Journal of Structural Biology **134,** 167–185 (2001)
doi:10.1006/jsbi.2001.4335, available online at http://www.idealibrary.com on IDEAL®

Fold Change in Evolution of Protein Structures

Nick V. Grishin

Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050

Received November 3, 2000, and in revised form February 15, 2001; published online May 23, 2001

Typically, protein spatial structures are more conserved in evolution than amino acid sequences. However, the recent explosion of sequence and structure information accompanied by the development of powerful computational methods led to the accumulation of examples of homologous proteins with globally distinct structures. Significant sequence conservation, local structural resemblance, and functional similarity strongly indicate evolutionary relationships between these proteins despite pronounced structural differences at the fold level. Several mechanisms such as insertions/deletions/substitutions, circular permutations, and rearrangements in β-sheet topologies account for the majority of detected structural irregularities. The existence of evolutionarily related proteins that possess different folds brings new challenges to the homology modeling techniques and the structure classification strategies and offers new opportunities for protein design in experimental studies.

Key Words: **circular permutation; insertion; deletion; molecular evolution; protein structure classification; conformational change; homology modeling.**

INTRODUCTION

After determination of the first handful of three-dimensional protein structures it became clear that spatial structure is more conserved than protein sequence (Chothia and Lesk, 1986; Doolittle, 1981; Flores *et al.*, 1993; Grishin, 1997; Holm and Sander, 1996, 1997c; Hubbard and Blundell, 1987). Numerous examples of very close structural resemblance in the absence of detectable sequence similarity have been catalogued (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995; Orengo *et al.*, 1997; Pearl *et al.*, 2000; Russell *et al.*, 1997, 1998). Many of them illustrate divergent evolution. Typically, structural features remain preserved long after sequence signal is lost to mutations, insertions, and deletions. Are there exceptions to this rule? Could one find proteins that in all likelihood share a common ancestor, but possess significant structural differences?

The most general structural similarity is described in terms of protein folds. According to SCOP (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995), proteins are classified within the same fold if they have the same major secondary structural elements in the same mutual orientation and with the same connectivity (topological connections). Application of this definition to real proteins could cause confusion due to subjectivity in deciding which elements are major. Nevertheless, through careful comparative analysis of proteins, one could identify structural units that occur recurrently and represent prototypes for the main folds, such as $(\beta\alpha)_8$-barrel, OB-fold, immunoglobulin, and Rossmann fold (Holm and Sander, 1996, 1997c; Lo Conte *et al.*, 2000; Murzin *et al.*, 1995). In many cases, however, fold definition remains an empirical approximate "art" and even the experts diagree on fold assignments for many proteins (Hadley and Jones, 1999; Holm and Sander, 1996, 1997a; Lo Conte *et al.*, 2000; Murzin *et al.*, 1995; Orengo *et al.*, 1997; Pearl *et al.*, 2000). The criteria used are often rather loose and are frequently based not only on structural data, but also on evolutionary and functional considerations (Murzin *et al.*, 1995; Orengo *et al.*, 1997). For the purpose of the current review, I try to apply the fold definition mentioned above, which is based exclusively on structural arguments.

Proteins of the same fold do not necessarily share a common ancestor. Major structural similarity could arise independently due to the limited number of acceptable spatial arrangements of secondary structural elements (Holm and Sander, 1997b; Ptitsyn and Finkelstein, 1981; Russell, 1998, Russell *et al.*, 1997, 1998). The purpose of this review is to demonstrate that the opposite scenario also takes place. Namely, there exist evolutionarily related proteins that contain major structural differences and thus these proteins could be attributed to different folds. Moreover, with the explosion of avail-

1047-8477/01 $35.00

able structural information during the past several years, such examples have been accumulating to allow for their classification into several distinct types. In late 1970s to early 1980s researchers described recurrent motifs of protein structures (Chothia, 1984; Chothia *et al.*, 1977; Levitt and Chothia, 1976; Richardson, 1977, 1981). Nowadays, we are at a point to discuss recurrent ways for proteins to transform from one structural motif to another in evolution. These ways, or mechanisms, are of exceptional interest since they have a profound impact on our understanding of the protein world. Practically, their existence indicates difficulties for homology modeling techniques that rely heavily on the assumption "similar sequences—similar structures." Indeed, it is common to assume that if similarity between two sequences is detected using profile-based methods, then these two sequences will have rather similar spatial structures. Additionally, the possibility of significant structural changes in evolution brings inconsistencies between sequence-based and structure-based protein classification schemes. The most fundamental questions, however, concern the evolution of protein structure, its relation to evolution of sequence and function, and mechanisms by which protein folds can change. These mechanisms remain largely unexplored both experimentally and theoretically.

HOMOLOGY AND SIMILARITY

Descent from a common ancestor, i.e., homology, can be hypothesized on the basis of similar properties detected in biological objects. For protein molecules, similarities could be reflected in sequence, structure, or function (Murzin, 1998; Thornton *et al.*, 1999). Comprehensive analysis of all these properties offers the best way to support homology convincingly. However, what are our options when the structural argument breaks down? How can we ensure that we are dealing with homology, when there are significant structural differences between proteins?

It was argued and largely accepted that statistically significant similarity detected from the sequence alone (without consideration of spatial structure) reflects descent from a common ancestor (Aravind and Koonin, 1999b; Doolittle, 1994; Murzin, 1998). There are at least three main reasons for this. First, sequence analysis methods, as powerful as they are, find only part of the homologs detectable by the combined structure–sequence–function approach (Aravind and Koonin, 1999b; Brenner *et al.*, 1998; Holm, 1998; Lo Conte *et al.*, 2000). The most divergent homologs are frequently missed in sequence similarity searches. Consequently, the homologs found by sequence-based techniques tend to be the closest ones in their sequences to a query protein. Thus if a sequence is detected in such searches, it is more likely to be from a pool of rather close homologs. Second, the space of possible sequences by far exceeds the space of allowed structures. Therefore it is unlikely that nature can independently find similar sequences that fold into a given structure (Doolittle, 1994). Third, programs that are routinely used in sequence similarity searches, such as PSI-BLAST (Altschul and Koonin, 1998; Altschul *et al.*, 1997), are based on amino acid similarity matrices. These matrices are either derived under evolutionary models (Dayhoff *et al.*, 1978) or computed from aligned homologous sequences (Henikoff and Henikoff, 1992) and thus are intended to find homologs.

The significant sequence similarity is almost always reflected in local structural resemblance in the regions of conserved sequence motifs. Although the structures might be globally different, local conformational similarities in the conserved motifs would hint at an evolutionary relationship. Additionally, weak sequence and structural similarities can be strengthened by a functional connection. Placement of a binding site, cofactor- or substrate-binding mode, and conserved catalytic residues could all support the hypothesized homology (Murzin, 1998).

More relaxed criteria for homology may be applied to multidomain proteins. Due to the evolutionary tendency to conserve domain architecture, the co-occurrence of domains increases the probability of homology between proteins. Typically, multidomain proteins have a major domain that is the principal functional unit of a molecule. This domain tends to be more conserved and usually shares significant sequence similarity between homologs. Other, smaller domains are less conserved. Sequence similarity between these smaller domains may be much weaker and might not be detectable by sequence search tools. However, if these smaller domains in the two proteins with homologous larger domains share some structural similarity, share residual sequence similarity, or interact with the larger domains in a similar manner, it is likely that the smaller domains are homologs.

Proteins discussed below to exemplify the fold change in evolution are selected carefully to ensure that homology is indeed the most likely scenario. Most of these proteins display statistically supported sequence similarity that can be detected by programs such as PSI-BLAST (Altschul and Koonin, 1998; Altschul *et al.*, 1997) or HMMer (Bateman *et al.*, 2000; Eddy, 1998). Four mechanisms that are believed to be the most common ways by which proteins can change their folds are discussed. Not all

known examples, but the most illustrative examples of proteins are provided to support each case.

ADDITION/DELETION/SUBSTITUTION OF STRUCTURAL ELEMENTS

Insertions and deletions (indels) together with single amino acid substitutions are the most common events in protein evolution. Indels are about an order of magnitude less frequent than residue substitutions (Benner *et al.*, 1993; Pascarella and Argos, 1992). One might look at indels as largely neutral or deleterious mishaps or as vehicles of progress: the right indel might relax structural tension accumulated in the course of amino acid substitutions. While the true role of indels in protein evolution remains to be investigated, it is quite clear that they offer a way that can potentially lead to significant and even drastic structural changes.

Luciferase

The most dramatic single-event change is revealed in a structural comparison of bacterial luciferase (Fisher *et al.*, 1995) (Fig. 1a) and a nonfluorescent flavoprotein (NFP)[1] encoded by the luxF gene of Photobacterium (Moore *et al.*, 1993) (Fig. 1b). Luciferase folds into a complete $(\beta\alpha)_8$-barrel. NFP contains an approximately 90-residue deletion that chops away two $\beta\alpha$ units and an α-helix (shown in red in Fig. 1a). To connect the remaining parts of the barrel and to complete the hydrophobic core in NFP, a single β-strand in an anti-parallel orientation (shown in red in Fig. 1b) occupies the place of the luciferase $\alpha\beta\alpha\beta\alpha$ unit (Holm and Sander, 1997c; Moore and James, 1994). The sequences of luciferase β subunit and NFP are 30% identical, the structures of the common regions are very similar, both proteins belong to the lux bacterial operon, and homology between them was suggested before the structure of luciferase was solved (Moore and James, 1994; Moore *et al.*, 1993). Indeed, the structure of NFP was determined first (Moore *et al.*, 1993), and the luciferase complete $(\beta\alpha)_8$-barrel has been modeled by homology using this structural information (Moore and James, 1994). Subsequent structure determination confirmed the prediction (Fisher *et al.*, 1995). While there is no doubt of homology between luciferase and NFP, the fold assignment of NFP remains a matter of definition and personal taste. As a common structural core between them, luciferase and NFP share only five $\beta\alpha$ units and a β-strand (b, Figs. 1a and 1b). Therefore, if the TIM-like $(\beta\alpha)_8$-barrel (Banner *et al.*, 1975) is defined as a fold, then NFP does not contain the same secondary structures with the same topology and thus should be classified as a separate, unique fold. In any phyletic classification, however, luciferase and NFP must be grouped as close relatives (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995). Deviations from the classical $(\beta\alpha)_8$-barrel topology are found in several other protein families, such as quinolinic acid phosphoribosyltransferases (Eads *et al.*, 1997), enolases (Lebioda *et al.*, 1989), and $(\beta\alpha)_7$ cellulases (Spezio *et al.*, 1993). Structural irregularities in these proteins were likely caused by insertion/deletion events.

Substitutions of Secondary Structures

The luciferase/NFP example illustrates how proteins can successfully accommodate significant structural changes in their cores after drastic deletions. More typically, however, indels are short, about one to five residues, and large ones occur at the periphery of the structure. In such cases, the core of the protein and thus its fold remain unchanged. Another common theme in protein evolution is a substitution of one secondary structural element by another. Such substitution may not reflect the actual substitution event at the gene level, where one piece of a sequence replaces another. Most likely many "substitutions" are caused by indels. For example, if a deletion occurs inside a helical segment that connects two fixed points in a protein structure, the resulting shorter segment would not be able to adopt a helical conformation and stretch between these fixed points. Such a deletion would be eliminated or would result in an apparent substitution of a helix by a loop or a strand. Essentially any two structural elements that fit between the two fixed points in a structure and complement the hydrophobic core could potentially substitute for each other. The most common substitutions of this kind are shown in Fig. 1c. The possibility of substitution of a β-strand by an α-helix has been experimentally demonstrated (Cordes *et al.*, 1999).

Lactate Dehydrogenase-NADH Peroxidase

The conversion (α-helix $\leftrightarrow$ 3-stranded β-meander) ($1 \leftrightarrow 4$ in Fig. 1c) is of particular interest since it seems to be rather common in doubly wound (Rossmann) fold proteins and affects one of the key elements in their structure. In a doubly wound $\alpha\beta\alpha$ sandwich, the protein chain starts in the middle of the β-sheet (β-strand a, Fig. 2b), travels outward, and then returns to the middle (β-strand d, Fig. 2b) via a connector. Classical Rossmann fold proteins,

[1] Abbreviations used: PDB, protein data bank; RMSD, root mean square deviation; KH, K homology; hnRNP, heterogeneous nuclear ribonucleoprotein; LDH, lactate dehydrogenase; NFP, nonfluorescent flavoprotein; HTH, helix-turn-helix; ODC, eukaryotic ornithine decarboxylase.

such as lactate dehydrogenase (LDH), contain only $\beta\alpha$ units and the connector is an α-helix (Adams *et al.*, 1970) (C, shown in red in Fig. 2b). PSI-BLAST searches initiated with lactate dehydrogenase sequences (Rossmann fold in a narrow sense and in SCOP) readily detect proteins of a different SCOP fold (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995), namely, the FAD/NAD(P)-binding domain. The sequence similarity is pronounced, with up to 26% identity between some members (Figs. 2a and 2b). For example, in NADH peroxidase from Enterococcus faecalis (Stehle *et al.*, 1991) and lactate dehydrogenase from Bacillus stearothermophilus (Wigley *et al.*, 1992), central β-strands a and d that form the core of the structure are composed of identical amino acids (five identical residues in a and 6 in d, Figs. 2a and 2b). The nucleotide-binding modes in these proteins are very similar and this similarity is reflected in the sequences: phosphate-binding glycine-rich loop GXGXXG, the signature of many Rossmann fold proteins, follows the β-strand a. Therefore it is likely that LDH and NADH peroxidase are homologous. However, the most significant structural difference between them, the apparent substitution of a connector helix C in LDH (shown in red in Fig. 2b) with a 3-stranded β-meander c′c″c‴ in NADH peroxidase (shown in red in Fig. 2a), results in a different, $\beta\beta\alpha$ (versus $\alpha\beta\alpha$) architecture and thus different folds for these homologs. Additional differences include the swap of the β-strand e and α-helices E and D between the two domains of NADH peroxidase (Fig. 2a).

Rossmann Fold-like Domains of ATP-Grasp Proteins

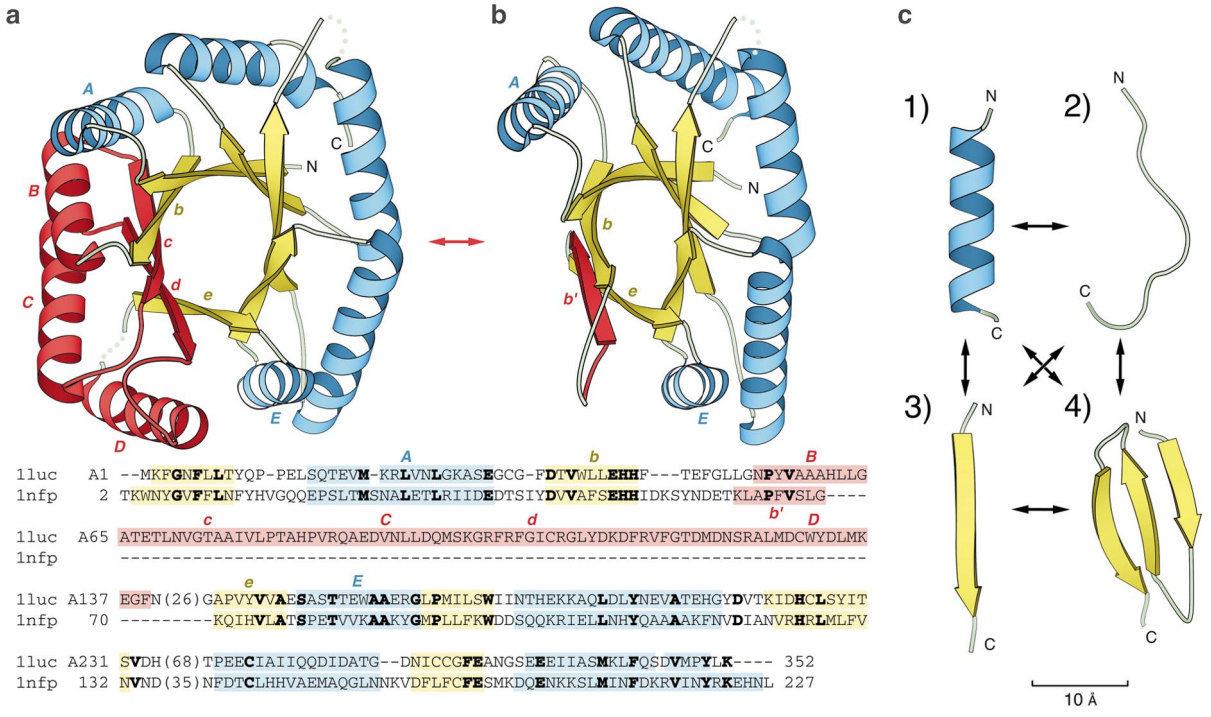
The complex pattern of indels and substitutions is revealed among proteins of the glutathione synthetase (ATP-grasp) superfamily. Most ATP-grasp proteins are readily detected by sequence similarity in the ATP-binding site (Galperin and Koonin, 1997). In many members of this superfamily, a substrate-binding domain is present N-terminal of the ATP-binding domains. This N-terminal domain contains a Rossmann-like core and with all likelihood is homologous between ATP-grasp proteins. However, this domain displays pronounced structural diversity in the peripheral region. Four representatives, Dala-Dala ligase (Fan *et al.*, 1994), biotin carboxylase (Waldrop *et al.*, 1994), synapsin (Esser *et al.*, 1998), and glutathione synthase (Yamaguchi *et al.*, 1993), are chosen here to illustrate the structural evolution of the domain (Figs. 3b, 3c, 3d, and 3e). Biotin carboxylase differs from Dala-Dala ligase by an additional $\beta\alpha$ unit at the edge of the β-sheet (Figs. 3b and 3c). The indel of a $\beta\alpha$ unit is typical for Rossmann proteins. Comparison of Dala-Dala ligase and synapsin reveals the (helix ↔ 3-meander) substitution (Figs. 3b and 3d), which is similar to that found in LDH and NADH peroxidase (Figs. 2a and 2b). The difference between biotin carboxylase and glutathione synthase can be interpreted as a (helix ↔ strand) substitution, where β-strand b‴ of glutathione synthase is topologically similar to α-helix B in biotin carboxylase (Figs. 3c and 3e) and is incorporated in the central β-sheet. Evolutionarily, however, it is more likely that glutathione synthase originated from the synapsin-like ancestor in which the β-hairpin b″c″ was extended to form strands b‴c (Figs. 3d and 3e). The carboxypeptidase A structure (Rees *et al.*, 1983), which might be related to the Dala-Dala ligase N-terminal domain, but is elaborated with many insertions, is also shown (Fig. 3a).

A Path from All-β to All-α Proteins

Analysis of protein structures shows clearly that indels/substitutions of a single structural element (α-helix, β-strand, loop, β-hairpin, $\beta\alpha$-init, 3-β-meander, etc.) are common events in molecular evolution. As a single event, such an indel/substitution may not affect a protein fold significantly, but gradual consecutive events could transform the structure beyond recognition. The effect of a series of indel/

FIG. 1. Insertions, deletions, and substitutions in the evolution of protein structures. (a) Bacterial luciferase (1luc); (b) nonfluorescent flavoprotein *luxF* (1nfp); (c) possible substitutions of secondary structural elements: (1) α-helix, (2) loop, (3) β-strand, (4) 3-stranded β-meander. References for each structure are given in the text. Ribbon diagrams were drawn by Bobscript (Esnouf, 1997), a modified version of Molscript (Kraulis, 1991). The structures were superimposed and then separated for clarity. N- and C-termini are labeled. The spatially equivalent structural elements are colored correspondingly in each group of structures. Insertions/deletions are shown in green or purple. Red is used to emphasize the structural change. Dotted lines symbolize long insertions or disordered regions. α-Helices and β-strands are labeled in upper- and lowercase boldface italic letters, respectively. Letter color matches the color of the secondary structure element. Red, orange, black, and blue arrows between structures symbolize clear evolutionary connection, possible evolutionary connection, absence of homology, and structural change within the same protein molecule, respectively. Structure-based sequence alignments are shown. The panel label, PDB entry code, and starting and ending residue numbers are given for each protein. Color shading and labels of secondary structure elements correspond to those in structure diagrams. Invariant amino acids are shown in boldface letters. PDB codes and protein names are shown below the alignments.

FIG. 2. α-Helix ↔ 3-stranded β-meander substitution in Rossmann fold-like proteins. (a) NADH peroxidase (1npx); (b) lactate dehydrogenase (1ldn). See legend to Fig. 1 for details.



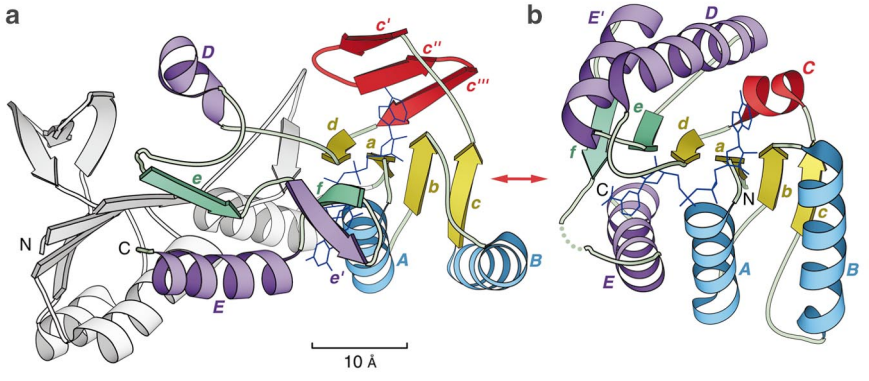

```
(a) 1luc   A1 --MKFGNFLLTYQP-PELSQTEVM-KRLVNLGKASEGCG-FDTVWLLEHHF---TEFGLLGNPYVAAAHLLG
(b) 1nfp    2 TKWNYGVFFLNFYHVGQQEPSLTMSNALETLRIIDEDTSIYDVVAFSEHHIDKSYNDETKLAPFVSLG----

(a) 1luc  A65 ATETLNVGTAAIVLPTAHPVRQAEDVNLLDQMSKGRFRFGICRGLYDKDFRVFGTDMDNSRALMDCWYDLMK
(b) 1nfp      ------------------------------------------------------------------------

(a) 1luc A137 EGFN(26)GAPVYVVAESASTTEWAAERGLPMILSWIINTHEKKAQLDLYNEVATEHGYDVTKIDHCLSYIT
(b) 1nfp   70 ---------KQIHVLATSPETVVKAAKYGMPLLFKWDDSQQKRIELLNHYQAAAAKFNVDIANVRHRLMLFV

(a) 1luc A231 SVDH(68)TPEECIAIIQQDIDATG--DNICCGFEANGSEEEIIASMKLFQSDVMPYLK---- 352
(b) 1nfp  132 NVND(35)NFDTCLHHVAEMAQGLNNKVDFLFCFESMKDQENKKSLMINFDKRVINYRKEHNL 227
```

1 1luc - Bacterial Luciferase 1nfp - Nonfluorescent Flavoprotein




```
(a) 1npx  149 VNNVVVIGSGYIGIEAAEAFAKAG--KKVTVIDILDRPLGVYLDKEFTDVLTEEMEA--------NNIT
(b) 1ldn  A20 GARVVVIGAGFVGASYVFALMNQGIADEIVLIDA---------NESKAIGDAMDFNHGKVFAPKPVDIW

(a) 1npx  208 IATGETVERYEGDGRVQKVVTDKNAYDADLVVVAVGVRP-----------NTAWLKGTLELHPNGLIK
(b) 1ldn  A80 HGDYD-----------------DCRDADLVVICAGANQKPGETRLDLVDKNIAIFRSIVESK------

(a) 1npx  265 TDEYMRTSEPDVFAVGD--ATLIKYN-------PADTEVNIA----LATNARKQGRFAVKNLEE  315
(b) 1ldn A124 ---VMASGFQGLFLVATNPVDILTYATWKFSGLPHERVIGSG(85)YYGIAMGLARVTRAILHN A265
```

2 1npx - NADH Peroxidase 1ldn Lactate Dehydrogenase

substitutions, one at each step, is illustrated by a path from an all-β to an all-α protein (Figs. 4a–4h). Unfortunately, in the current set of available protein structures it is not possible to find a convincing example of such a path composed entirely of evolutionarily related proteins. Thus, some steps in the presented path (black arrows) do not necessarily reflect connections between homologs. However, each subsequent protein is different from the previous protein by a single imaginary indel or substitution. A path similar to that shown in Figs. 4a–4h might have occurred in nature and will hopefully reveal itself when more protein structures become available.

The first two structures in this series, C-terminal domains of two amylases, namely, bacterial α-amylase (Machius *et al.*, 1995) and G4 amylase from Pseudomonas stutzeri (Morishita *et al.*, 1997), are almost certainly homologous, and yet CATH (Orengo *et al.*, 1997; Pearl *et al.*, 2000) places them not only in two different topological groups, and not even in two different architectures, but in two different classes: mainly Beta and Alpha Beta for α and G4 amylase, respectively. The major domain of both amylases is a $(\beta\alpha)_8$-barrel that is characterized by strong sequence similarity (28%). In both amylases, the smaller C-terminal domain interacts with the major $(\beta\alpha)_8$-barrel domain using the face of the conserved β-sheet abcf (Figs. 4a and 4b). The possible deletion of a β-hairpin e′e″ (shown in red in Fig. 4a) in α-amylase strips β-strand d off its H-bonding partner. The segment in G4 amylase structurally equivalent to d is folded as an α-helix D (Fig. 4b). The fold of the G4 amylase C-terminal domain is more similar to the N-terminal domain of Ser/Thr-Tyr protein kinases (Owen *et al.*, 1995) than to the original α-amylase β-barrel; the only difference is a longer helix D in kinases (Fig. 4c). Indeed, CATH places this domain in the same topological group with protein kinases. C-terminal domains of amylases exemplify a major structural change in evolution by one-step deletion: not only is the fold of the domain changed, but its class is also switched from all-β to $\alpha + \beta$.

The substitution of the β-strand a in the protein kinase N-terminal domain-like fold (Fig. 4c) by an α-helix A leads to the fold of the sonic hedgehog N-terminal signaling domain (Hall *et al.*, 1995) (Fig. 4d). Insertion of a helix C between the β-strand c and α-helix D creates the topology of a winged helix-turn-helix (HTH) domain of the catabolic gene activator protein (Schultz *et al.*, 1991) (Fig. 4e). There is no evidence that these transitions are between homologs. However, the examples in Figs. 4e–4h are nucleic acid-binding domains (Feng *et al.*, 1994; Schultz *et al.*, 1991; Wilson *et al.*, 1992; Xing *et al.*, 1997) that bind to the major groove through an α-helix D. These proteins contain an HTH motif detectable by sequence similarity in most of them (Aravind and Koonin, 1999a) with the possible exception of ribosomal protein L11 (Xing *et al.*, 1997) (Fig. 4g). Thus there is sequence and functional evidence for homology. The transitions from Figs. 4e to 4h are through successive deletions of β-strands: b, f, and e. Without the β-strand e, β-strand c does not have a partner to share H-bonds with, and the last structure, the HIN recombinase homeodomain (Feng *et al.*, 1994) (Fig. 4h), contains only α-helices.

Again, the path shown in Figs. 4a–4h does not necessarily reflect the actual evolutionary events but rather illustrates some principles that are possible in protein evolution and could be implemented in protein design. Recently, it was experimentally demonstrated that by replacing 50% of amino acids in a single step, it is possible to go from all-α to all-β protein or vice versa (Dalal *et al.*, 1997a, b; Jones *et al.*, 1996; Yuan and Clarke, 1998). However, these experiments do not really model the evolution that occurs through gradual, step-by-step changes, with all intermediates being fully foldable proteins (Blanco *et al.*, 1999). To create such an evolutionarily relevant path from all-α to all-β domains would be the next challenge for protein designers.

CIRCULAR PERMUTATION

Amino and carboxyl termini of protein domain structures are frequently placed in close proximity (Goldenberg, 1989; Thornton and Sibanda, 1983). Such an arrangement favors circular permutations, which are changes in protein connectivity that can be visualized through ligation of the termini and cleavage at another site. The first report of naturally occurring circular permutation describes precisely that process, namely, cleavage and ligation of the two fragments in concanavalin A as a result of post-translational modification (Bowles *et al.*, 1986; Carrington *et al.*, 1985; Cunningham *et al.*, 1979). However, most circular permutations occur at the gene level and can be detected by the existence of two homologs with different connectivities: the connectivity of one can be transformed to the connectivity of another by an imaginary circular permutation. A large number of possible circular permutations at the gene level have been reported and reviewed (Goldenberg, 1989; Lindqvist and Schneider, 1997; Pan and Uhlenbeck, 1993; Russell and Ponting, 1998). Although circular permutations do not change the spatial arrangement of secondary structural elements and side-chain packing and thus are easily tolerated in experimental studies (Ay *et al.*, 1998; Graf and Schachman, 1996; Luger *et al.*, 1989;

Otzen and Fersht, 1998; Pan and Uhlenbeck, 1993; Viguera *et al.*, 1995), they alter connectivity between secondary structures and thus, according to the classical fold definition, result in a protein fold change.

C2 Domains

One of the first well-documented examples of circular permutation at the gene level concerns C2 domains. The C2 domain is a Ca^{2+}-binding module present mainly in proteins participating in signal transduction. The crystal structure of the phospholipase C_δ C2 domain (Essen *et al.*, 1996) revealed topological differences compared to the previously determined structure of the synaptotagmin I C2 domain (Sutton *et al.*, 1995) (Figs. 5a and 5b). The first β-strand in synaptotagmin C2 (a, shown in red in Figs. 5a and 5b) corresponds to the last β-strand in phospholipase C2 and the topologies of the two domains are related by a circular permutation. Strong sequence similarity shared among the members of the C2 family (Nalefski and Falke, 1996) is reflected in close structural similarity (Pappa *et al.*, 1998) (Figs. 5a and 5b) and indicates homology. Sequence analysis of the C2 domain family revealed the presence of the two subfamilies typified by synaptotagmin and phospholipase C2 domains. However, only a single β-strand, which covers about 12% of the protein chain length, is involved in the circular permutation of the C2 domain. Are there examples where a more significant portion of the protein is permuted?

Saposins

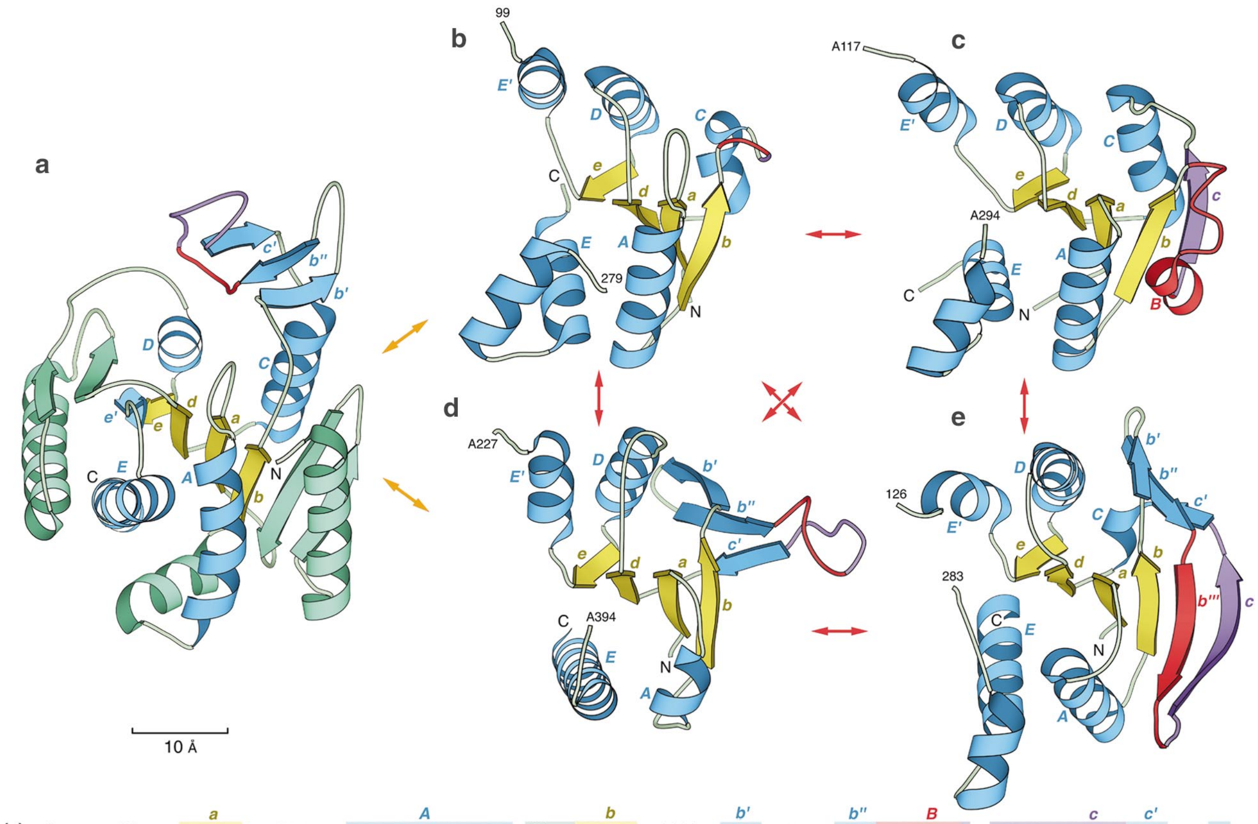
The most striking case of circular permutations is probably made by saposin-like domains (Ponting and Russell, 1995). These domains interact with lipid membranes and participate in a variety of physiological processes (Egas *et al.*, 2000; Tatti *et al.*, 1999). The structure of NK-lysin is composed of five α-helices arranged in a "folded leaf" architecture (Liepinsh *et al.*, 1997) (Fig. 5c). The recently solved structure of an aspartic proteinase prophytepsin (Kervinen *et al.*, 1999) confirmed the circular permutation in a "swaposin" domain (Fig. 5d). About half of the domain chain (α-helices A and B, shown in red in Figs. 5c and 5d) participates in this permutation event. Permutation in saposins has been detected by the sequence analysis before either structure became available, which leaves no doubt about homology between NK-lysin and swaposin. The aspect that is of particular interest is that the saposin–swaposin example provides clues to the actual mechanisms of this permutation. The swaposin sequence contains a 30-residue insert at the region corresponding to the N- and C-termini of NK-lysin. This insert is disordered in prophytepsin crystals. The "hybrid" origin of the swaposin domain has been suggested by Ponting and Russell (1995). Saposin domains occur as tandem repeats with linker regions in between. The N-terminal part of the swaposin molecule is likely to correspond to the C-terminal part of one repeat, the unstructured insert corresponds to the "intersaposin" linker, and the C-terminal part of swaposin is probably derived from the N-terminal segment of the next swaposin repeat.

Methyltransferases

The next example suggests that the mechanism of permutations proposed for saposins may be widespread. A large superfamily of *S*-adenosylmethionine-dependent methyltransferases is characterized by two highly conserved motifs that map to the C-termini of the β-strands a and d (Figs. 5e and 5f). It has been noticed that methyltransferase subfamilies differ in the sequence order of these motifs (Malone *et al.*, 1995; Wilson, 1992), suggesting a circular permutation. This permutation incorporates a significant part of the molecule. Two $\alpha\beta$ units and an α-helix (EaAbB, shown in red in Figs. 5e and 5f) appear at the N-terminus of the adenine-specific DNA methyltransferase (Schluckebier *et al.*, 1997) and are C-terminal in *Pvu*II DNA methyltransferase (Gong *et al.*, 1997). The evolutionary model for the origin of this permutation assumes the presence of a tandem protein that originated by a gene duplication and in-frame fusion. Subsequent introduction of the new start codon in the middle of the first gene copy and a stop codon at the equivalent position in the second gene copy can generate circularly permuted variants (Jeltsch, 1999). The proposed intermediate, a tandem methyltransferase, has been observed in nature. The FokI methyltransferase

FIG. 3. Rossmann fold-like domains of ATP-grasp proteins and zinc-carboxypeptidase. (a) Carboxypeptidase A (2ctc); (b) Dala-Dala ligase (2dln); (c) biotin carboxylase (1bnc); (d) synapsin (1auv); (e) glutathione synthase (1gsa). See legend to Fig. 1 for details.

FIG. 4. A path from all-β to all-α proteins. (a) *Bacillus licheniformis* α-amylases, C-terminal domain (1bpl); (b) *Pseudomonas stutzeri* G4-amylase C-terminal domain (2amg); (c) γ-subunit of glycogen phosphorylase kinase N-terminal domain (1phk); (d) sonic hedgehog N-terminal signaling domain (1vhh); (e) catabolite gene activator protein (CAP), C-terminal domain (1cgp); (f) biotin repressor N-terminal domain (1bia); (g) ribosomal protein L11 C-terminal domain (1fow); (h) HIN recombinase DNA-binding domain (1hcr). See legend to Fig. 1 for details.

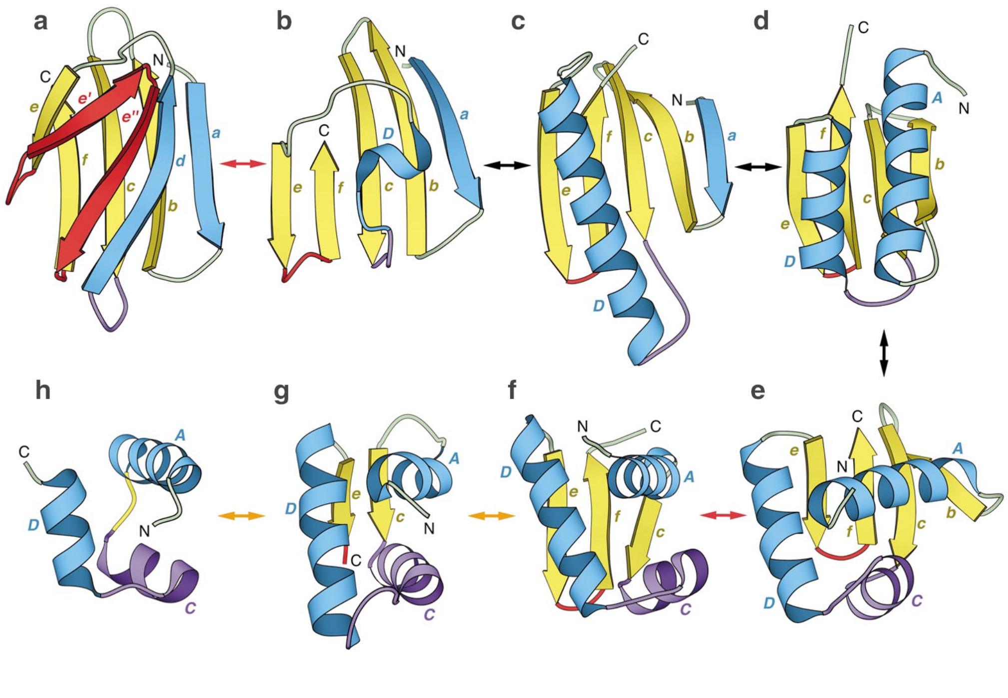



```
                      a                         A                            b              b'          b''          B               c    c'
(a) 2ctc   A58  NRPAIWIDLGIHSREW-ITQATGVWFAKKFTEDYGQ(9)SMDIFLEIVT(14)WRKTRSVTSSSLCVGVDANRNWDAGF--GKAGASSSPCSETYHGKYANSEV
(b) 2dln    A1  MTDKIAVLLGGTSAEREVSLNSGAAVLAGLREGG------IDAYPVDPK-------------------------------------------------------
(c) 1bnc    A1  MLDKIVIAN---------RGEIALRILRACKELG------IKTVAVHSS--------------------ADRDLKHVL--------LADETVCIGPAPSVKSYL
(d) 1auv  A114  --RVLLVIDEPH--------TDWAKYFKGKKIHGEI---DIKVEQAEFS----DLNLVAHAN--GGFSVDMEVLRN------------GVKVVR----------
(e) 1gsa     2  --IKLGIVMDPIANI-NIKKDSSFAMLLEAQRRG------YELHYMEMG----DLYLIN-------GEARAHTRTLNVKQNYEEWFSFVGEQDL----------
                                                                                                  b'''

                         C         d                    D           e   e'                                  E
(a) 2ctc  A175  EVKSIVDFVKDHGNFKAFLSIHSYSQL(47)QASGGSIDWSYNQGIKYSFTFELRDTGRYGFLL---PASQIIPTAQETWL-----GVLTIMEHTLNN--  A307
(b) 2dln   A44  --EVDVTQLKSMGFQKVFIALHGRG------GEDGTLQGMLELMGLPYTGSGVMASALSMDKL(180)SHSLVPMAARQAGMSFSQLVVRILELAD----  A306
(c) 1bnc   A62  NIPAIISAAEITGAVAIHPGYGFL-------SENANFAEQVERSGFIFIGPKAETIRLMGDKV(176)-QVEHPVTEMITGVD---LIKEQLRIAAGQP-  A320
(d) 1auv  A177  -------SLK---PDFVLIRQHAFSMAR--NGDYRSLVIGLQYAGIPSI-NSLHSVYNFCDKP(167)-LIGDHQDEDKQL-----IVELVVNKMAQA--  A417
(e) 1gsa    76  -------PLAD--LDVILMRKDPPFDTE--FIYATYILERAEEKGTLIV-NKPQSLRDCNEKL(160)SPTCIREIEAEFPVS---ITGMLMDAIEARLQ   314
                       c'                                   E'
```

3 2ctc - Carboxypeptidase A 2dln - Dala-Dala Ligase 1bnc - Biotin Carboxylase 1auv - Synapsin 1gsa - Glutathione Synthase

---



4

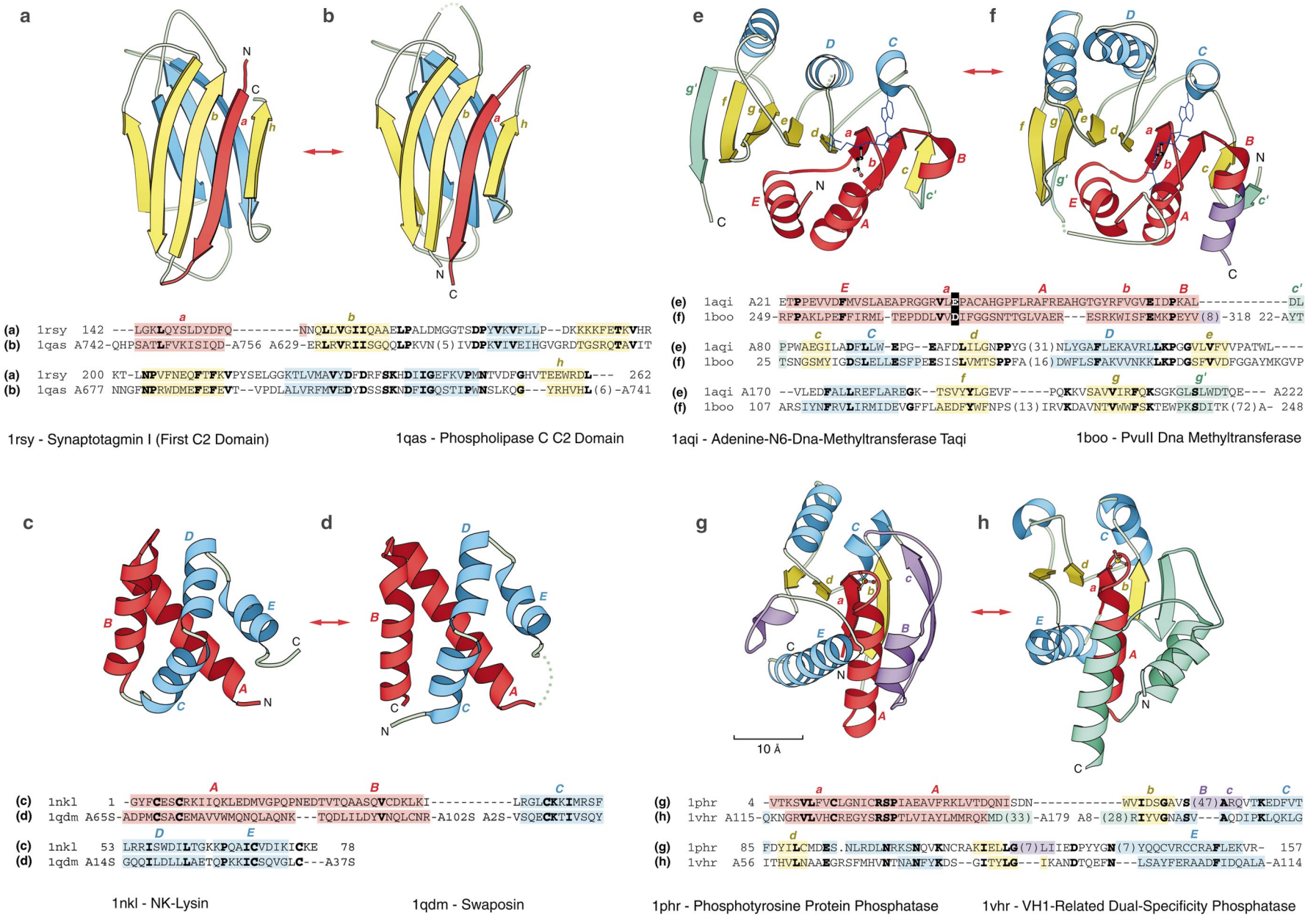


FIG. 5. Circular permutations. (a) Synaptotagmin C2 domain (1rsy); (b) phospholipase C C2 domain (1qas); (c) NK-lysin (1nkl); (d) swaposin (1qdm); (e) adenine-specific DNA methyltransferase (1aqi); (f) *Pvu*II DNA methyltransferase (1boo); (g) phosphotyrosine protein phosphatase (1phr); (h) VH1 dual-specificity phosphatase (1vhr). See legend to Fig. 1 for details.

polypeptide chain is made up of two fused active enzymes (Leismann *et al.*, 1998).

Protein Phosphatases

C2 domains (all β), saposins (all α), and methyltransferases (α/β) show that circular permutations occur naturally in proteins of different structural classes and occupy different fractions of the polypeptide chain. However, these circular permutations are the only events that change the protein fold in the example above. Finally, we discuss a possible circular permutation case in which detection is complicated by several large insertions. Phosphotyrosine protein phosphatases/sulfurtransferases are characterized by the common catalytic mechanism and a conserved motif CX_5R in their active site (Fauman *et al.*, 1998). However, due to significant structural differences, they have been classified into several distinct folds in SCOP (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995). We limit our discussion to two of them, namely, families I and II of the phosphotyrosine phosphatases. Low-molecular-weight phosphotyrosine protein phosphatase (Su *et al.*, 1994) is a type I enzyme that possesses a recognizable doubly wound Rossmann-like fold (Fig. 5g) with the active site within the first $\beta\alpha$ unit aA. Dual-specificity protein phosphatase VHR (Yuvaniyama *et al.*, 1996), a representative of family II, is characterized by a unique topology (Fig. 5f), which can, however, be converted to the type I enzyme by deletions/insertions and a circular permutation. This conversion results in a superposition of the active sites and produces sequence alignment with 20% identity. The $\beta\alpha$ unit shown in red in Figs. 5g and 5h and is placed at the N-terminus of phosphatase I and is C-terminal in phosphatase II. An additional $\alpha\beta$ unit in phosphatase I (Bc, shown in purple in Fig. 5g) is deleted in phosphatase II. Phosphatase II contains N- and C-terminal extensions (shown in green in Fig. 5f). It is likely that these phosphatases share an evolutionary origin with yet another family, rhodanese/CDC phosphatase (Bordo *et al.*, 2000; Fauman *et al.*, 1998), that is related to them by a different circular permutation.

Causes for Permutations

Circular permutations are frequent events in protein evolution that represent a common way to generate protein chains with different topologies. Their wide distribution is facilitated by the proximity on N- and C-termini in protein spatial structures and a high frequency of gene duplications resulting in tandem repeats. However, permutations might be easily missed during sequence and structure analyses, since commonly used methods of similarity detection do not take them into account (Uliel *et al.*, 1999). What could be the reason for evolutionary fixation of a circular permutation? Some of them are likely to be neutral. However, many domains that are inserted in other proteins appear to be circularly permuted in comparison with their homologs that exist as a separate polypeptide chain. It is possible that there are some specific folding reasons for this. It is also clear that circularly permuted variants will expose different regions to the solvent, which might participate in the domain interface if the inserted domain is not permuted (Ponting and Russell, 1995). Circular permutation may offer a mechanism for generating diversity analogous to recombination due to the hybrid nature of the permuted protein that is composed of the two segments from slightly different genes.

STRAND INVASION/WITHDRAWAL

Indels/substitutions and circular permutations are well-known and extensively studied events in protein evolution. Acting together, they could result in transformation of a protein fold beyond recognition. However, as individual small-step changes they usually do not introduce drastic alterations to the protein structure, especially in the conserved core regions. The next two sections discuss events that involve rearrangement of hydrogen bonds in the β-sheets and as such might affect the very core regions in proteins. The disruption (creation) of hydrogen bonds at the edge of a planar β-sheet by deletions/insertions that involve peripheral β-strands is typical in globular proteins and was discussed above. The disruption in H-bonding of internal β-strands, which are believed to stabilize the protein structure, appears to be problematic. The insertion of a β-strand(s) into existing β-sheets that requires H-bond breakage and formation of the H-bonding pattern on both sides of the inserted β-strand(s) is termed “invasion” here.

Lipocalins

Additions to the edge are possible only for planar β-sheets in which there is an edge to start with. In β-barrels, an insertion that is structured as a β-strand H-bonding to the existing circular β-sheet would inevitably cause invasion. Such an insertion would be considered a minor change for a planar β-sheet, but causes drastic topological differences for a barrel. Lipocalins, β-barrels that bind hydrophobic ligands in their interior, offer an example of such invasion. Retinol-binding protein forms an 8-stranded β-barrel (Zanotti *et al.*, 1993) (Fig. 6a). Retionic acid-binding protein (Kleywegt *et al.*, 1994), however, folds as a 10-stranded β-barrel (Fig. 6b).

From the classical fold definition standpoint, such a difference warrants placement of retinol- and retinoic acid-binding proteins in two different folds. The following similarities, however, indicate homology between these proteins. First, the structural similarity in the common regions is pronounced, which includes conserved length and tilt of the β-strands. Second, both proteins are functionally similar and bind lipids inside the β-barrel. The difference between their folds can be explained by an invasion of a β-hairpin d′d″ (shown in red in Figs. 6a and 6b) between β-strands d and e of retinol-binding protein (Figs. 6a and 6b). The β-barrel of retinol-binding protein can be viewed as being composed of two halves (shown in yellow and blue in Fig. 6a; potential duplication). Each half is structured as a 4-stranded β-sheet. Notably, the insertion in the retinoic acid-binding protein sequence is placed between these two halves. By means of hairpin invasion, lipocalins change the size of their interior cavity and can adapt to binding ligands of different shapes and sizes. A similar mechanism is likely to be used by porins (Koebnik *et al.*, 2000).

Serpins

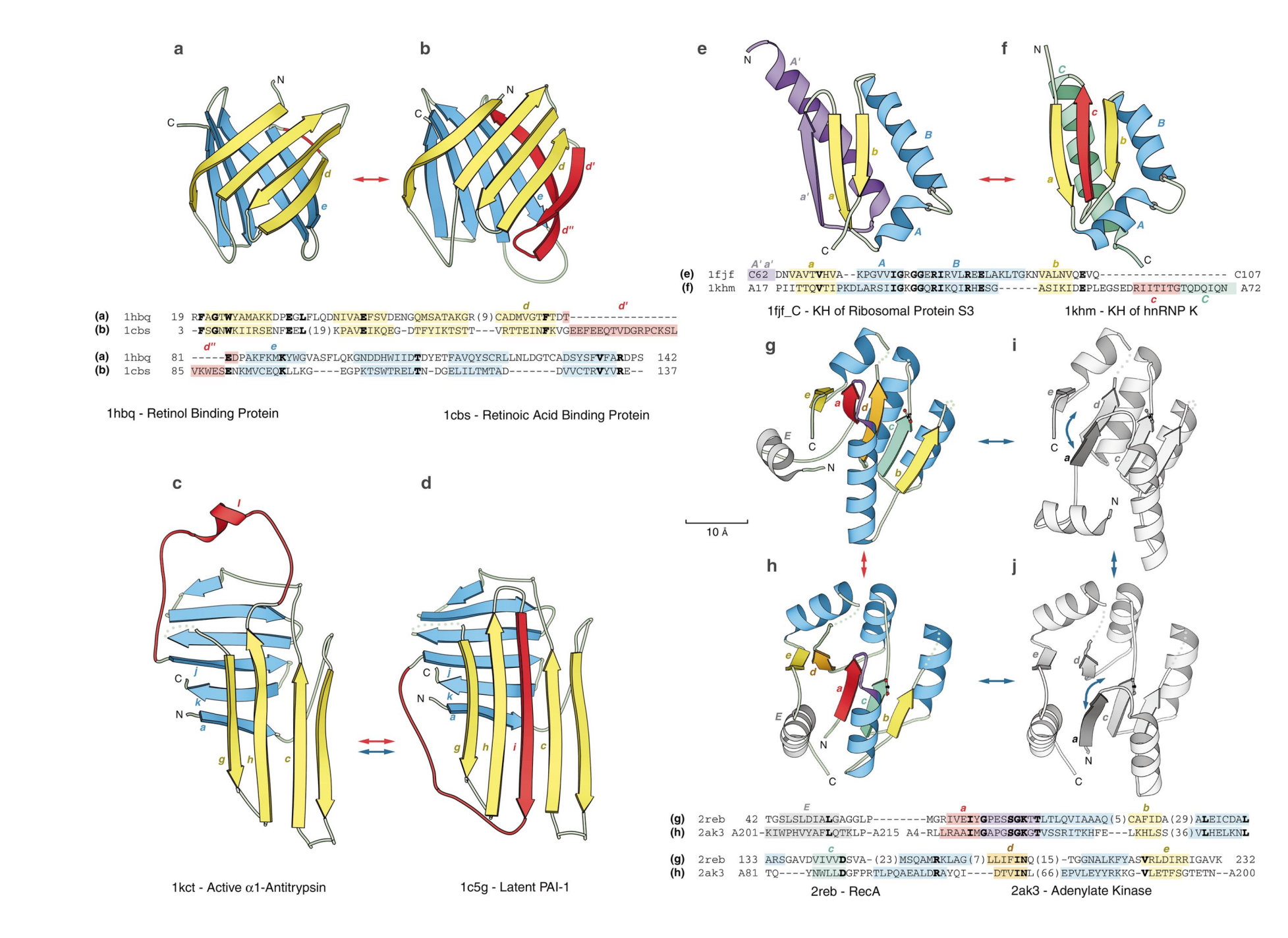
There is a unique protein family that can help us to understand possible mechanisms of strand invasion/withdrawal. In serine protease inhibitors, serpins, β-strand invasion occurs as a natural physiological process within the same molecule (Whisstock *et al.*, 1998). The active form of the inhibitor illustrated by the structure of α1-antitrypsin (Song *et al.*, 1995) (Fig. 6c) contains a long loop that connects β-strands h and j. The central part of this loop is folded as an α-helix i. This region interacts with the protease causing inhibition (Huntington *et al.*, 2000). In the latent form of the inhibitor shown for PAI-1 (Mottonen *et al.*, 1992) (Fig. 6d), into which some active serpins convert spontaneously with time, the loop is inserted into the β-sheet between β-strands h and c forming β-strand i. This drastic structural change was first postulated on the basis of a proteolytically cleaved serpin structure in which the β-strand i was inserted in the β-sheet and two sequentially consecutive residues between which cleavage occurs were separated by about 70 Å (Engh *et al.*, 1990; Loebermann *et al.*, 1984). Serpins fold into a metastable structure that can exist in at least two drastically different conformations. Such structures may arise in protein evolution as a consequence of substitutions/indels. Since these metastable proteins still fold and might be functional, they will not be eliminated. Further mutations may result in the fixation of one of the conformations and cause a change of fold.

KH Domains

The next example illustrates exactly this point, and structural change between the two folds of the K homology (KH) domain is similar to the difference between active and latent serpins. The KH module is a widespread RNA-binding motif. KH was first biochemically characterized in the major pre-mRNA-binding protein K (hnRNP K) and described as a 45- to 50-amino-acid repeat detected by sequence similarity in a number of RNA-binding proteins (Siomi *et al.*, 1993). Siomi *et al.* (1993) noted that similarity was particularly strong with ribosomal protein S3. Analysis of spatial structures of KH domains in hnRNP K (Baber *et al.*, 1999) and S3 (Wimberly *et al.*, 2000) reveals that they are topologically dissimilar and thus belong to different protein folds (Figs. 6e and 6f) (Grishin, 2001). The KH motif region covers two β-strands and two α-helices (aABb in Figs. 6e and 6f). The two distinct topologies might have arisen from an ancestral KH-motif protein by N- and C-terminal extensions, or one of the existing topologies may have evolved from the other by extension, displacement, and deletion. The S3 domain contains an N-terminal extension folded in an αβ unit (A′a′, shown in purple in Fig. 6e). The hnRNP K domain has a C-terminal βα extension (cC in Fig. 6f). The β-strand c of this extension appears to be inserted between β-strands a and b (Figs. 6e and 6f). Additionally, KH domains of S3 and hnRNP K give an example of homologs that diverged to different topologies while converging to the same architecture (3-stranded β-sheet with three α-helices on one side).

P-loop ATPases

P-loop NTPases are probably the most diversified and abundant protein superfamily (Saraste *et al.*, 1990; Wolf *et al.*, 1999). These enzymes are characterized by Walker A (P-loop) and B motifs (Walker *et al.*, 1982). The detection of proteins with these motifs by sequence analysis tools is relatively straightforward and a monophyletic origin of the P-loop NTPases has been proposed (Koonin, 1993; Neuwald *et al.*, 1999). Representative structures for the most of the distinct families of P-loop NTPases have been determined (Lo Conte *et al.*, 2000; Murzin *et al.*, 1995). All these structures display a certain resemblance: they have αβα sandwich architecture with the central mainly parallel β-sheet and are composed of βα units. However, connectivity between these βα units is not the same in different NTPase families. The notable difference is in relative location of Walker A and B motifs. The two rather common β-sheet topologies are illustrated by the struc-


a
b
N
C
d
e
d'
d''
(a) 1hbq 19 RFAGTWYAMAKKDPEGLFLQDNIVAEFSVDENGQMSATAKGR(9)CADMVGTFIDT---------- 81 -----EDPAKFKMKYWGVASFLQKGNDDHWIIDTDYETFAVQYSCRLLNLDGTCADSYSFVFARDPS 142
(b) 1cbs 3 -FSGNWKIIRSENFEEL(19)KPAVEIKQEG-DTFYIKTSTT---VRTTEINFKVGEEFEEQTVDGRPCKSL 85 VKWESENKMVCEQKLLKG----EGPKTSWTRELTN-DGELILTMTAD-------DVVCTRVYVRE-- 137
1hbq - Retinol Binding Protein
1cbs - Retinoic Acid Binding Protein
c
d
l
j
k
a
g
h
c
i
1kct - Active α1-Antitrypsin
1c5g - Latent PAI-1
e
f
A'
a'
A
B
C
(e) 1fjf C62 DNVAVTVHVA--KPGVVIGRGGERIRVLREELAKLTGKNVALNVQEVQ---------- C107
(f) 1khm A17 PIITTQVTIPKDLARSIIGKGGQRIKQIRHESG------ASIKIDEPLEGSEDRIITITGTQDQIQN A72
1fjf_C - KH of Ribosomal Protein S3
1khm - KH of hnRNP K
g
i
h
j
E
10 Å
(g) 2reb 42 TGSLSLDIALGAGGLP--------MGRIVEIYGPESSGKTTLTLQVIAAAQ(5)CAFIDA(29)ALEICDAL
(h) 2ak3 A201-KIWPHVYAFLQTKLP-A215 A4-RLLRAAIMGAPGSGKGTVSSRITKHFE---LKHLSS(36)VLHELKNL
(g) 2reb 133 ARSGAVDVIVVDSVA-(23)MSQAMRKLAG(7)LLIFINQ(15)-TGGNALKFYASVRLDIRRIGAVK 232
(h) 2ak3 A81 TQ----YNWLLDGFPRTLPQAEALDRAYQI----DTVINL(66)EPVLEYYRKKG-VLETFSGTETN--A200
2reb - RecA
2ak3 - Adenylate Kinase

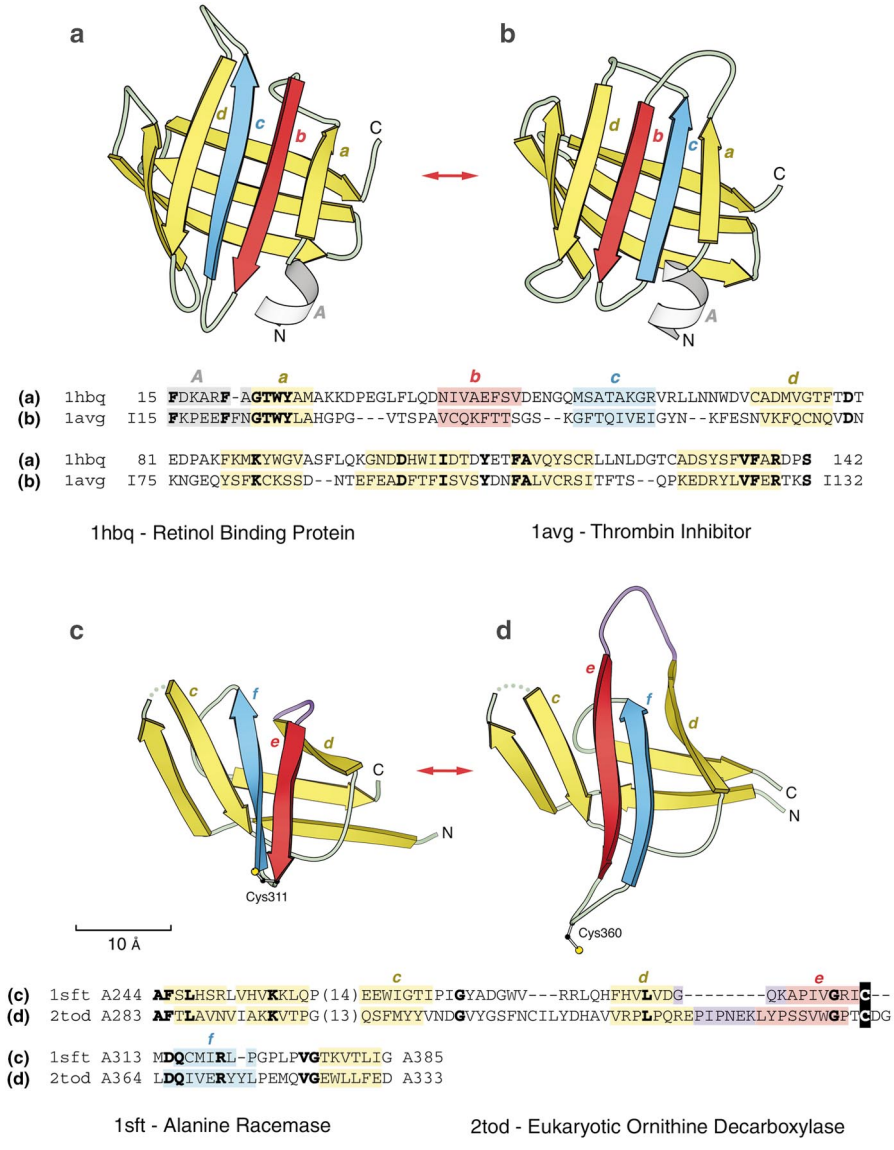


FIG. 7. Hairpin flips/swaps. (a) Retinol-binding protein (1hbq); (b) thrombin inhibitor triabin (1avg); (c) alanine racemase C-terminal domain (1sft); (d) eukaryotic ornithine decarboxylase C-terminal domain (2tod). See legend to Fig. 1 for details.

tures of RecA protein (Story *et al.*, 1992) and adenylate kinase (Diederichs and Schulz, 1991) (Figs. 6g and 6h). The Walker A motif (shown in purple in Figs. 6f and 6g) follows the first β-strand in both structures (a, shown in red in Figs. 6f and 6g). Walker B with the conserved acidic Mg^{2+}-binding residue is structured as a β-strand (c, shown in green in Figs. 6g and 6h). In adenylate kinase, this β-strand c is adjacent to the Walker A β-strand a and forms H-bonds with it (Fig. 6h). In RecA, β-strands a and c are separated by a β-strand d (Fig. 6g). The difference between the two arrangements of the β-sheet can be rationalized in terms of strand invasion/withdrawal. To convert from RecA topology to adenylate kinase topology, the β-strand a can be taken out of the β-sheet from its location between the β-strands e and d (Fig. 6i) and inserted in between strands d and c (Fig. 6j). Such interconversion

FIG. 6. Strand invasions/withdrawal. (a) Retinol-binding protein (1hbq); (b) retinoic acid-binding protein (1cbs); (c) active α1-antitrypsin (1kct); (d) latent PAI-1 (1c5g); (e) KH domain of ribosomal protein S3 (1fjf, chain C); (f) KH domain of hnRNP K; (g) RecA ATPase domain (2reb); (i) and (j) hypothetical intermediates; (h) adenylate kinase (2ak3). See legend to Fig. 1 for details.

postulates the existence of a metastable intermediate that, like serpins, can exist in both states.

β-HAIRPIN FLIP/SWAP

Another type of structural rearrangement in β-sheets can be described in terms of internal swapping of the β-strands. This is more commonly observed for the two β-strands adjacent to each other and forming a β-hairpin. Such a rearrangement creates (or removes) crossing loops and may be a common mechanism in generating unusual topologies in β-sheet proteins.

Lipocalins

The structure of thrombin inhibitor triabin (Fuentes-Prior *et al.*, 1997) revealed an unexpected irregularity in comparison to its homologs, the lipocalins. A typical lipocalin, for example, retinol-binding protein, is folded as an 8-stranded up-and-down β-barrel (Zanotti *et al.*, 1993) (Fig. 7a). Triabin shares significant and easily detectable sequence similarity with retinol-binding proteins and structurally is more similar to them than to any other protein family. However, the N-terminal regions of the structures (β-strands abcd in Figs. 7a and 7b) are topologically distinct and thus triabin and retinol-binding protein can be classified into different folds. To convert the up-and-down topology of retinol-binding protein to the unique topology of triabin, one can imagine a "flip" of a β-hairpin bc 180° around its axis (Figs. 7a and 7b). Mechanistically, however, such a flip is hardly distinguishable from the "swap" of the two β-strands with each other. In any event, the interchange of the two β-strands resulted in a parallel arrangement of b and d on the one hand and of a and c on the other hand (versus an up-and-down all-antiparallel topology) and created crossing loops in the triabin structure. This flip/swap is likely to have functional reasons (Murzin, 1998). Triabin does not bind ligands inside the barrel as most other lipocalins do, but functions as a protease inhibitor. The β-strands flip/swap in triabin guides a loop between the β-strands c and d across the open end of the β-barrel. The loop blocks the entrance of a lipocalin ligand-binding site (Fuentes-Prior *et al.*, 1997; Murzin, 1998).

Alanine Racemase—Eukaryotic Ornithine Decarboxylase

Homology between alanine racemase and eukaryotic ornithine decarboxylase (ODC) has been detected by sequence analysis (Grishin *et al.*, 1995). The structures of these PLP-dependent enzymes confirmed the presence of a βα-barrel domain and revealed unexpected structural differences in the C-terminal β-barrel domain (Kern *et al.*, 1999; Shaw *et al.*, 1997). Similarly to the lipocalin example, the topologies of the β-barrel domains in alanine racemase and ODC are related by a hairpin flip/swap (Grishin *et al.*, 1999) (hairpin ef in Figs. 7c and 7d). In alanine racemase, the β-barrel has a common Greek key topology with an all-antiparallel β-sheet (Fig. 7c). In ODC, the β-barrel is open and contains parallel β-strands and crossing loops (Fig. 7d). Surprisingly, however, the sequence similarity is the strongest in the region of the flip/swap (Figs. 7c and 7d). This similarity indicates that the rearrangement is a swap rather than a flip, since the side chains pointing toward the core in one structure are not pointing outward in the other structure. Since the β-hairpin ef contains an active site cysteine residue, the swap is likely to have functional reasons and is correlated with the differences in the mutual orientation of the two domains in alanine racemase and ODC. When βα-barrel domains of alanine racemase and ODC are superimposed, the β-barrel domains in the two structures are related by a 30° rotation. A hairpin swap compensates for this difference by restoring the position of the cysteine in the active site (Grishin *et al.*, 1999).

HOMOLOGY: GLOBAL OR LOCAL?

The strongest sequence similarity usually resides in several conserved motifs. Typically, there is a functional reason for conservation of motifs. For example, in P-loop NTPases, the Walker A motif facilitates NTP binding and Walker B incorporates a Mg^{2+} ligand (Walker *et al.*, 1982). It is a usual assumption that a pronounced sequence similarity in motif regions signifies homology. However, if the motifs are few and short, and the rest of the protein does not look similar between the potential homologs neither in sequence nor in structure, the question about local versus global homology arises. Indeed, two scenarios are possible. The first scenario assumes that the localized region of homology is inserted into unrelated structural templates giving rise to structural differences in "homologous proteins" (local homology). The second scenario explains the differences through gradual changes in a structure by transforming homologous regions with mutations and deletions and acquiring nonhomologous segments by insertions (global homology).

In any case, it becomes clear that domains are not the only units of homology. A domain could potentially contain the segments of homology embedded into regions acquired independently between proteins (local homology concept) (Fetrow and Godzik, 1998). On the other hand, if one prefers, segments acquired independently could be embedded into the

regions of homology (global homology concept). We would like to argue that there is no clear-cut distinction between these two scenarios and every homology is local to some extent. It seems likely that evolutionarily conserved domains that contain few indels are homologous throughout almost the entire length. Evolutionarily plastic proteins, in which only a few structural segments are functionally crucial, sustain significant evolutionary drift. Lysozyme-like proteins can be an example (Holm and Sander, 1997c; Lo Conte *et al.*, 2000; Murzin *et al.*, 1995). Most of the structural segments in these enzymes differ between family representatives and do not necessarily need to be homologous. Among existing protein structures, one can find almost every intermediate between the two extremes from highly localized to global homology. For example, local motifs such as Walker A (Matte *et al.*, 1996) and helix-turn-helix (Aravind and Koonin, 1999a) could be incorporated into larger structural templates that differ between proteins.

Additionally, it should be noted that even structural similarity around the conserved motifs does not necessarily indicate homology between these regions. A short alien segment could potentially substitute for a segment of similar size within a gene. Long-range interactions in protein folding are likely to force this alien segment into the conformation of the segment it has substituted for. Indeed, context-dependent secondary structure formation has been observed experimentally. It has been shown that the 11-residue sequence folds as an α-helix in one position, but adopts a β-sheet conformation in a different position of the sequence of the protein G IgG-binding domain (Minor and Kim, 1996). At present, methods for analyzing irregularities in protein structures and sequences discussed here are not well developed. However, with the development of novel approaches that specifically take these irregularities into account (Uliel *et al.*, 1999), and with the explosion of structural information that is expected from the structural genomics projects (Sali, 1998; Service, 2000), a more complete and comprehensive picture will emerge.

FOLD CHANGE IN EVOLUTION: PRACTICAL IMPLICATIONS

There are at least three main lessons learned from the example discussed in this review. First, the existence of homologs that can be detected by sequence analysis methods but fold into different structures brings additional challenges to homology modeling techniques. Second, it introduces difficulties and potential contradictions in the protein classification schemes. Third, understanding mechanisms of naturally occurring fold change would facilitate protein design.

Two types of protein structural classification are conceivable (May, 1999). In phenetic classification, only structural similarity is taken into account. Phyletic classification is based on evolutionary relationships between proteins. Only phyletic classification appears to be natural. Since structures can change substantially in evolution, a contradiction between the two approaches inevitably arises. How can we resolve the problem? One approach would be to modify the fold definitions to incorporate structural differences between homologous proteins. The simpler way, however, is to accept that homology and fold similarity can go their separate ways, despite being strongly correlated. Thus, structural classification of proteins cannot be phyletic. On the contrary, classification of protein structures could have an evolutionary basis (Lo Conte *et al.*, 2000; Murzin, 1998; Murzin *et al.*, 1995). In such a classification, different folds within one superfamily are allowed. Classification of structures according to their evolutionary history would be more similar to a functional classification, in which many different protein folds can be used to perform the same chemical function (Thornton *et al.*, 1999). The main difficulty with any phyletic classification is the loss of evolutionary signal with time. Indeed, if the structural rearrangement occurred rather recently, or there are significant functional restraints on the two proteins, the signal remains and enables us to detect the relationship. If the signal is lost, it might not be possible to link two proteins.

In summary, analysis of available protein spatial structures revealed that there is no strict correlation between homology and fold similarity. Homologous proteins can have different folds, and mechanisms such as insertions/deletions/substitutions, circular permutations, strand invasions/withdrawals, and hairpin flips/swaps emerge as leading causes for globally different protein structures within homologous families.

The author is grateful to Hong Zhang and Tammiko Jones for critically reading the manuscript and helpful comments.

REFERENCES

Adams, M. J., Ford, G. C., Koekoek, R., Lentz, P. J., McPherson, A., Jr., Rossmann, M. G., Smiley, I. E., Schevitz, R. W., and Wonacott, A. J. (1970) Structure of lactate dehydrogenase at 2–8 Å resolution, *Nature* **227,** 1098–1103.

Altschul, S. F., and Koonin, E. V. (1998) Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases, *Trends Biochem. Sci.* **23,** 444–447.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.* **25,** 3389–3402.

Aravind, L., and Koonin, E. V. (1999a) DNA-binding proteins and evolution of transcription regulation in the archaea, *Nucleic Acids Res.* **27,** 4658–4670.

Aravind, L., and Koonin, E. V. (1999b) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches, *J. Mol. Biol.* **287,** 1023–1040.

Ay, J., Hahn, M., Decanniere, K., Piotukh, K., Borriss, R., and Heinemann, U. (1998) Crystal structures and properties of de novo circularly permuted 1,3-1,4-β-glucanases, *Proteins* **30,** 155–167.

Baber, J. L., Libutti, D., Levens, D., and Tjandra, N. (1999) High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor, *J. Mol. Biol.* **289,** 949–962.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D., and Waley, S. G. (1975) Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data, *Nature* **255,** 609–614.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000) The Pfam protein families database, *Nucleic Acids Res.* **28,** 263–266.

Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins, *J. Mol. Biol.* **229,** 1065–1082.

Blanco, F. J., Angrand, I., and Serrano, L. (1999) Exploring the conformational properties of the sequence space between two proteins with different folds: An experimental study, *J. Mol. Biol.* **285,** 741–753.

Bordo, D., Deriu, D., Colnaghi, R., Carpen, A., Pagani, S., and Bolognesi, M. (2000) The crystal structure of a sulfurtransferase from *Azotobacter vinelandii* highlights the evolutionary relationship between the rhodanese and phosphatase enzyme families, *J. Mol. Biol.* **298,** 691–704.

Bowles, D. J., Marcus, S. E., Pappin, D. J., Findlay, J. B., Eliopoulos, E., Maycox, P. R., and Burgess, J. (1986) Posttranslational processing of concanavalin A precursors in jackbean cotyledons, *J. Cell Biol.* **102,** 1284–1297.

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Natl. Acad. Sci. USA* **95,** 6073–6078.

Carrington, D. M., Auffret, A., and Hanke, D. E. (1985) Polypeptide ligation occurs during post-translational modification of concanavalin A, *Nature* **313,** 64–67.

Chothia, C. (1984) Principles that determine the structure of proteins, *Annu. Rev. Biochem.* **53,** 537–572.

Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5,** 823–826.

Chothia, C., Levitt, M., and Richardson, D. (1977) Structure of proteins: Packing of alpha-helices and pleated sheets, *Proc. Natl. Acad. Sci. USA* **74,** 4130–4134.

Cordes, M. H., Walsh, N. P., McKnight, C. J., and Sauer, R. T. (1999) Evolution of a protein fold in vitro, *Science* **284,** 325–328.

Cunningham, B. A., Hemperley, J. J., Hopp, T. P., and Edelman, G. M. (1979) Favin versus concanavalin A: Circularly-permuted amino acid sequences, *Proc. Natl. Acad. Sci. USA* **76,** 3218–3222.

Dalal, S., Balasubramanian, S., and Regan, L. (1997a) Protein alchemy: Changing beta-sheet into alpha-helix, *Nat. Struct. Biol.* **4,** 548–552.

Dalal, S., Balasubramanian, S., and Regan, L. (1997b) Transmuting alpha helices and beta sheets, *Fold Des.* **2,** R71–R79.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) A model of evolutionary change in proteins, *in* Atlas of Protein Sequences and Structures Dayhoff, M. O. (Ed.), Vol. 5, Suppl. 3, pp. 345–352. National Biomedical Research Foundation, Washington, DC.

Diederichs, K., and Schulz, G. E. (1991) The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution, *J. Mol. Biol.* **217,** 541–549.

Doolittle, R. F. (1981) Similar amino acid sequences: Chance or common ancestry? *Science* **214,** 149–159.

Doolittle, R. F. (1994) Convergent evolution: The need to be explicit, *Trends Biochem. Sci.* **19,** 15–18.

Eads, J. C., Ozturk, D., Wexler, T. B., Grubmeyer, C., and Sacchettini, J. C. (1997) A new function for a common fold: The crystal structure of quinolinic acid phosphoribosyltransferase, *Structure* **5,** 47–58.

Eddy, S. R. (1998) Profile hidden Markov models, *Bioinformatics* **14,** 755–763.

Egas, C., Lavoura, N., Resende, R., Brito, R. M., Pires, E., Pedroso De Lima, M. C., and Faro, C. (2000) The saposin-like domain of the plant aspartic proteinase precursor is a potent inducer of vesicle leakage, *J. Biol. Chem.* **275,** 38190–38196.

Engh, R. A., Wright, H. T., and Huber, R. (1990) Modeling the intact form of the alpha 1-proteinase inhibitor, *Protein Eng.* **3,** 469–477.

Esnouf, R. M. (1997) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities, *J. Mol. Graph. Model.* **15,** 133–138.

Essen, L. O., Perisic, O., Cheung, R., Katan, M., and Williams, R. L. (1996) Crystal structure of a mammalian phosphoinositide-specific phospholipase C delta, *Nature* **380,** 595–602.

Esser, L., Wang, C. R., Hosaka, M., Smagula, C. S., Sudhof, T. C., and Deisenhofer, J. (1998) Synapsin I is structurally similar to ATP-utilizing enzymes, *EMBO J.* **17,** 977–984.

Fan, C., Moews, P. C., Walsh, C. T., and Knox, J. R. (1994) Vancomycin resistance: Structure of D-alanine:D-alanine ligase at 2.3 Å resolution, *Science* **266,** 439–443.

Fauman, E. B., Cogswell, J. P., Lovejoy, B., Rocque, W. J., Holmes, W., Montana, V. G., Piwnica-Worms, H., Rink, M. J., and Saper, M. A. (1998) Crystal structure of the catalytic domain of the human cell cycle control phosphatase, Cdc25A, *Cell* **93,** 617–625.

Feng, J. A., Johnson, R. C., and Dickerson, R. E. (1994) Hin recombinase bound to DNA: The origin of specificity in major and minor groove interactions, *Science* **263,** 348–355.

Fetrow, J. S., and Godzik, A. (1998) Function driven protein evolution. A possible proto-protein for the RNA-binding proteins, *Pac. Symp. Biocomput.* **3,** 485–496.

Fisher, A. J., Raushel, F. M., Baldwin, T. O., and Rayment, I. (1995) Three-dimensional structure of bacterial luciferase from *Vibrio harveyi* at 2.4 Å resolution, *Biochemistry* **34,** 6581–6586.

Flores, T. P., Orengo, C. A., Moss, D. S., and Thornton, J. M. (1993) Comparison of conformational characteristics in structurally similar protein pairs, *Protein Sci.* **2,** 1811–1826.

Fuentes-Prior, P., Noeske-Jungblut, C., Donner, P., Schleuning, W. D., Huber, R., and Bode, W. (1997) Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug, *Proc. Natl. Acad. Sci. USA* **94,** 11845–11850.

Galperin, M. Y., and Koonin, E. V. (1997) A diverse superfamily of

enzymes with ATP-dependent carboxylate-amine/thiol ligase activity, *Protein Sci.* **6,** 2639–2643.

Goldenberg, D. P. (1989) Circularly permuted proteins, *Protein Eng.* **2,** 493–495.

Gong, W., O'Gara, M., Blumenthal, R. M., and Cheng, X. (1997) Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment, *Nucleic Acids Res.* **25,** 2702–2715.

Graf, R., and Schachman, H. K. (1996) Random circular permutation of genes and expressed polypeptide chains: Application of the method to the catalytic chains of aspartate transcarbamoylase, *Proc. Natl. Acad. Sci. USA* **93,** 11591–11596.

Grishin, N. V. (1997) Estimation of evolutionary distances from protein spatial structures, *J. Mol. Evol.* **45,** 359–369.

Grishin, N. V. (2001) KH domain: one motif, two folds, *Nucleic Acids Res.* **29,** 638–643.

Grishin, N. V., Osterman, A. L., Brooks, H. B., Phillips, M. A., and Goldsmith, E. J. (1999) X-ray structure of ornithine decarboxylase from *Trypanosoma brucei:* The native structure and the structure in complex with alpha-difluoromethylornithine, *Biochemistry* **38,** 15174–15184.

Grishin, N. V., Phillips, M. A., and Goldsmith, E. J. (1995) Modeling of the spatial structure of eukaryotic ornithine decarboxylases, *Protein Sci.* **4,** 1291–1304.

Hadley, C., and Jones, D. T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP, *Structure Fold Des.* **7,** 1099–1112.

Hall, T. M., Porter, J. A., Beachy, P. A., and Leahy, D. J. (1995) A potential catalytic site revealed by the 1.7-Å crystal structure of the amino-terminal signalling domain of Sonic hedgehog, *Nature* **378,** 212–216.

Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.

Holm, L. (1998) Unification of protein families, *Curr. Opin. Struct. Biol.* **8,** 372–379.

Holm, L., and Sander, C. (1996) Mapping the protein universe, *Science* **273,** 595–603.

Holm, L., and Sander, C. (1997a) Dali/FSSP classification of three-dimensional protein folds, *Nucleic Acids Res.* **25,** 231–234.

Holm, L., and Sander, C. (1997b) Decision support system for the evolutionary classification of protein structures, *Ismb* **5,** 140–146.

Holm, L., and Sander, C. (1997c) New structure—Novel fold? *Structure* **5,** 165–171.

Hubbard, T. J., and Blundell, T. L. (1987) Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling, *Protein Eng.* **1,** 159–171.

Huntington, J. A., Read, R. J., and Carrell, R. W. (2000) Structure of a serpin-protease complex shows inhibition by deformation, *Nature* **407,** 923–926.

Jeltsch, A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases, *J. Mol. Evol.* **49,** 161–164.

Jones, D. T., Moody, C. M., Uppenbrink, J., Viles, J. H., Doyle, P. M., Harris, C. J., Pearl, L. H., Sadler, P. J., and Thornton, J. M. (1996) Towards meeting the Paracelsus Challenge: The design, synthesis, and characterization of paracelsin-43, an alpha-helical protein with over 50% sequence identity to an all-beta protein, *Proteins* **24,** 502–513.

Kern, A. D., Oliveira, M. A., Coffino, P., and Hackert, M. L. (1999) Structure of mammalian ornithine decarboxylase at 1.6 Å resolution: Stereochemical implications of PLP-dependent amino acid decarboxylases, *Structure Fold Des.* **7,** 567–581.

Kervinen, J., Tobin, G. J., Costa, J., Waugh, D. S., Wlodawer, A., and Zdanov, A. (1999) Crystal structure of plant aspartic proteinase prophytepsin: Inactivation and vacuolar targeting, *EMBO J.* **18,** 3947–3955.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., and Jones, T. A. (1994) Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-*trans*-retinoic acid and a synthetic retinoid, *Structure* **2,** 1241–1258.

Koebnik, R., Locher, K. P., and Van Gelder, P. (2000) Structure and function of bacterial outer membrane proteins: Barrels in a nutshell, *Mol. Microbiol.* **37,** 239–253.

Koonin, E. V. (1993) A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication, *Nucleic Acids Res.* **21,** 2541–2547.

Kraulis, P. J. (1991) MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures, *J. Appl. Crystallogr.* **24,** 946–950.

Lebioda, L., Stec, B., and Brewer, J. M. (1989) The structure of yeast enolase at 2.25-Å resolution. An 8-fold beta + alpha-barrel with a novel beta beta alpha alpha (beta alpha) 6 topology, *J. Biol. Chem.* **264,** 3685–3693.

Leismann, O., Roth, M., Friedrich, T., Wende, W., and Jeltsch, A. (1998) The *Flavobacterium okeanokoites* adenine-N6-specific DNA-methyltransferase M.FokI is a tandem enzyme of two independent domains with very different kinetic properties, *Eur. J. Biochem.* **251,** 899–906.

Levitt, M., and Chothia, C. (1976) Structural patterns in globular proteins, *Nature* **261,** 552–558.

Liepinsh, E., Andersson, M., Ruysschaert, J. M., and Otting, G. (1997) Saposin fold revealed by the NMR structure of NK-lysin, *Nat. Struct. Biol.* **4,** 793–795.

Lindqvist, Y., and Schneider, G. (1997) Circular permutations of natural protein sequences: Structural evidence, *Curr. Opin. Struct. Biol.* **7,** 422–427.

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., and Chothia, C. (2000) SCOP: A structural classification of proteins database, *Nucleic Acids Res.* **28,** 257–259.

Loebermann, H., Tokuoka, R., Deisenhofer, J., and Huber, R. (1984) Human alpha 1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function, *J. Mol. Biol.* **177,** 531–557.

Luger, K., Hommel, U., Herold, M., Hofsteenge, J., and Kirschner, K. (1989) Correct folding of circularly permuted variants of a beta alpha barrel enzyme in vivo, *Science* **243,** 206–210.

Machius, M., Wiegand, G., and Huber, R. (1995) Crystal structure of calcium-depleted *Bacillus licheniformis* alpha-amylase at 2.2 Å resolution, *J. Mol. Biol.* **246,** 545–559.

Malone, T., Blumenthal, R. M., and Cheng, X. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes, *J. Mol. Biol.* **253,** 618–632.

Matte, A., Goldie, H., Sweet, R. M., and Delbaere, L. T. (1996) Crystal structure of *Escherichia coli* phosphoenolpyruvate carboxykinase: A new structural family with the P-loop nucleoside triphosphate hydrolase fold, *J. Mol. Biol.* **256,** 126–143.

May, A. C. (1999) Toward more meaningful hierarchical classification of protein three-dimensional structures, *Proteins* **37,** 20–29.

Minor, D. L., Jr., and Kim, P. S. (1996) Context-dependent sec-

ondary structure formation of a designed protein sequence, *Nature* **380,** 730–734.

Moore, S. A., and James, M. N. (1994) Common structural features of the luxF protein and the subunits of bacterial luciferase: Evidence for a (beta alpha) 8 fold in luciferase, *Protein Sci.* **3,** 1914–1926.

Moore, S. A., James, M. N., O'Kane, D. J., and Lee, J. (1993) Crystal structure of a flavoprotein related to the subunits of bacterial luciferase, *EMBO J.* **12,** 1767–1774.

Morishita, Y., Hasegawa, K., Matsuura, Y., Katsube, Y., Kubota, M., and Sakai, S. (1997) Crystal structure of a maltotetraose-forming exo-amylase from *Pseudomonas stutzeri, J. Mol. Biol.* **267,** 661–672.

Mottonen, J., Strand, A., Symersky, J., Sweet, R. M., Danley, D. E., Geoghegan, K. F., Gerard, R. D., and Goldsmith, E. J. (1992) Structural basis of latency in plasminogen activator inhibitor-1, *Nature* **355,** 270–273.

Murzin, A. G. (1998) How far divergent evolution goes in proteins, *Curr. Opin. Struct. Biol.* **8,** 380–387.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* **247,** 536–540.

Nalefski, E. A., and Falke, J. J. (1996) The C2 domain calcium-binding motif: Structural and functional diversity, *Protein Sci.* **5,** 2375–2390.

Neuwald, A. F., Aravind, L., Spouge, J. L., and Koonin, E. V. (1999) AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes, *Genome Res.* **9,** 27–43.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH—A hierarchic classification of protein domain structures, *Structure* **5,** 1093–1108.

Otzen, D. E., and Fersht, A. R. (1998) Folding of circular and permuted chymotrypsin inhibitor 2: Retention of the folding nucleus, *Biochemistry* **37,** 8139–8146.

Owen, D. J., Noble, M. E., Garman, E. F., Papageorgiou, A. C., and Johnson, L. N. (1995) Two structures of the catalytic domain of phosphorylase kinase: An active protein kinase complexed with substrate analogue and product, *Structure* **3,** 467–482.

Pan, T., and Uhlenbeck, O. C. (1993) Circularly permuted DNA, RNA and proteins—A review, *Gene* **125,** 111–114.

Pappa, H., Murray-Rust, J., Dekker, L. V., Parker, P. J., and McDonald, N. Q. (1998) Crystal structure of the C2 domain from protein kinase C-delta, *Structure* **6,** 885–894.

Pascarella, S., and Argos, P. (1992) Analysis of insertions/deletions in protein structures, *J. Mol. Biol.* **224,** 461–471.

Pearl, F. M., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P., Thornton, J. M., and Orengo, C. A. (2000) Assigning genomic sequences to CATH, *Nucleic Acids Res.* **28,** 277–282.

Ponting, C. P., and Russell, R. B. (1995) Swaposins: Circular permutations within genes encoding saposin homologues, *Trends Biochem. Sci.* **20,** 179–180.

Ptitsyn, O. B., and Finkelstein, A. V. (1981) Similarities in protein topologies: Evolutionary divergence, functional convergence or principles of folding, *Q. Rev. Biophys.* **13,** 339–386.

Rees, D. C., Lewis, M., and Lipscomb, W. N. (1983) Refined crystal structure of carboxypeptidase A at 1.54 Å resolution, *J. Mol. Biol.* **168,** 367–387.

Richardson, J. S. (1977) β-Sheet topology and the relatedness of proteins, *Nature* **268,** 495–500.

Richardson, J. S. (1981) The anatomy and taxonomy of protein structure, *Adv. Protein Chem.* **34,** 167–339.

Russell, R. B. (1998) Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution, *J. Mol. Biol.* **279,** 1211–1227.

Russell, R. B., and Ponting, C. P. (1998) Protein fold irregularities that hinder sequence analysis, *Curr. Opin. Struct. Biol.* **8,** 364–371.

Russell, R. B., Saqi, M. A., Bates, P. A., Sayle, R. A., and Sternberg, M. J. (1998) Recognition of analogous and homologous protein folds—Assessment of prediction success and associated alignment accuracy using empirical substitution matrices, *Protein Eng.* **11,** 1–9.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A., and Sternberg, M. J. (1997) Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation, *J. Mol. Biol.* **269,** 423–439.

Sali, A. (1998) 100,000 protein structures for the biologist, *Nat. Struct. Biol.* **5,** 1029–1032.

Saraste, M., Sibbald, P. R., and Wittinghofer, A. (1990) The P-loop—A common motif in ATP- and GTP-binding proteins, *Trends Biochem. Sci.* **15,** 430–434.

Schluckebier, G., Kozak, M., Bleimling, N., Weinhold, E., and Saenger, W. (1997) Differential binding of *S*-adenosylmethionine *S*-adenosylhomocysteine and Sinefungin to the adenine-specific DNA methyltransferase M. TaqI, *J. Mol. Biol.* **265,** 56–67.

Schultz, S. C., Shields, G. C., and Steitz, T. A. (1991) Crystal structure of a CAP–DNA complex: The DNA is bent by 90 degrees, *Science* **253,** 1001–1007.

Service, R. F. (2000) Genomics. Structural biology gets a $150 million boost, *Science* **289,** 2254–2255. [News]

Shaw, J. P., Petsko, G. A., and Ringe, D. (1997) Determination of the structure of alanine racemase from *Bacillus stearothermophilus* at 1.9-Å resolution, *Biochemistry* **36,** 1329–1342.

Siomi, H., Matunis, M. J., Michael, W. M., and Dreyfuss, G. (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif, *Nucleic Acids Res.* **21,** 1193–1198.

Song, H. K., Lee, K. N., Kwon, K. S., Yu, M. H., and Suh, S. W. (1995) Crystal structure of an uncleaved alpha 1-antitrypsin reveals the conformation of its inhibitory reactive loop, *FEBS Lett.* **377,** 150–154.

Spezio, M., Wilson, D. B., and Karplus, P. A. (1993) Crystal structure of the catalytic domain of a thermophilic endocellulase, *Biochemistry* **32,** 9906–9916.

Stehle, T., Ahmed, S. A., Claiborne, A., and Schulz, G. E. (1991) Structure of NADH peroxidase from *Streptococcus faecalis* 10C1 refined at 2.16 Å resolution, *J. Mol. Biol.* **221,** 1325–1344.

Story, R. M., Weber, I. T., and Steitz, T. A. (1992) The structure of the *E. coli* recA protein monomer and polymer, *Nature* **355,** 318–325.

Su, X. D., Taddei, N., Stefani, M., Ramponi, G., and Nordlund, P. (1994) The crystal structure of a low-molecular-weight phosphotyrosine protein phosphatase, *Nature* **370,** 575–578.

Sutton, R. B., Davletov, B. A., Berghuis, A. M., Sudhof, T. C., and Sprang, S. R. (1995) Structure of the first C2 domain of synaptotagmin I: A novel Ca^{2+}/phospholipid-binding fold, *Cell* **80,** 929–938.

Tatti, M., Salvioli, R., Ciaffoni, F., Pucci, P., Andolfo, A., Amoresano, A., and Vaccaro, A. M. (1999) Structural and membrane-binding properties of saposin D, *Eur. J. Biochem.* **263,** 486–494.

Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M.

(1999) Protein folds, functions and evolution, *J. Mol. Biol.* **293,** 333–342.

Thornton, J. M., and Sibanda, B. L. (1983) Amino and carboxy-terminal regions in globular proteins, *J. Mol. Biol.* **167,** 443–460.

Uliel, S., Fliess, A., Amir, A., and Unger, R. (1999) A simple algorithm for detecting circular permutations in proteins, *Bioinformatics* **15,** 930–936.

Viguera, A. R., Blanco, F. J., and Serrano, L. (1995) The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics, *J. Mol. Biol.* **247,** 670–681.

Waldrop, G. L., Rayment, I., and Holden, H. M. (1994) Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase, *Biochemistry* **33,** 10249–10256.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.* **1,** 945–951.

Whisstock, J., Skinner, R., and Lesk, A. M. (1998) An atlas of serpin conformations, *Trends Biochem. Sci.* **23,** 63–67.

Wigley, D. B., Gamblin, S. J., Turkenburg, J. P., Dodson, E. J., Piontek, K., Muirhead, H., and Holbrook, J. J. (1992) Structure of a ternary complex of an allosteric lactate dehydrogenase from *Bacillus stearothermophilus* at 2.5 Å resolution, *J. Mol. Biol.* **223,** 317–335.

Wilson, G. G. (1992) Amino acid sequence arrangements of DNA-methyltransferases, *Methods Enzymol.* **216,** 259–279.

Wilson, K. P., Shewchuk, L. M., Brennan, R. G., Otsuka, A. J., and Matthews, B. W. (1992) *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains, *Proc. Natl. Acad. Sci. USA* **89,** 9257–9261.

Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit, *Nature* **407,** 327–239.

Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. (1999) Distribution of protein folds in the three superkingdoms of life, *Genome Res.* **9,** 17–26.

Xing, Y., Guha Thakurta, D., and Draper, D. E. (1997) The RNA binding domain of ribosomal protein L11 is structurally similar to homeodomains, *Nat. Struct. Biol.* **4,** 24–27.

Yamaguchi, H., Kato, H., Hata, Y., Nishioka, T., Kimura, A., Oda, J., and Katsube, Y. (1993) Three-dimensional structure of the glutathione synthetase from *Escherichia coli* B at 2.0 Å resolution, *J. Mol. Biol.* **229,** 1083–1100.

Yuan, S. M., and Clarke, N. D. (1998) A hybrid sequence approach to the Paracelsus challenge, *Proteins* **30,** 136–143.

Yuvaniyama, J., Denu, J. M., Dixon, J. E., and Saper, M. A. (1996) Crystal structure of the dual specificity protein phosphatase VHR, *Science* **272,** 1328–1331.

Zanotti, G., Berni, R., and Monaco, H. L. (1993) Crystal structure of liganded and unliganded forms of bovine plasma retinol-binding protein, *J. Biol. Chem.* **268,** 10728–10738.