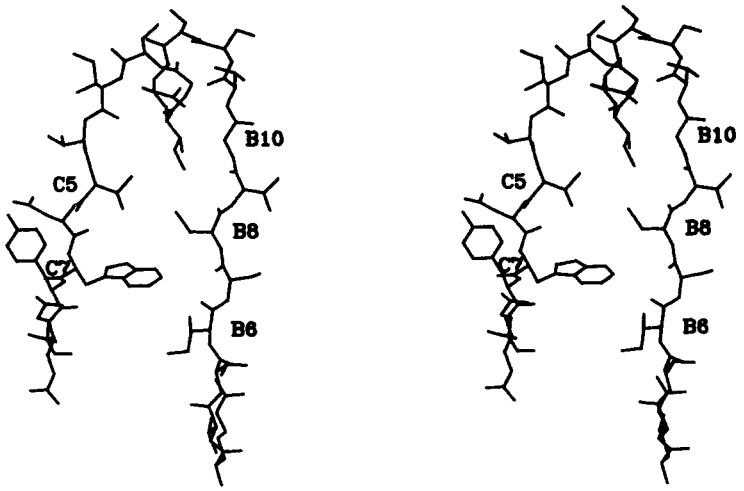
SOD NUCLEUS SOD NUCLEUS

Fig. 4. The β-zipper structures as observed in **(A)** hecy-3d, **(B)** SOD, **(C)** ACX, **(D)** $KOLV_L$, **(E)** $KOLV_H$, **(F)** $KOLC_L$, **(G)** $KOLC_H$. Figure continues through page 289.



C

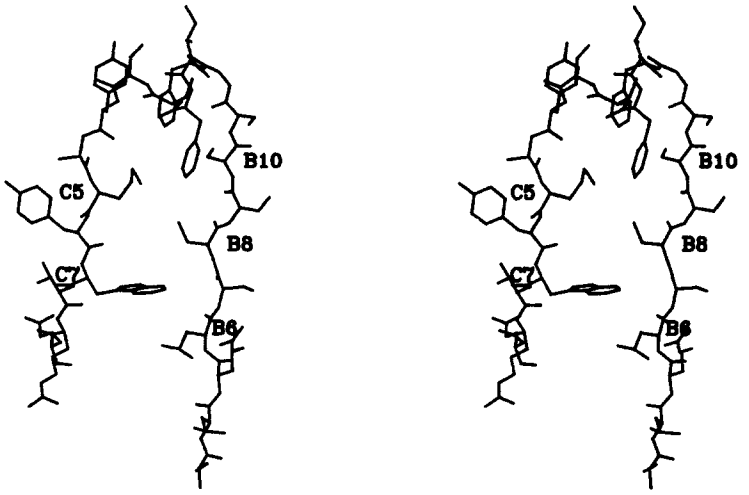
ACX NUCLEUS ACX NUCLEUS



D

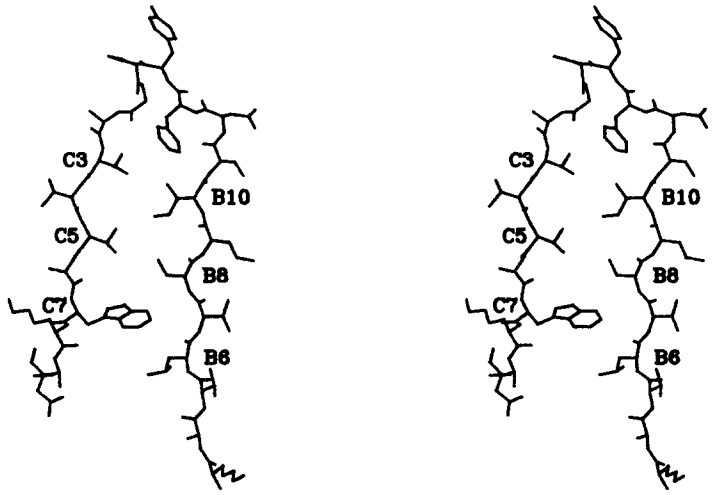
KOLVL NUCLEUS KOLVL NUCLEUS

Fig. 4C and D. Legend appears on page 286.



E

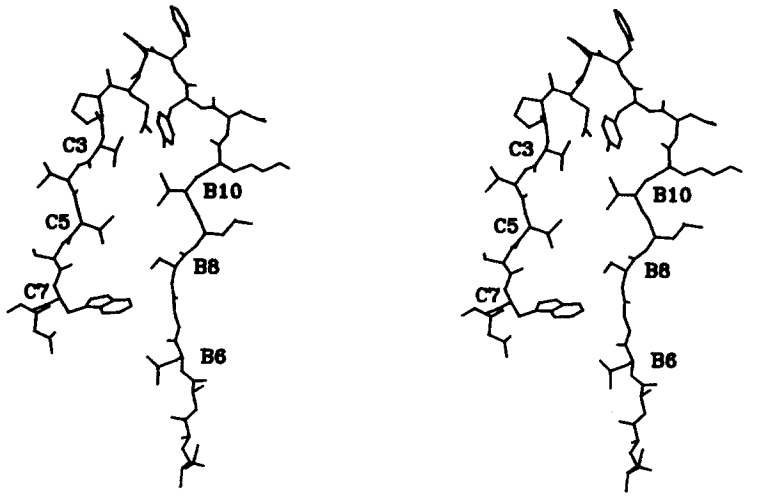
KOLVH NUCLEUS KOLVH NUCLEUS



F

KOLCL NUCLEUS KOLCL NUCLEUS

Fig. 4E and F. Legend appears on page 286.



G

KOLCH NUCLEUS KOLCH NUCLEUS

Fig. 4G. Legend appears on page 286.

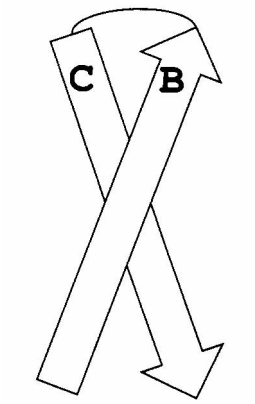


Fig. 5. Schematic drawing of the β-zipper, showing the angle between the strands as a result of β-strand twist.

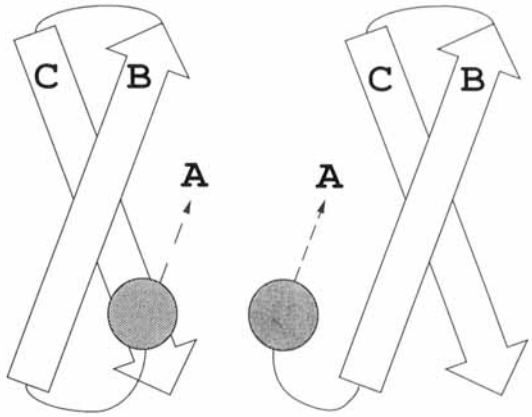


Fig. 6. The two possible positions for adding the A-strand to the β-zipper. Note that in the observed handedness, the addition of the A-strand leads to an immediate stabilization by contacts with both the B-strand and the C-strand.

conserved short BC loop and the absence of small residues at positions B6, C5, and C7 with the kinetics of protein folding will be discussed.

Protein Folding

A number of different theoretical models exist that describe the sequence of events occurring during protein folding. Two basic models are discussed by Karplus and Weaver.[26] One of the models, the "diffusion–collision model," assumes that two or more metastable nuclei are formed independently, which then diffuse together and coalesce into a stable native-like structure. The other model, the "random search-nucleation and chain propagation model,"[27–29] assumes that a single nucleus is formed, which then grows by sequential addition of chain segments. A combination of both processes might of course occur during folding.

A common characteristic of both models is that protein folding is initiated by the formation of one or more folding nuclei. An important feature of such nuclei is that they contain amino acids that are close together in the linear sequence. One reason for this

TABLE IV. Observed β-Zipper Patterns in Greek Key β-Barrels*

Protein	NS†	First strand pattern B4	B6	B8	B10	Range of lengths‡	Second strand pattern C3	C5	C7	C9	Start of β-zipper residue§
HECY-3D	7**	L x	L x	L x	L	6–6		L x	L		Phe-460
SOD	12		L x	A x	L	4–5		L x	L		Val-29
ACX	3	L x	L			3–4		L x	L		Val-18
IgG V	60	L†† x	L x	C		8–11		L‡‡ x	W		Leu-18
IgG C	24		L x	C x	L§§	5–7	L*** x	L x	W		Leu-143
γ-Crystallin†††	24		L x	L		4–4		L x	L		Ile-35
Prealbumin	4	L L	L x	L		9–9	L x	L x	L x	L	Leu-12
Azurin	7	L x	L x	L		14–14		L L	L		Phe-29
Plastocyanin	16		L x	L		8–8		L L	L		Ile-27

*Each β-zipper pattern is given as a "first strand pattern" and a "second strand pattern," separated by the minimum and maximum length of the loop between the β-strands. The one letter code used represents A, Gly, Ala, Val, Ile, Leu, Met, Phe, Tyr, or Trp; L, Val, Ile, Leu, Met, Phe, Tyr, or Trp; C, Cys; W, Trp; X, any residue.

†NS is the number of sequences considered per family.

‡Loop length is defined as the number of residues between β-strands as given by the DSSP program. The first number is the minimum length of the loop observed in the family of proteins; the latter number is the maximum loop length.

§Residue specification of the first residue of the β-zipper pattern as it occurs in the structures listed in Table I.

**Not all sequences are complete.

††Two exceptions occur (1 × Gly, 1 × Arg).

‡‡Three exceptions occur (2 × Ser, 1 × Arg).

§§Six exceptions occur (6 × Ala); in 5 cases this is "compensated" by an additional valine at position B4.

***One exception occurs (Ala).

†††The given pattern occurs in both the N- and the C-terminal barrel.

is that the time needed to form a folding nucleus increases rapidly with increasing distance between components of the nucleus.[26] Second, the loss in entropy by fixing two peptide segments is proportional to the logarithm of the number of residues in between the two segments.[30,31] Therefore the most likely nucleation structures are thought to consist of a single α-helix or two or three secondary structural elements connected by short loops. (e.g., α, α–α, β–β, β–α–β topologies).[26,32]

A Proposal for Folding Initiation of IgG-Like Greek Key β- Barrels

Based on the nucleus properties described above, our observation that the B- and C-strands are connected by a conserved short loop suggests that the B- and C-strands might well be involved in a nucleation mechanism common to all seven β-barrels. We propose that this nucleation site is similar to the B-loop-C unit as observed in the native protein. The B-loop-C unit has a β-arch conformation and so consists of two β-strands that are not hydrogen bonded to each other (see Fig. 4). The extended conformation of the nucleus strands is therefore not the result of hydrogen bond formation. It is, however, a conformation that gives an efficient packing of the hydrophobic side chains. To obtain a sufficiently stable nucleus these side chains should not be small.

We have named the β-arch conformation with a short loop and stabilized by a core of large hydrophobic side-chains a "β-zipper." This term is suggested by Figure 4A and D–G, where B-loop–C units are seen stacking their hydrophobic side chains in an alternating manner. For the other proteins studied, the native structure no longer contains a "perfect," closely packed, β-zipper. It is, however, not difficult to imagine that initially a β-zipper type packing occurred, which in a later stage of the folding process is rearranged into the conformation observed in the completely folded protein. The disruption of the β-zipper in SOD by a Gly at position B8 (see Fig. 4B) will be discussed later.

Alternative Folding Schemes for Greek Key β-Barrel Proteins

A number of other folding schemes for Greek key β-barrel proteins, based on a β-hairpin nucleus, have been suggested in the literature. A hypothesis in which folding was initiated by the formation of a single long ribbon of two antiparallel hydrogen bonded β-strands was first proposed by Richardson[33] and subsequently taken up and elaborated by several authors such as Ptitsyn[34,35] and Richardson.[36] McLachlan proposed a number of folding pathways based on the assumption that Greek key β-barrels have originated from a gene duplication followed by a gene fusion.[37] One of the folding proposals for the IgG variable domains started with the pairing of the B, C, E, and F-strands, forming a four-stranded nucleus. It is, however, not clear from the article if this pairing proceeds via the addition of β-strands to a β-zipper-like structure or via the addition of two β-hairpin structures (B–E and C–F).

To decide on a theoretical basis which kind of nu-

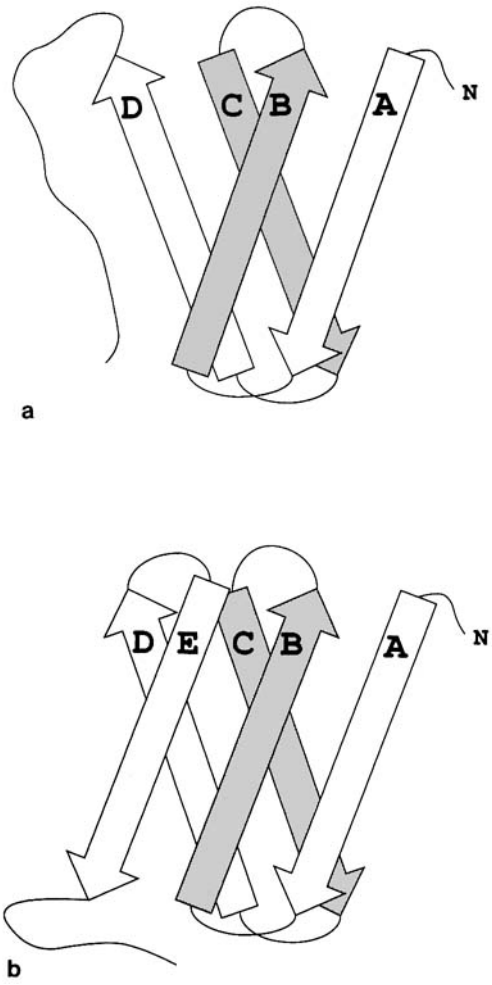


Fig. 7a and b.

cleus, β-zipper, or β-hairpin occurs in nature, one cannot look only at the relative stabilities of both nuclei conformations. After all, the only requirements for a good nucleation site are that, under folding conditions, it can be formed rapidly, is present for a reasonable percentage of time, and has a reasonable likelihood of stimulating successful protein folding. The presence of a more stable alternative conformation is thus irrelevant as long as that conformation does not trigger successful protein folding.

To obtain some insight in the stabilization of the β-zipper by hydrophobic interactions, we calculated the difference in exposed hydrophobic surface of positions B6, B8, B10, C5, and C7 in hecy-3d, when in the β-zipper conformation and when both strands were taken apart. This gave a difference of 242 Å^2. With current estimates of $\Delta G_{transfer}$ in the order of 25 to 60 cal mol^{-1} Å^{-2},[38] this would give a stabilization energy of at least 6 kcal/mol. This indicates that the β-zipper is likely to be formed in the still unfolded protein. In the following we will concentrate on a possible sequence of events occurring after the putative initial formation of the β-zipper.

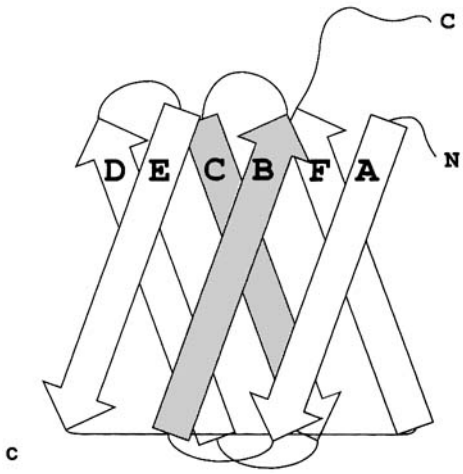


Fig. 7. Hypothetical folding pathway for Greek key β-barrel proteins, based on β-zipper nucleation. **(a)** Addition of A- and D-strands. **(b)** Addition of E-strand next to the B-strand. **(c)** Addition of F-strand next to the C-strand.

Extension of the β-Zipper to the Greek Key β-Barrel

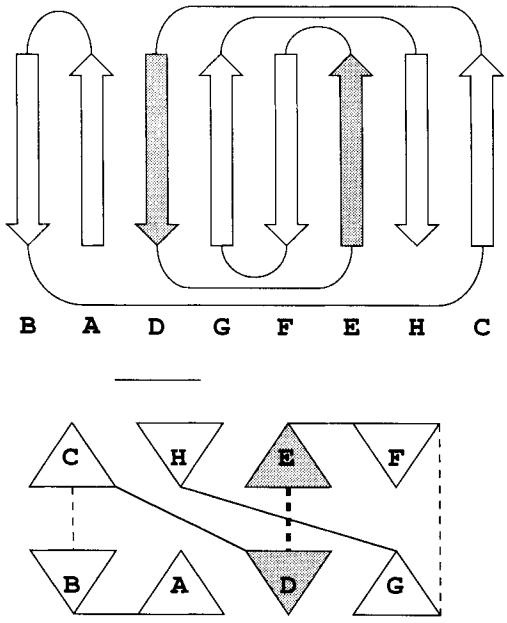
If we assume that the β-zipper is the nucleation site in all Greek key β-barrels studied in this paper, how does this common nucleus then lead to a unique native fold? To give an explanation for this, we would like to give a hypothetical folding pathway based on three assumptions:

1. Each folding step tries to optimize hydrophobic interactions.
2. Hydrogen bond donors and acceptors do not become buried without forming a hydrogen bond.
3. Residues that are close in sequence to the already present partial structure are preferably added first.

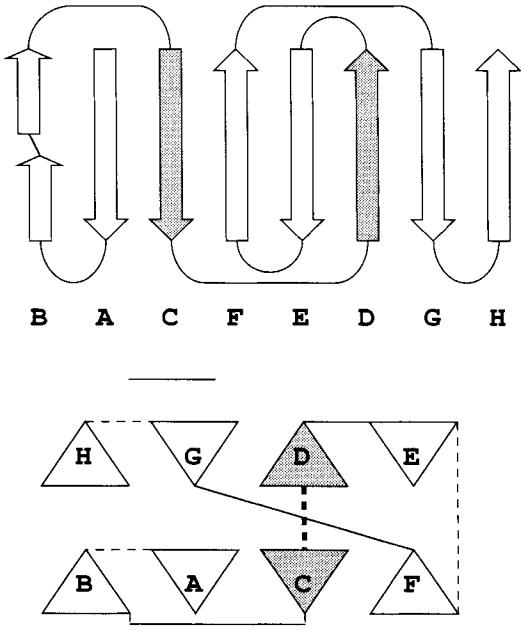
Since the β-zipper strands are already in an extended conformation, the most likely structures to be added to it are other β-strands, giving an efficient main chain hydrogen bonding pattern (rule 2). The first strands to be added are most likely strands A and D (rule 3). In principle these strands can bind in two positions, either to the left or to the right of strands B and C (see Fig. 6). Observed Greek key β-barrels show, however, a unique handedness.

The unique handedness can be explained by considering the β-strand twist. Due to this twist there is an angle between the strands in the β-zipper. This feature, which was already shown in an article by Chothia and Janin,[39] is represented schematically in Figure 5.

In Figure 6 the addition of the A-strand on both sides of the B-strand is depicted. The difference between both possibilities is that in the observed hand-

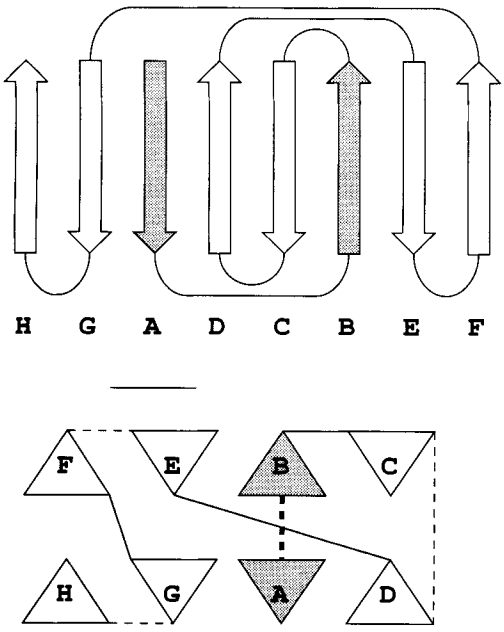


a GAMMA CRYSTALLIN BARREL



c AZURIN/PLASTOCYANIN BARREL

Fig. 8. Topology diagrams for **(a)** γ-crystallin (both barrels of γ-crystallin have this topology), **(b)** prealbumin, **(c)** azurin, and plastocyanin.

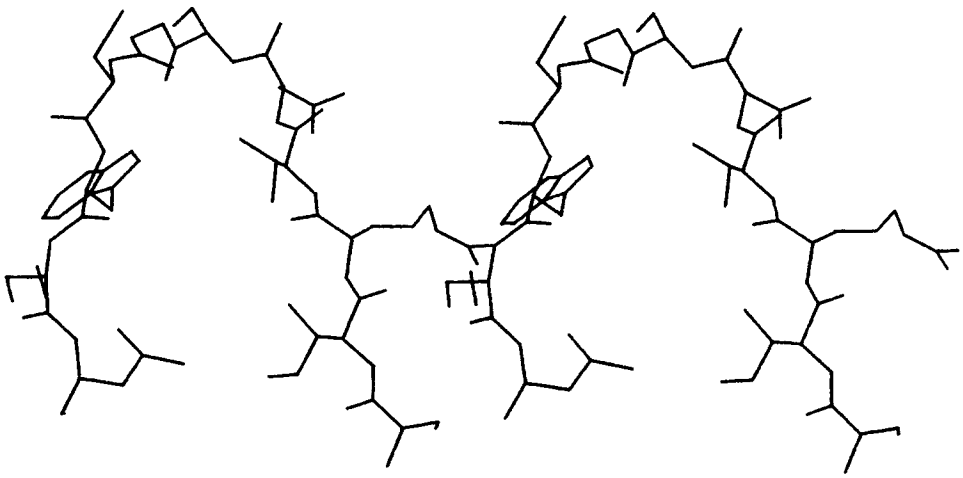


b PREALBUMIN BARREL

Fig. 8a and b.

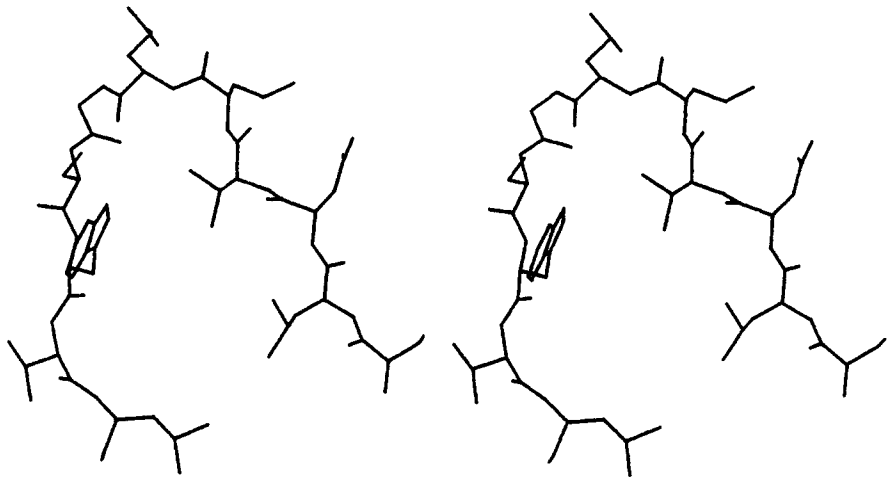
edness the A-strand is immediately stabilized by contacts with both the B- and C-strand, whereas in the inverted handedness stabilization by the C-strand occurs only in the final stage of adding A to B. The same reasoning applies to adding strand D to C. The observed positions are therefore kinetically more favorable. In the now four-stranded structure, two attractive positions for adding another β-strand are left open: in the upper sheet next to the B-strand and in the lower sheet next to the C-strand (see Figure 7A). Both positions offer extensive hydrophobic interactions. In the absence of another strand N-terminal to the A-strand, The E-strand will now be added next to the B-strand (rule 3).

One face of the β-zipper has now been buried from the solvent by strands D and E (see Fig. 7B). Residues occupying core positions in these latter strands can be hydrophilic, since the polar head groups can be pointing outward into solution. However, due to these polar groups, further extension of the β-barrel to this side of the β-zipper becomes blocked, since further extension would bury these polar groups. Also at the top of the barrel the hydrophobic core is screened from the solvent by the B–C and D–E loops. The bottom side and the position next to the C-strand are, however, still exposed. Both exposed surfaces can now be buried at the same time by addition of the E–F loop and the F-strand (see Fig. 7C). The



A

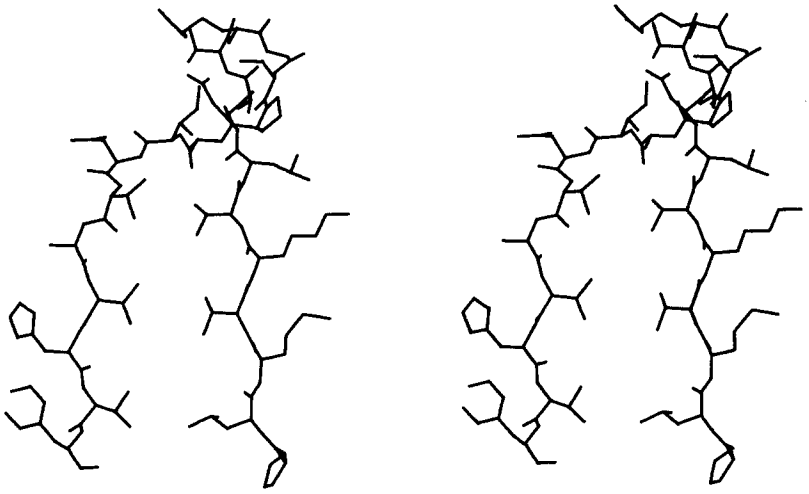
CRYST DOMAIN 1 CRYST DOMAIN 1



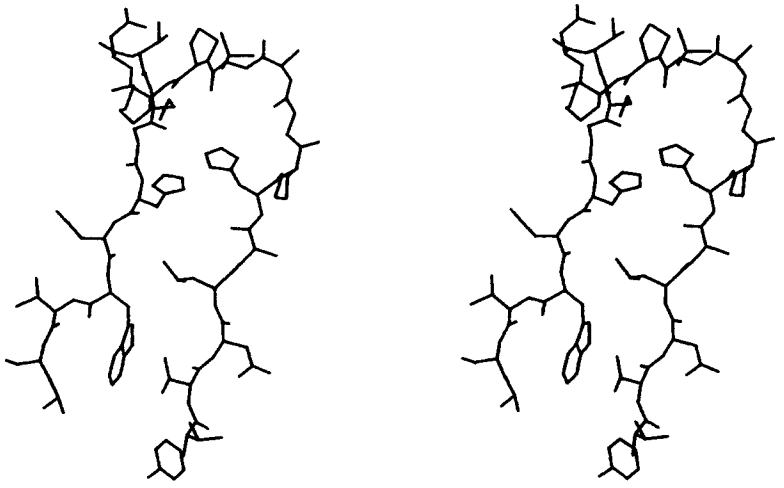
B

CRYST DOMAIN 2 CRYST DOMAIN 2

Fig. 9. The β-zipper structures as observed in **(A)** γ-crystallin, residues Ser-34–Leu-44 (N-terminal barrel), **(B)** γ-crystallin, residues Ser-123–Leu-133 (C-terminal barrel), **(C)** prealbumin, residues Pro-11–Phe-33, **(D)** azurin, residues Phe-29–Leu-50, **(E)** azurin alternative β-zipper, residues Ile-20–Asn-32, **(F)** plastocyanin, residues Ile-27–Asp-42, **(G)** plastocyanin alternative β-zipper, residues Glu-18–Lys-30. (The residue ranges refer to the structures listed in Table I.) Figure continues through page 296.

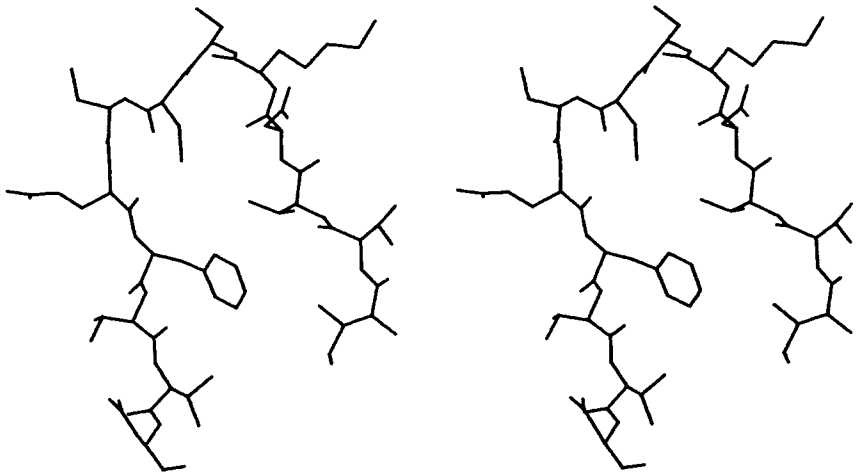


C PREALBUMIN PREALBUMIN



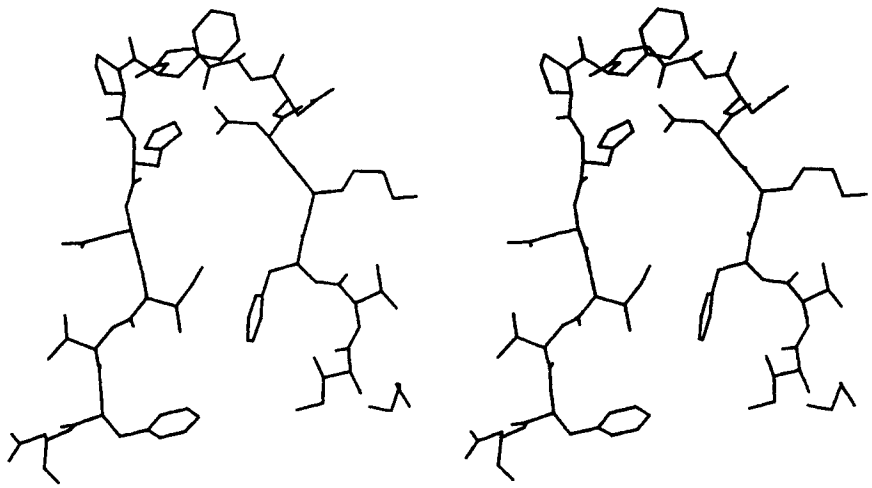
D AZURIN AZURIN

Fig. 9C and D. Legend appears on page 293.



E

AZURIN2 AZURIN2



F

PLASTOCYANIN PLASTOCYANIN

Fig. 9E and F. Legend appears on page 293.

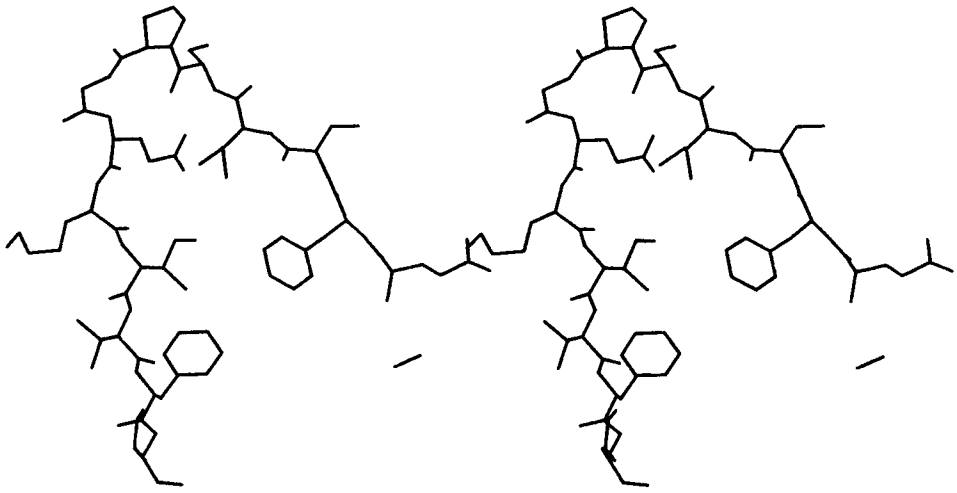


Fig. 9G. Legend appears on page 293.

six-stranded structure now arrived at is the basic topology for all Greek key β-barrels with an IgG fold.

Monte Carlo Protein Folding Simulations

Additional support for the protein folding pathway given above and the role of the β-zipper comes from recent Monte Carlo protein folding simulations.[40] The model used for the simulations consisted of 74 residues represented by 74 beads placed on a diamond lattice. In the "protein" sequence 6 stretches were defined, having a slight preference for the extended conformation and containing hydrophilic and hydrophobic residues in alternating positions. A few other parameters were present in the model, but no target potential for the native structure was included. This model folded reproducibly into the Greek key β-barrel topology. From 14 observed complete folding transitions, 9 were initiated by the B–C β-zipper. For the other 5 transitions the C–D β-hairpin was the nucleus. It was also observed that β-zipper formation was sometimes stimulated by the initial formation of an A–B β-hairpin, after which the A-strand was released again from the β-zipper. In the case of SOD, where a glycine (Gly B8) disrupts the close packing of the β-zipper (see Fig. 4B), initial formation of the A–B β-hairpin is probably essential. The gap created by the glycine is then nicely filled by an isoleucine (Ile-A7).

Other Greek Key β-Barrels

The β-zipper provides a plausible mechanism by which a Greek key β-barrel is formed by adding β-strands to both sides of the β-zipper. It is, however, not necessary that after formation of the BCDEF-structure, subsequent addition of more strands should always lead to the IgG topology. Therefore we investigated whether or not the β-zipper also occurs in Greek key β-barrel proteins with different topologies, like γ-crystallins, prealbumins, azurins, and plastocyanins. The first two families have topologies that are rather different from the IgG topology, whereas the azurins and the related plastocyanins are more similar to the IgG topology. Topology diagrams for these families are given in Figure 8A to C.

For these four additional protein families many aligned sequences were used to look for a conserved β-zipper motif (see Materials and Methods). For each family, a β-zipper motif was indeed found between two β-strands in the center of opposite β-sheets. In most cases it was also the only conserved motif or the most obvious one. Only for the azurins and the related plastocyanins a second β-zipper pattern with an even shorter loop was found (β-strands B, C, and the connecting loop, see Fig. 8). Interestingly, however, it appeared that although the N-terminal part of strand B is lying in the same β-sheet as strand C, its C-terminal part switches to the other sheet and indeed makes the hydrophobic interactions of a β-zipper structure. (see Fig. 9E and G). A list of observed β-zipper motifs for all families is given in Table IV. Stereo pictures of β-zipper structures in non IgG-like Greek key β-barrel proteins are given in Figure 9. For γ-crystallin the β-zippers of both domains are shown and for azurin

and plastocyanin also the alternative β-zipper is depicted.

CONCLUSION

Our study suggests that the similar Greek key β-barrel structure found in four unrelated protein families is due to a similar nucleation event. Observation of the β-zipper in another four Greek key β-barrel protein families with topologies that differ from the IgG topology indicates that this nucleation event could be a universal feature for Greek key β-barrel proteins.

Considering the very limited degree of conserved sequence properties, especially in the residues that do not belong to the nucleation structure, it seems that the formation of the β-barrel does not impose stringent sequence restrictions. This might be inherent to the nature of the β-zipper structure, which, with its four free rows of main chain hydrogen bond donors and acceptors surrounding a hydrophobic core, could be acting as a template to which other β-strands are then easily added. The rather low sequence requirements could also explain why the Greek key β-barrel topology occurs quite often in protein structures. The β-zipper is, however, not the only determinant for Greek key formation, since the β-zipper sequence pattern is also observed in proteins without a Greek key β-barrel.

We are aware of the fact that the proposed β-zipper mechanism is fully hypothetical and cannot be definitely proven by the kind of analysis described in this article. Several new experimental procedures, including NMR, stopped flow experiments, and genetic engineering have, however, appeared to be promising tools to gain more insight in protein folding and could be used to test the validity of the proposed β-zipper protein folding pathway.

ACKNOWLEDGMENTS

We like to thank Dr. Anne Volbeda for his work leading to the refined structure of *Panulirus interruptus* hemocyanin. This research was supported by the Dutch Foundation for Chemical Research (SON) with financial aid from the Netherlands Organisation for Scientific Research (NWO).

REFERENCES

1. Gaykema, W.P.J., Hol, W.G.J., Vereijken, J.M., Soeter, N.M., Bak, H.J., Beintema, J.J. 3.2 Å structure of the copper-containing, oxygen-carrying protein *Panulirus interruptus* haemocyanin. Nature (London) 309:23–29, 1984.
2. Gaykema, W.P.J., Volbeda, A., Hol, W.G.J. Structure determination of *Panulirus interruptus* haemocyanin at 3.2 Å resolution: Successful phase extension by sixfold density averaging. J. Mol. Biol. 187:255–275, 1986.
3. Volbeda, A., Hol, W.G.J. Crystal structure of hexameric haemocyanin from *Panulirus interruptus* refined at 3.2 Å resolution. J. Mol. Biol. 209:249–279, 1989.
4. Samy, T.S.A., Hahm, K., Modest, E.J., Lampman, G.W., Keutmann, H.T., Umezawa, H., Herlihy, W.C., Gibson, B.W., Carr, S.A., Biemann, K. Primary structure of macromomycin, an antitumor antibiotic protein. J. Biol. Chem. 258:183–191, 1983.
5. Richardson, J.S., Richardson, D.C., Thomas, K.A. Similarity of three-dimensional structure between the immunoglobulin domain and the copper,zinc superoxide dismutase subunit. J. Mol. Biol. 102:221–235, 1976.
6. Getzoff, E.D., Tainer, J.A., Stempien, M.M., Bell, G.I., Hallewell, R.A. Evolution of CuZn superoxide dismutase and the Greek key β-barrel structural motif. Proteins 5: 322–336, 1989.
7. Lesk, A.M., Chothia, C. Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. 160:325–342, 1982.
8. Taylor, W.R. Identification of protein sequence homology by concensus template alignment. J. Mol. Biol. 188:233–258, 1986.
9. Van Roey, P., Beerman, T.A. Crystal structure analysis of auromomycin apoprotein (macromomycin) shows importance of protein side chains to chromophore binding selectivity. Proc. Natl. Acad. Sci. U.S.A. 86:6587–6591, 1989.
10. Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S., Richardson, D.C. Determination and analysis of the 2 Å structure of copper, zinc superoxide dismutase. J. Mol. Biol. 160:181–217, 1982.
11. Pletnev, V.Z., Kuzin, A.P., Trakhanow, S.D., Kostetsky, P.V. Three-dimensional structure of actinoxanthin. IV. A 2.5 Å resolution. Biopolymers 21:287–300, 1982.
12. Marquart, M. Deisenhofer, J., Huber, R., Palm, W. Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 and 1.9 Å resolution. J. Mol. Biol. 141: 369–391, 1980.
13. Summers, L., Wistow, G., Narebor, M., Moss, D., Lindley, P., Slingsby, C., Blundell, T., Bartunik, H., Bartels, K. X-ray studies of the lens specific proteins. The crystallins. Pept. Protein Rev. 3:147–168, 1984.
14. Blake, C.C.F., Geisow, M.J., Oatley, S.J., Rerat, B., Rerat, C. Structure of prealbumin, secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 angstroms. J. Mol. Biol. 121:339–356, 1978.
15. Adman, E.T., Jensen, L.H. Structural features of azurin at 2.7 angstroms resolution. Isr. J. Chem. 21:8–12, 1981.
16. Guss, J.M., Freeman, H.C. Structure of oxidized poplar plastocyanin at 1.6 angstroms resolution. J. Mol. Biol. 169: 521–563, 1983.
17. Linzen, B., Soeter, N.M., Riggs, A.F., Schneider, H.-J., Schartau, W., Moore, M.D., Yokota, F., Behrens, P.Q., Nakashima, H., Tagaki, T., Nemoto, T., Vereijken, J.M., Bak, H.J., Beintema, J.J., Volbeda, A., Gaykema, W.P.J., Hol, W.G.J. The structure of arthropod hemocyanins. Science 229:519–524, 1985.
18. Neuteboom, B., Jekel, P.A., Hofstra, R.M.W., Beintema, J.J. Sulfhydryl groups and disulfide bridges in subunit c of *Panulirus interruptus* hemocyanin. Biochim. Biophys. Acta 998:126–130, 1989.
19. Atlas of protein sequence and structure. Dayhof, M.O. (ed). National Biomedical Research Foundation: 200–204, 1978.
20. Jones, T.A. Interactive computer graphics: FRODO. Methods Enzymol. 115:157–171, 1985.
21. Rao, S.T., Rossmann, M.G. Comparison of super-secondary structures in proteins. J. Mol. Biol. 76:241–256, 1973.
22. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
23. Chothia, C., Lesk, A.M. Evolution of proteins formed by β-sheets. I. Plastocyanin and azurin. J. Mol. Biol. 160:309–323, 1982.
24. Lim, W.A., Sauer, R.T. Alternative packing arrangements in the hydrophobic core of λ repressor. Nature (London) 339:31–36, 1989.
25. Lim, V.I. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J. Mol. Biol. 88:857–872, 1974.
26. Karplus, M., Weaver, D.L. Protein-folding dynamics. Nature (London) 260:404–406, 1976.
27. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. U.S.A. 70:697–701, 1973.
28. Levinthal, C. Molecular model-building by computer. Sci. Am. 214:42–52, 1966.

29. Levinthal, C. Are there pathways for protein folding? J Chim. Phys. 65:44–45, 1968.
30. Poland, D.C., Scheraga, H.A. Statistical mechanics of noncovalent bonds in polyamino acids. VIII. Covalent loops in proteins. Biopolymers 3:379–399, 1985.
31. Chan, H.S., Dill, K.A. Intrachain loops in polymers: Effect of excluded volume. J. Chem. Phys. 90:492–509, 1988.
32. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature (London) 261:552–557, 1976.
33. Richardson, J.S. β-Sheet topology and the relatedness of proteins. Nature (London) 268:495–500, 1977.
34. Ptitsyn, O.B., Finkelstein, A.V. Self-organization of proteins and the problem of their three-dimensional structure prediction. In: "Protein Folding." Jaenicke, R (ed). Amsterdam: Elsevier/North-Holland Biochemical Press, 1980: 101–115.
35. Ptitsyn, O.B. Protein folding: General physical model. FEBS Lett. 131:197–201, 1981.
36. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
37. McLachlan, A.D. Pseudo-symmetric structural elements and the folding of domains In: "Protein Folding." Jaenicke, R. (ed). Amsterdam: Elsevier/North-Holland Biomedical Press, 1980:79–99.
38. Sharp, K.A. The hydrophobic effect. Curr. Opinion Struct. Biol. 1:171–174, 1991.
39. Chothia, C., Janin, J. Relative orientations of close-packed β-pleated sheets in proteins. Proc. Natl. Acad. Sci. U.S.A. 78(7):4146–4150, 1981.
40. Skolnick, J., Kolinski, A. Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six-member, Greek key β-barrel proteins. J. Mol. Biol. 212:787–817, 1989.