New structure — novel fold?

Liisa Holm and Chris Sander

Address: EMBL-EBI, Wellcome Trust Genome Campus, Cambridge
CB10 1SD, UK.

E-mail: holm@embl-ebi.ac.uk
sander@embl-ebi.ac.uk

Electronic identifier: 0969-2126-005-00165

Structure 15 February 1997, 5:165–171

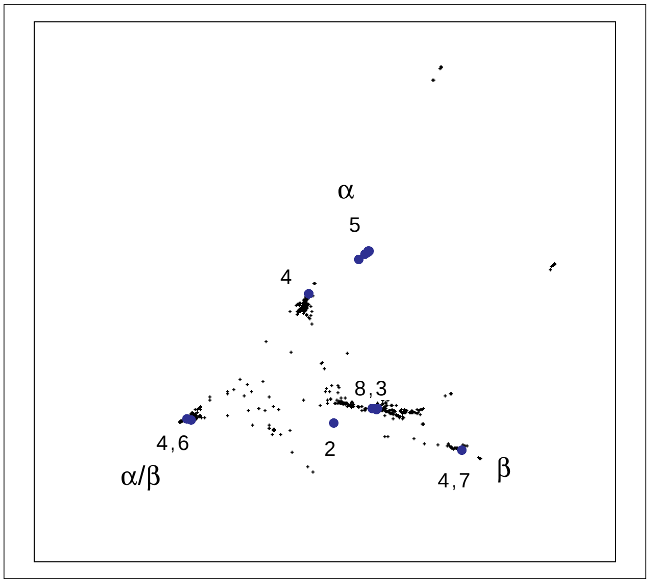
Introduction

The total variety of protein structures is considerably smaller than the variety of protein sequences. This is commonly explained as the result of physical limitations and the evolutionary history of natural proteins. As more structures are solved, observations of the recurrence of similar three-dimensional (3D) chain traces between proteins which are dissimilar in sequence are becoming increasingly frequent. Describing the structural similarity relationships between a new structure and previously known structures has therefore become an essential follow-up of protein structure determination. For example, claiming that a new structure represents a novel fold requires the exclusion of significant similarity with any other known structure [1]. In light of the complexity of comparing 3D shapes, and with the number of protein structures known in atomic detail already exceeding five thousand and doubling every two years, this may seem a daunting task. Fortunately, the solution is provided by a new generation of automatic computer algorithms for protein structure comparison that allow continuous monitoring and classification of the rapidly increasing flow of new structures [2–5]. Here, we briefly review the Dali system for structure classification (accessible for structure comparison in 3D via e-mail to dali@embl-ebi.ac.uk or addressing http://www.embl-ebi.ac.uk/dali). We illustrate how protein structure database searching can lead to evolutionary discoveries or the identification of new types of protein architecture. We also revisit the popular question of what constitutes a fold or fold class and introduce the concepts of minimal functional core and minimal structural core.

Fold space

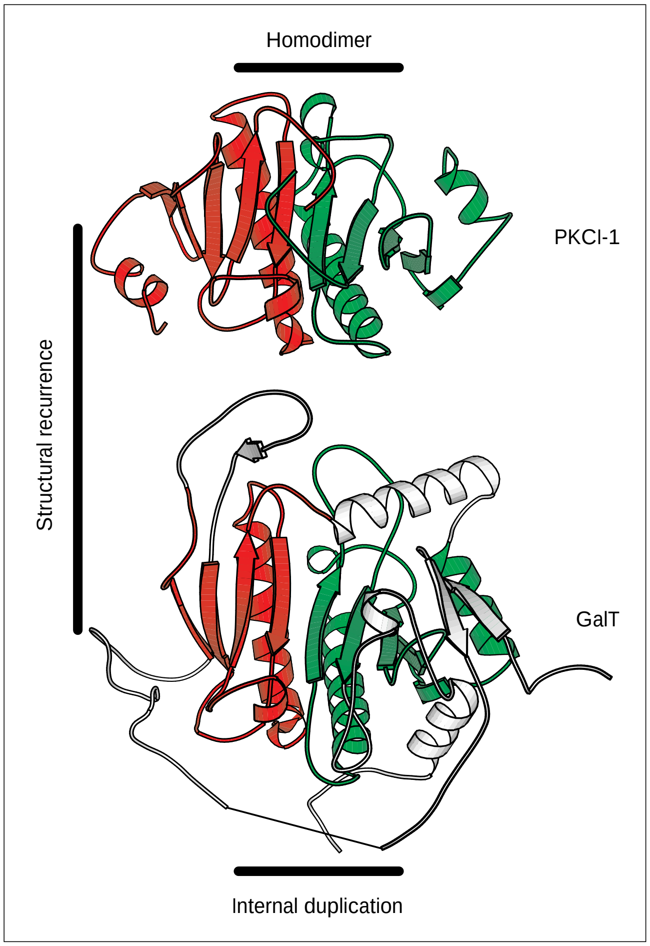
The following two statements appear only subtly different, but exemplify two different conceptual frameworks of fold classification: one states that the structure of luciferase is a triosephosphate isomerase (TIM) barrel, the other, that luciferase resembles triosephosphate isomerase. The former statement goes with traditional and widely used protein structure taxonomies. Loosely speaking, these taxonomies are based on the assertion that there exist abstract descriptors which group together structures related by common evolutionary origin and/or common physical principles which can be identified in single structures by qualitative visual analysis [6–8]. The latter statement goes with a new conceptual framework, that makes no theoretical assumptions but simply describes the structural data by pairwise geometrical relationships in the context of an all-on-all comparison. This approach results in an internally consistent, objective, overall classification of protein structures with interesting implications for protein evolution and folding. Conceptually, the all-on-all table of similarities positions each structure as a point in a high-dimensional fold space (Fig. 1). In the emerging picture of fold space, neighbours at close range represent protein families related through strong functional constraints (e.g. myoglobin and haemoglobin). Neighbours at an intermediate range are related by structure similarity that does not necessarily support similar function (e.g. globin and membrane-insertion domains of bacterial

Figure 1



An image of fold space. Quantification of pairwise structural similarities in an all-on-all comparison allows one to position each domain structure relative to the others in an abstract, high-dimensional fold space. The long-range distribution of different architectures (where α, β and αβ refer to secondary structure composition) is revealed in a projection down onto the plane based on multivariate scaling. Mathematically, an eigenvalue problem is solved; the input is the matrix of all-on-all structural similarities, the axes are the two dominant eigenvectors [2]. Each point is a domain of a representative set of sequence-unique protein structures with less than 25% pairwise sequence identity. Proximity in the plot corresponds to correlated structural neighbourhoods. Large circles denote the position of structures shown in the other figures (identified by figure number).

Figure 2



Structural recurrence with evolutionary implications. Structure database searching can lead to considerable information gain through the unification of remotely related protein families into structurally and functionally conserved superfamilies. Shown here is the striking similarity between a member of the histidine triad (HIT) family of proteins (PKCI-1 [16]; top), and galactose-1-phosphate uridylyltransferase (GalT [17]; bottom). The intermolecular symmetry of the PKCI-1 homodimer is reflected in the approximate intramolecular symmetry between the duplicated subdomains of GalT (red and green regions, additional elements in GalT are in white). The structural alignment reveals conserved functional residues, in spite of a low overall sequence identity of 16%. The discovery of a remote evolutionary relationship permits deductions about the active site of HIT proteins based on the biochemically characterized GalT proteins [11]. (All ribbon diagrams were plotted with Molscript [19].)

toxins), and the long-range distribution of structures corresponds to general architectural types. How are the neighbour relations defined?

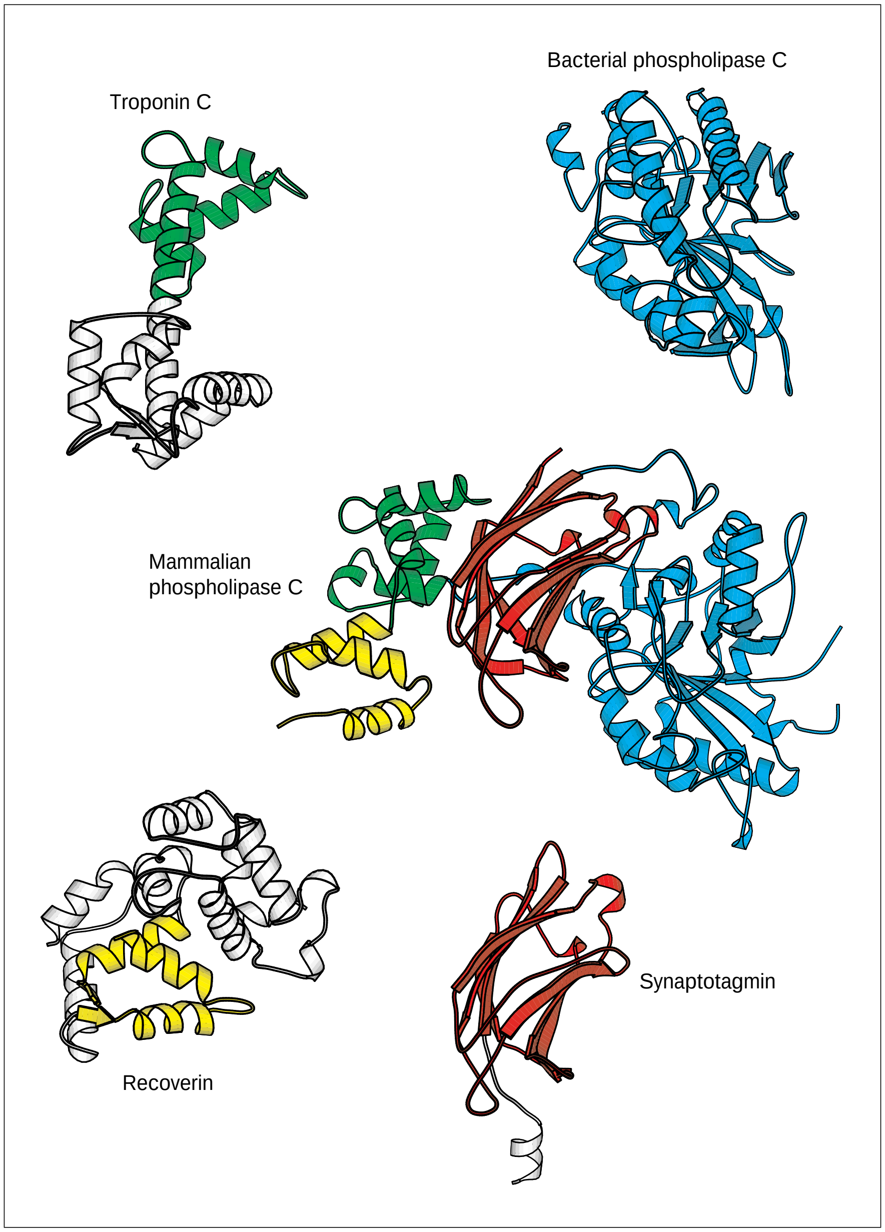
Quantifying neighbour relations

Intuitively, very close structural neighbours have complete and precise structural overlap, while common regions between more distant neighbours have larger variation and cover a smaller part of the structures. The quantitative analysis of neighbour relations involves two steps: structural alignment optimizing a geometrical similarity score (i.e. finding the structurally common parts); and estimating the statistical significance of the match (effectively penalizing for structurally unique parts). The first step is implemented in Dali by comparing matrices of intramolecular distances between residue centres (i.e. Cα atoms) with the goal of identifying substructures that are described by submatrices with small distance deviations. The quantitative geometrical measure of structural similarity is the weighted sum of similarities over equivalent intramolecular distances in the common core. This formulation leads to a combinatorial optimization problem: which residues in protein A paired with which residues in protein B maximize the similarity score? The algorithmic solution to this problem [2] detects recurrent substructures between two proteins. Similar substructures composed of α helices and β sheets recur at all size levels and not necessarily in compact units. The statistical significance of a match depends on the similarity score and the size of the structures being compared. We have empirically calibrated this dependence against a large set of pairwise comparisons, and express the strength of structural similarities in terms of Z scores (i.e. standard deviations above the mean). In practice, we almost invariably find that functionally homologous structures score higher than similar structures that have no functional relationship. In other words, the structural similarity scores appear to capture an evolutionary signal above that due to possible physical convergence of protein architectures.

An example of database searching

The computational tools described above enable structural biologists to fish out the structural neighbours of a newly solved structure from the Protein Data Bank (PDB). Let us look at a concrete example that illustrates the fast developments and potential for biological discoveries in structure databases. The histidine triad (HIT) protein family was first defined based on the presence of a conserved sequence motif. The family is highly conserved in all forms of life which suggests it has an important cellular function, even though it is not known which cellular processes these proteins are involved in. The structure of one member of the HIT family, protein kinase C interacting protein 1 (PKCI-1), was solved last year. On release of the coordinates by the PDB [1], in June 1996, scanning the structure database found no significant neighbours. Indeed, about 15% of new, sequence-unique structures represent structurally unique folds [9]. However, within less than five months a close structural neighbour of PKCI-1, galactose-1-phosphate uridylyltransferase (GalT), was revealed in routine processing of the PDB updates: GalT covers practically all of the HIT dimer structure (Fig. 2). Further analysis of the structurally phased sequence alignments revealed functional signatures leading to family unification of HIT and GalT. This analysis also lead to

Figure 3



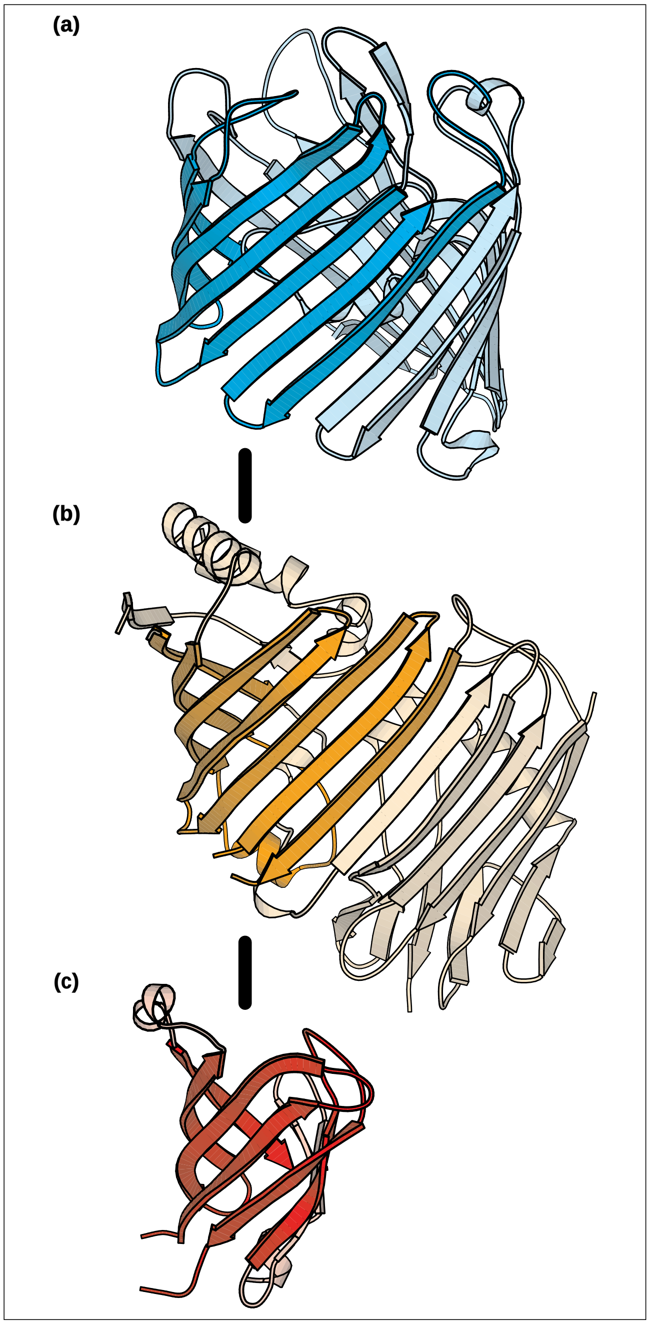
Structural recurrence of physically compact units. Domains are commonly perceived as the basic units of protein folding, function and evolution. Their basic characteristics are a compact globular shape and their recurrence in different structural contexts in many proteins. This principle is illustrated by mammalian phosphoinositide-specific phospholipase C enzymes (PI-PLC) which act as signal transducers, generating two second messengers: inositol-1,4,5-triphosphate and diacylglycerol. The crystal structure of phospholipase Cδ1 (middle) reveals a multidomain protein incorporating recurrent domains (modules) shared by many signalling proteins [20]. The closest structural neighbour for each domain of PI-PLC is shown in matching colour. For example, there are two Ca^{2+}-binding EF-hand domains which also recur in tandem in the proteins recoverin and troponin C. Only the catalytic domain is shared with bacterial phospholipase C.

detailed function predictions for HIT proteins [10]. About 35% of sequence-unique new structures take up membership in existing functionally related superfamilies [9], but surprising discoveries of remote evolutionary relationships, such as that between HIT and GalT, are of course much rarer [11–14]. Curiously, the structure of GalT appeared in print one year before that of the HIT protein (in reverse order of their release by the PDB) but the similarity appears to have gone unnoticed; in fact, quite different vocabulary (β meander [15] versus antiparallel half-barrel [16]) was used to describe the two structures. Database processing was essential in this discovery.

All-on-all comparison: recurrent domains

The detection of structure neighbours is not only useful for comparing single structures against the database, but the automatic methods can also be applied to perform all-on-all comparisons of a complete set of known protein structures. However, there is one technical problem that must be solved before the results of all-on-all structure comparison can be turned into an objective structure classification. Visual inspection of protein structures has shown that most larger proteins are constructed in a modular fashion being comprised of globular, compact, often loosely connected units called domains. This phenomenon is beautifully

Figure 4



Structural recurrence without compactness. Protein structures are composed mainly of α and β secondary structure elements, leading to structural recurrences at all size levels and not necessarily coincident with compact domains. The structures of **(a)** porin (16-stranded β barrel), **(b)** β-lactamase (flat β sheet) and **(c)** streptavidin (eight-stranded β barrel) are structural neighbours because the curls and twists of the β sheet are locally similar in each. However, the overlapping pieces are not compact, and the structures are usually considered as three different folds. The completeness of the overlaps is reflected in the statistical significance of the similarity score that takes the size of the matched units into account. The Dali Z scores (in units of standard deviations above the mean) are 5 and 3 for the pairs (a,b) and (b,c), respectively. (Streptavidin is structurally more closely related to the lipocalins than to β-lactamase.)

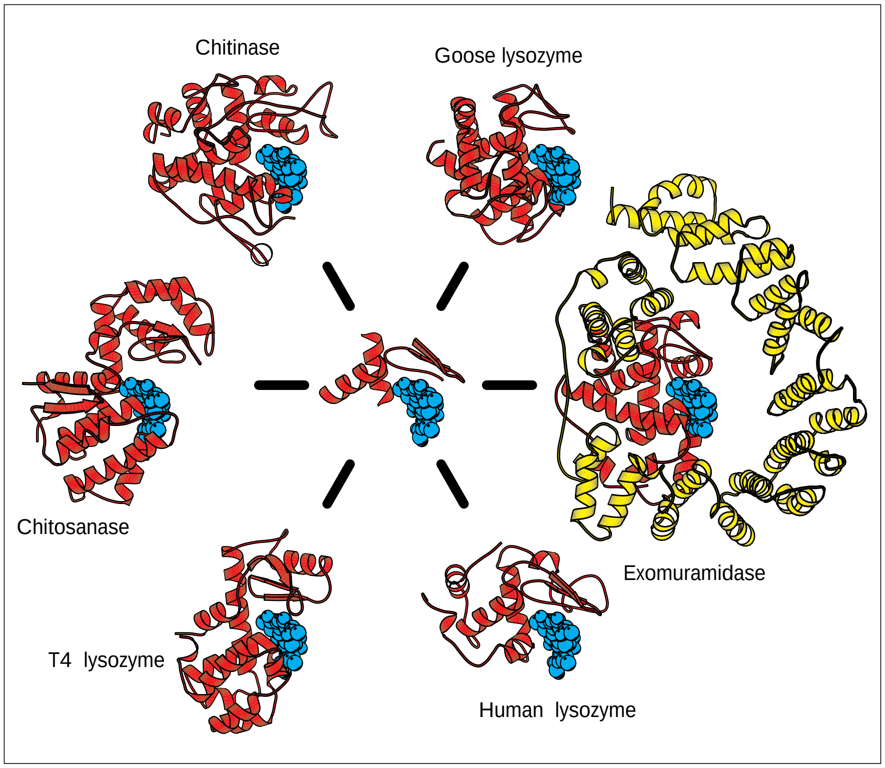
illustrated by mammalian phosphoinositide-specific phospholipase (Fig. 3). The crystal structure of phospholipase C is comprised of four domains, and similar domains have been found to occur in several other proteins with relatively sharp boundaries between the recurrent substructures. Classically, domains are thought of as capable of folding independently, shuffling in evolution between different proteins and often carrying with them a distinct biological function. Because of these important properties, structural classifications are commonly defined at the level of visually identified domains [6–8]. How can protein structures be objectively cut into domains, given the 3D coordinates?

The two most useful criteria for characterizing domains are physical compactness and structural recurrence, and they should be used in combination to resolve conflicts in domain cutting. Conflicts arise because the recurrent substructures detected computationally do not necessarily coincide with visual parsing of 3D structures into physically compact domains (Fig. 4). In other cases, such as the HIT/GalT pair (Fig. 2), human experts' domain definitions may be influenced by evolutionary considerations. It would seem plausible that the HIT protomer folds up before dimerization and, for consistency, GalT should then be viewed as a two-domain structure rather than a single-domain structure as suggested in the original crystallographic report [17]. In our solution to the domain cutting problem, compact units are defined as having minimal surface and maximal interior residue–residue contacts. The compact units are derived by dividing up a protein structure into a tree of putative folding units. An optimal division of a given set of structures maximizes the appearance of recognizably similar substructures (compact units) in different proteins, with recurrence quantified as the sum of Z scores between the selected units. The result is a small set of large substructures in terms of which most protein structures can be parsimoniously described. Structural similarities within this set of domains form the basis for the analysis of fold space (Fig. 1).

Folds and minimal cores

The term 'fold' is commonly used to group together domains with a structural core formed by secondary structure elements in a similar mutual orientation and connectivity (i.e. similar topology in protein jargon). How do these folds relate to structural neighbourhoods detected by the quantitative all-on-all comparison? In general, the structural neighbourhoods in fold space do not form distinct clusters. However, for the purpose of clustering domains into discrete equivalence groups, more or less in agreement with the common notion of fold, we have employed the simple process of average linkage clustering [2], stopping at an empirically chosen cut-off in structural similarity. Let us now ask: what structural principles, if any, are visible in these groups? Although the groups are selected on the

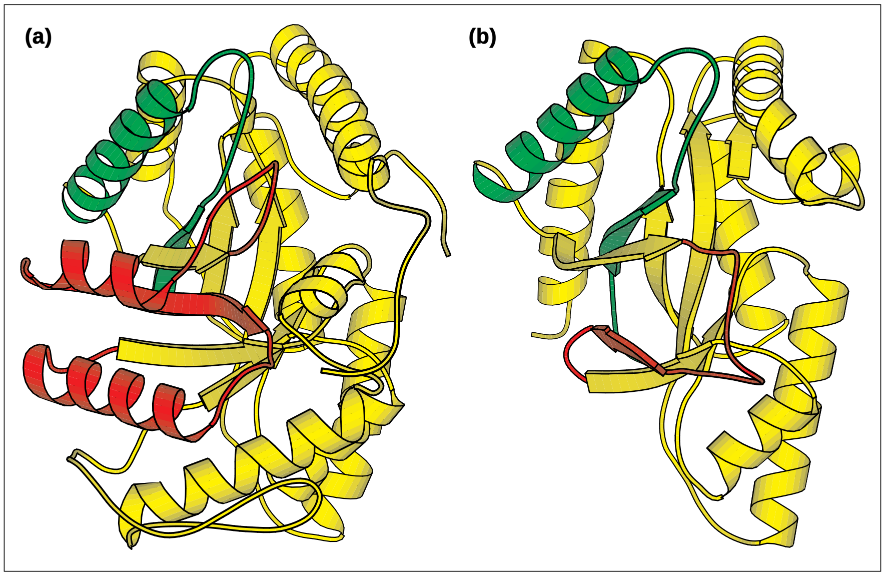
Figure 5



Minimal functional core. In an analysis of the similarity relationships detectable among all known structures, the most clearly visible organizing principle is that of structural conservation due to a continuity of functional constraints in evolution. However, as long as the architectural support of the active site remains, large-scale deviations are quite acceptable. This principle is beautifully illustrated by the six incarnations of the lysozyme fold (red), arranged in a ring around the invariantly conserved substructure. The structurally invariant portion defines the minimal functional core of a divergently related superfamily. The functional core (shown in the centre and cut out from the structure of human lysozyme) consists of a three-stranded β sheet to the right of the substrate, *N*-acetylglucosamine (NAG; blue), and two helices crossing above NAG. To give a visual cue, the NAG substrate from the human lysozyme is shown in structurally equivalent positions in each of the other structures. Overall the structures show considerable variation, and the minimal functional core is surprisingly small; in this case the functional core is not by itself a compact structural unit.

basis of similarity of 3D shapes, the prominent impression is one of rather surprising variability in regions around the central core regions. Structural variability is observed both in groups of structures that are related by functional constraints, and in large groups related by structural similarity but without any apparent biological connection. In a process analogous to analyzing conserved sequence positions in multiple sequence alignments, we can define minimal cores based on the presence of structurally equivalent residues in multiple structure alignments.

Figure 6



Variations around structural themes. **(a)** The enzyme responsible for bioluminescence, bacterial luciferase [21], has the classical $(\beta\alpha)_8$-barrel topology. **(b)** The non-fluorescent flavoprotein, LuxF [22], is evolutionarily related to luciferase. Surprisingly, the structure of LuxF entirely skips one stave of the barrel (variable regions shown in red, N-terminal βα subunits in green). Depending on the point of view, the closed $(\beta\alpha)_8$-barrel topology either can be supported by, or is compatible with, two alternative modes of relaxed sidechain packing [23].

Figure 7

Permission was not granted to reproduce this figure in the electronic version.

The minimal structural core. The common structural features of 65 immunoglobulin-like domains were mapped onto a central member of the cluster. These immunoglobulin-like domains have detectably similar 3D shapes with seven to nine β strands arranged in two β sheets but flexible topological connectivity of the edge strands. Consequently, the structurally invariant core comprises only the two central β ladders. Colour coding changes from blue (regions aligned to the central member in 100% of members of the cluster) to red (less than 50% occupancy). This way of displaying variability within a cluster illustrates how little need be conserved and raises a question about the size scale at which principles of protein structure formation apply. (Figure reproduced from [2], with permission.)

Functional constraints are believed to result in the retention of structural features during evolutionary divergence. Nevertheless, deviations around the common core between remote relatives can involve as much as half of the structure, as in the comparison of glycogen phosphorylase to DNA-glucosyltransferase [11]. A comparison of different lysozyme structures reveals variations in even larger fractions of the proteins (Fig. 5). We define the ‘minimal functional core’ as regions which are invariantly conserved in all members of a superfamily of functionally and structurally related proteins. In lysozymes, the minimal functional core consists of a small β sheet located on one side of the substrate and two α helices at the back of the substrate. It seems plausible that these elements are the minimal set required to sustain substrate binding and catalysis. The minimal functional core of lysozymes does not have the characteristics of a structural unit that would be capable of folding independently. Indeed, each incarnation of ‘the lysozyme fold’ uses additional elements as structural support to the functional core, so that it would be difficult to define the essential elements from a single structure. (Reassuringly, structure comparison detects sufficiently strong similarities due to the active-site region that lysozymes are detected as forming a group distinct from other protein structures.)

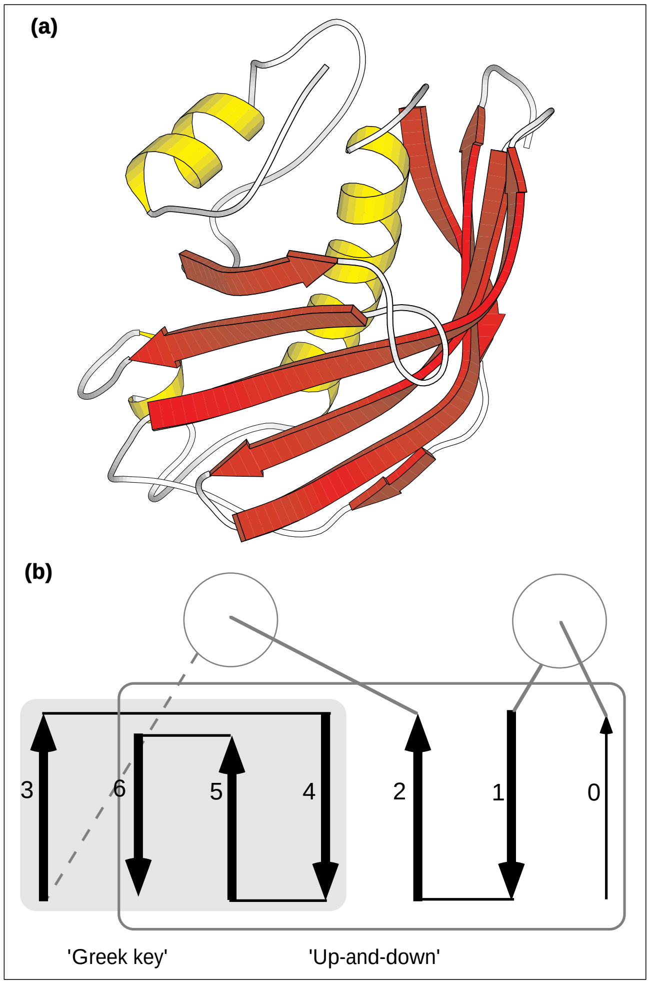
Many classical folds are associated with a particular topological pattern; we will discuss here variations to such patterns. The nicely symmetrical $(\beta\alpha)_8$, alias TIM, barrel topology groups together structures with very different ellipticity of the barrel, and also includes surprising deviations from the consensus strand connectivity (Fig. 6; [17]). A total of 65 immunoglobulin-like domains have been identified and are known to have β-sandwich structures comprised of seven to nine β strands. However, if the minimal structural core of these domains is defined as the positionally invariant elements identified in a multiple structure alignment, then the minimal core only comprises four strands (Fig. 7). The minimal structural core of the immunoglobulin-like domains, interpreted in terms of the protein folding process, is suggestive of a folding nucleus onto which the rest of the structure readily collapses in a sequence-specific manner. This suggestion derives from a conceptual leap stating that elements stably recurring in evolution also stably occur in the process of folding. Experiments testing this hypothesis would be most welcome.

Complementary to defining minimal cores for multimembered groups, one may ask whether folds defined as unique are really singular. The distribution of structures into the different clusters is highly skewed [2] and our operational fold definition leads to a large number of unique folds, one of which is shown in Figure 8. Even though the overall domain fold is unique (with the similarity threshold currently in use), the topology diagram reveals that the domain is composed of very common topological motifs at the subdomain level. If we stay at the domain level of structure classification, we do not yet see any signs of saturation in the number of new domain types [2]. In attempting to describe the spectrum of all possible protein structures, charting fold space at the subdomain level (yet above single secondary structure elements, of which there are two types) may be more productive [18].

Conclusions

Protein sequence and function data, derived from genome projects, and 3D structure data from experimental structural biology will soon yield the complete catalogue of natural proteins. Structure neighbouring is the basis of information services available on the World Wide Web that provide biologists with the tools to search for similarities

Figure 8



Unique fold but ambiguous topological classification. **(a)** The structure of β-hydroxydecanoyl thiol ester dehydrase [24]. The structure is an antiparallel β sheet clasping around a central α helix and defines a unique 'hot dog' fold [24] in the PDB, even though there are structural neighbours with which it has partial overlap. **(b)** Topologically, the structure of β-hydroxydecanoyl thiol ester dehydrase may be described as an up-and-down motif with an insertion (α–β3), or as a Greek key motif with an N-terminal appendix (β strands are shown as arrows, α helices as circles). This illustrates one difficulty in the prediction (enumeration) of admissible but yet unobserved protein structures (i.e. the flexible ways in which common topological motifs can blend into each other).

between a newly determined protein structure and known structures in the PDB. Such searches can lead to the identification of new types of protein architecture or to the discovery of unexpected evolutionary relations between protein families.

Structure classification classically focuses on structural domains and discrete domain folds. Exhaustive all-on-all structure comparison leads to a new view of fold space with less sharp boundaries between different folds. This analysis also raises doubts as to whether the search for grand folding principles in operation at the domain level of structural hierarchy will be fruitful. Instead, the analy-sis leads to the concepts of minimal structural cores and minimal functional cores. We propose that it may be productive to view protein evolution in terms of diver-sions around these minimal cores and that analysis of the minimal cores may help us to understand key events in protein folding.

References

1. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542.
2. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* 273, 595–602.
3. Alexandrov, N.N., Takahashi, K. & Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci.* 3, 866–875.
4. Sali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. *J. Mol. Biol.* 212, 403–428.
5. Orengo, C.A., Flores, T.P., Taylor, W.R. & Thornton, J.M. (1993). Identification and classification of protein fold families. *Protein Eng.* 6, 485–500.
6. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552–558.
7. Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34, 167–339.
8. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.
9. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* 20, 478–480.
10. Holm, L. & Sander, C. (1987). Enzyme HIT. *Trends Biochem. Sci.*, in press.
11. Holm, L. & Sander, C. (1995). Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. *EMBO J.* 14, 1287–1293.
12. Holm, L. & Sander, C. (1995). DNA polymerase β belong to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* 20, 345–347.
13. Holm, L. & Sander, C. (1994). Structural similarity between plant endochitinase and lysozymes from animal and phage: an evolutionary connection. *FEBS Lett.* 340, 129–132.
14. Jabri, E., Carr, M.B., Hausinger, R.P. & Karplus, P.A. (1995). The crystal structure of urease from *Klebsiella aerogenes* at 2.2 Å resolution. *Science* 268, 998–1004.
15. Lima, C.D., Klein, M.G., Weinstein, I.B. & Hendrickson, W.A. (1996). Three-dimensional structure of human protein kinase C interacting protein 1, a member of the HIT family of proteins. *Proc. Natl. Acad. Sci.* 93, 535–5362.
16. Wedekind, J.E., Frey, P.A. & Rayment, I. (1995). Three-dimensional structure of galactose-1-phosphate uridylyltransferase from *Escherichia coli* at 1.8 Å resolution. *Biochemistry* 34, 11049–11061.
17. Lebioda, L. & Stec, B. (1989). The structure of yeast enolase at 2.25 Å resolution. An eightfold $\beta+\alpha$ barrel with a novel $\beta\beta\alpha\alpha(\beta\alpha)_6$ topology. *J. Biol. Chem.* 264, 3685–3693.
18. Efimov, A.V. (1995). Structural similarity between two-layer α/β and β proteins. *J. Mol. Biol.* 245, 402–415.
19. Kraulis, P. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24, 946–950.
20. Essen, L.O., Perisic, O., Cheung, R., Katan, M. & Williams, R.L. (1996). Crystal structure of a mammalian phosphoinositide-specific phospholipase Cδ. *Nature* 380, 595–602.
21. Fisher, A.J., Rauschel, F.M., Baldwin, T.O. & Rayment, I. (1995). Three-dimensional structure of bacterial luciferase from *Vibrio harveyi* at 2.4 Å resolution. *Biochemistry* 34, 6581–6586.
22. Moore, S.A. & James, M.N.G. (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.
23. Lesk, A.M., Bränden, C.I. & Chothia, C. (1989). Structural principles of α/β barrel proteins: the packing of the interior of the sheet. *Proteins* 5, 139–148.
24. Leesong, M., Henderson, B.S., Gillig, J.R., Schwab, J.M. & Smith, J.L. (1996). Structure of a dehydratase-isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. *Structure* 4, 253–264.