Classification of proteins: Available structural space for molecular modeling

ARTICLE *in* METHODS IN MOLECULAR BIOLOGY (CLIFTON, N.J.) · JANUARY 2012

Impact Factor: 1.29 · DOI: 10.1007/978-1-61779-588-6_1 · Source: PubMed

Antonina Andreeva
University of Cambridge

Chapter 1

Classification of Proteins: Available Structural Space for Molecular Modeling

Antonina Andreeva

Abstract

The wealth of available protein structural data provides unprecedented opportunity to study and better understand the underlying principles of protein folding and protein structure evolution. A key to achieving this lies in the ability to analyse these data and to organize them in a coherent classification scheme. Over the past years several protein classifications have been developed that aim to group proteins based on their structural relationships. Some of these classification schemes explore the concept of structural neighbourhood (structural continuum), whereas other utilize the notion of protein evolution and thus provide a discrete rather than continuum view of protein structure space. This chapter presents a strategy for classification of proteins with known three-dimensional structure. Steps in the classification process along with basic definitions are introduced. Examples illustrating some fundamental concepts of protein folding and evolution with a special focus on the exceptions to them are presented.

Key words: Protein domain, Protein motif, Protein repeat, Oligomeric complex, Protein classification, Conformational changes, Chameleon sequences, Fold decay, Fold transitions, Circular permutation

1. Introduction

Over five decades have passed from the time when the first three-dimensional structure of globular protein, myoglobin, was solved (1). Since this pioneering work, the determination of protein structures has seen tremendous increase. The largest repository of structural data, the Protein Data Bank (2), currently holds more than 70,000 protein structures. This wealth of structural data provides unprecedented opportunity to study and better understand the molecular mechanisms of protein function and evolution. A key to achieving this lies in the ability to analyse these data and organize them in a coherent classification scheme.

Andrew J.W. Orry and Ruben Abagyan (eds.), *Homology Modeling: Methods and Protocols*, Methods in Molecular Biology, vol. 857, DOI 10.1007/978-1-61779-588-6_1, © Springer Science+Business Media, LLC 2012

The notion of protein structure classification has emerged from early studies aiming to elucidate the basic principles of protein folding and protein structure evolution. In the late 1970s, Chothia and coworkers pioneered the division of protein structures into four major classes, based on their secondary structure composition and demonstrated that simple geometrical principles govern their mutual arrangement into distinct architectures (3–5). In the early 1980s, in the "Anatomy and Taxonomy of Protein Structure," Jane Richardson has provided the first general classification scheme for protein structures founded on their architecture and topological details (6, 7).

Several protein structure classifications were developed in the 1990s. Liisa Holm and Chris Sander established the Families of Structurally Similar Proteins (FSSP), a fully automatic classification based on structural alignments generated using Dali algorithm (8). FSSP explored the concept of structural neighbourhood and thus creating continuum rather than discrete view of protein structure space. Similarly, the Molecular Modeling DataBase (MMDB) developed at National Center for Biotechnology Information (NCBI) provided a look at the structural neighbourhood but based on the VAST structure comparison algorithm (9). Nearly at the time of the FSSP and MMDB development, the Structural Classification of Proteins (SCOP) database was created at LMB Cambridge by Alexey Murzin, Steven Brenner, Tim Hubbard, and Cyrus Chothia (10). The notion of protein evolution, embodied in SCOP, allowed to create discrete groupings of proteins based not only on their structural similarity but also on their common evolutionary origin. Like in the Linnaean taxonomy, discrete units (domains) were grouped hierarchically on the basis of their common structural and evolutionary relationships. Soon after the release of SCOP, another protein structural classification, Class, Architecture, Topology, Homology (CATH), was developed at UCL London by Orengo et al. (11, 12). Similar to SCOP, the CATH database organized protein domains into hierarchical levels but in contrast to SCOP, used a semi-automatic, rather than manual approach for classification. Each of these classifications remains widely used today and became invaluable resource in many areas of protein structure research.

This chapter discuses a methodology for classification of proteins with known structure. Steps in the classification process along with basic definitions are introduced. Examples illustrating some fundamental concepts of protein folding and evolution, with a special focus on the exceptions to them, are presented. At the end, an overview of the widely used classifications is given.

2. Materials

Automated methods for sequence and structure comparison are indispensible part of protein structure classification process. The most commonly used comparison tools along with the sequence and structural data resources are listed in Table 1. The reader is directed to the references therein for more details about algorithms and descriptions of databases.

3. Units of Protein Classification

Structural similarities between proteins can arise at different levels of protein structure organization. These similarities can be local, comprising only a few secondary structural elements, or global, extending to the entire tertiary or quaternary structure. Each of these structural similarities can indicate biologically relevant relationships between proteins and thus provide important insights into protein function and structure evolution.

This section aims to describe basic units of protein structure classification. Beside protein domain that is most commonly used, additional units of classification, namely motif, repeat, and protein complex are introduced.

3.1. Protein Domain

Domain, as a general feature of protein three-dimensional structure, was primary described by Wetlaufer in terms of regions of polypeptide chain that can enclose in a compact volume and fold autonomously (13). Wetlaufer also introduced the concept of continuous and discontinuous structural regions and proposed an approach for defining domains. Later on, Rossmann based on his observations on dehydrogenases proposed that domains represent genetic units which in the course of evolution have been transferred and combined with other structurally distinct domains to produce functionally different but related proteins (14). These, in essence, conceptually different approaches to delineate domains have evolved in a broad definition of domain as a unit of folding, structure, function, and evolution.

Generally, one or more of the following criteria can be used to define protein domain:

1. A compact, globular region of structure that is semi-independent of the rest of the polypeptide chain (structural domain); this region can consist of one or more segments of the polypeptide chain, the entire polypeptide chain or several polypeptide chains.

Table.1. Databases and tools for protein analysis.

Sequence databases	
Uniprot (*141*)	http://www.uniprot.org
NCBI (*142*)	http://www.ncbi.nlm.nih.gov/
Structure databases	
PDB (*2*)	http://www.pdb.org
Protein structure classifications	
SCOP (*10*)	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH (*12*)	http://www.cathdb.info/
SISYPHUS (*28*)	http://sisyphus.mrc-cpe.cam.ac.uk/
3Dcomplex (*27*)	http://www.3Dcomplex.org
Structural neighbourhoods	
MMDB (*142*)	http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure
FSN (*137*)	http://fatcat.burnham.org/fatcat-cgi/cgi/FSN/fsn.pl
Dali DB (*135*, *143*)	http://ekhidna.biocenter.helsinki.fi/dali/start
COPS (*136*)	http://cops.services.came.sbg.ac.at/
Tools for analysis	
Tools for sequence comparison and similarity searches	
BLAST & PSIBLAST (*85*)	http://www.ncbi.nlm.nih.gov/blast
FASTA3 (*144*)	http://www.ebi.ac.uk/Tools/fasta33
HMMER (*86*)	http://selab.janelia.org/
Tools for structure comparison and similarity searches	
Dali (*143*)	http://ekhidna.biocenter.helsinki.fi/dali_server/
VAST (*145*)	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
SSAP (*146*)	http://www.cathdb.info
FATCAT (*147*)	http://fatcat.burnham.org/
CE (*148*)	http://cl.sdsc.edu/
Mammoth (*149*)	http://ub.cbm.uam.es/mammoth/mult/
Topmatch (*150*)	http://topmatch.services.came.sbg.ac.at/TopMatchFlex.php
TM-align (*151*)	http://zhanglab.ccmb.med.umich.edu/TM-align/
Other resources	
DisProt (*84*)	http://www.disprot.org/
PROSITE (*26*)	http://www.expasy.org/prosite
Consurf (*140*)	http://consurf.tau.ac.il/
Database of Membrane proteins (*152*)	http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
Pratt (*38*)	http://www.ebi.ac.uk/Tools/pratt/index.html
Jalvew (*139*)	http://www.jalview.org/

2. A region of protein that occurs in nature either in isolation or in more than one context of multidomain proteins (evolutionary domain).
3. A region of protein structure that is associated with a particular function (functional domain).

Often when dividing a protein structure into domains not all of these criteria can simultaneously be satisfied. Structural domains, for instance, may not be associated with a particular function or evolutionary domains can consists of two or more structural domains. Similarly, some protein functional domains can contain more than one structural domain. One example of functional domain composed of two structural domains is the structure of D-aminopeptidase DppA that consists of an N-terminal 5-stranded α/β/α domain and a C-terminal 5-stranded β/α domain (Fig. 1) (15). The active site of this enzyme is located in a cleft between the two domains that comprises the most conserved part of the protein. The functionally active protein requires the presence of two domains. None of these domains exists on its own or in combination with other domains and therefore the evolutionary domain spans over the two structural domains.

The selection of criteria used for defining domains should depend on the type of analysis for which domains will be used. For protein structure analysis and structure comparison searches, the domain defined as a structural unit is more appropriate. Some structural domains, however, might not be suitable for sequence

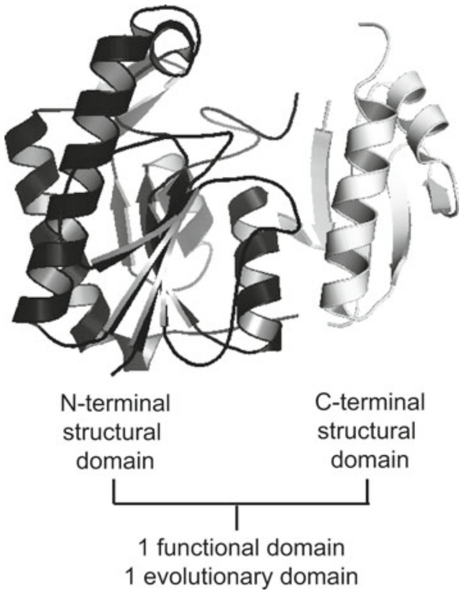

Fig.1. Domains in the structure of D-aminopeptidase DppA (pdb 1hi9).

analysis particularly when the domain consists of two or more discontinuous segments or the domain boundaries disrupt a highly conserved sequence motif that can be crucial for detection of proteins' homologs.

Assignment of novel domains can be done by visual inspection or by using automated methods. Over the past years, several methods for automatic detection of domains have been devised (16–25). Many of them, however, disagree in their domain definitions. The problem with these methods arises from the fact that there is no simple quantitative definition of protein domain. One approach to tackle with this problem is by combining the results of several independent automatic domain definition programmes with visual inspection. This strategy has been implemented by the authors of CATH, in which domains are assigned by using the results of three different methods PUU (18), Domak (20), and DETECTIVE (22) in combination with manual validation. Domains can also be assigned by similarity to already known domains by using either sequence or structure comparison tools.

3.2. Other Units of Classification

Most classifications use the protein domain as classification unit. Within the classification scheme, domains are usually organized hierarchically depending on their structural and evolutionary relationships. The units described here, add extra complexity to the hierarchical presentation of relationships between proteins. They can be classified either separately (as in refs. 26, 27) or as interrelationships within the hierarchical scheme (as in ref. 28).

3.2.1. Protein Motifs

Protein motif is a local, relatively small, contiguous region within a protein polypeptide chain that can be distinguish by a well-defined set of properties (structural and/or functional). There are two types of motifs: sequence and structural. Sequence motif represents a conserved amino acid sequence pattern that is common to a group of proteins. The conservation of the amino acid residues within the motif sometimes can be strict and also may be defined within a certain group, e.g., hydrophobic, polar, or charged. The unique sequence features reflect structural and/or functional constraints and hence sequence motifs usually reside in regions of polypeptide chain that are important for the protein either to perform its tasks or to adopt particular three-dimensional conformation.

Structural motif is regarded as a combination of a few secondary structural elements with a specific geometric arrangement. In contrast to protein domain, it lacks compactness and a well-defined hydrophobic core. Typical examples for structural motifs are Greek-key motif found in β-sandwiches (29), helix-turn-helix (HTH) motif (30), helix-hairpin-helix (HhH) motif (31), etc. Structural motifs were thought that cannot fold independently if they are expressed separately from the rest of the protein. However, recently the HTH motif of engrailed homeodomain was found to fold independently in solution and having essentially the same structure as in the full-length protein (32). This finding allows arguing that some structural motifs may act as a folding template and increase the likelihood for a successful non-homologous recombination (reviewed in ref. 33).

Quite often, but not always a local sequence motif resides in a local structural motif. Some sequence motifs, however, can span over dissimilar structural motifs. For instance, a number of cytochrome c proteins contain a sequence motif defined by C-X2-C-H pattern that binds heme via two invariant Cys residues and coordinates heme iron via conserved His residue. This heme-binding sequence motif spans over regions that have different conformations as shown in Fig. 2. Similarly, (pro)aerolysin and α-hemolysin share a common sequence motif described with [KT]-X2-N-W-X2-T-[DN]-T pattern. Both proteins have globally distinct structures and the sequence motif resides in structurally dissimilar regions.

Similar sequence and structural motifs can be found in structurally distinct proteins. This can result in significant sequence hits between proteins which structures are globally dissimilar. Some of these motifs, however, are of particular interest since they are frequently related to function. Some examples of such motifs are KH motif (34), HTH motif (30), nucleotide-binding motif (35), Ca-binding (DxDxDG) motif (36), P-loop motif (37), etc. The P-loop motif, for instance, is a Gly-rich sequence motif that comprises a flexible loop between a β-strand and an α-helix. This motif is involved in binding of mononucleotides, e.g., ATP, GTP, and directly interacts with one of the phosphate groups. Detection of this motif by sequence analysis tools is relatively straightforward. Several topologically different structures are found to contain the P-loop motif. Another example is the "nucleophile elbow and

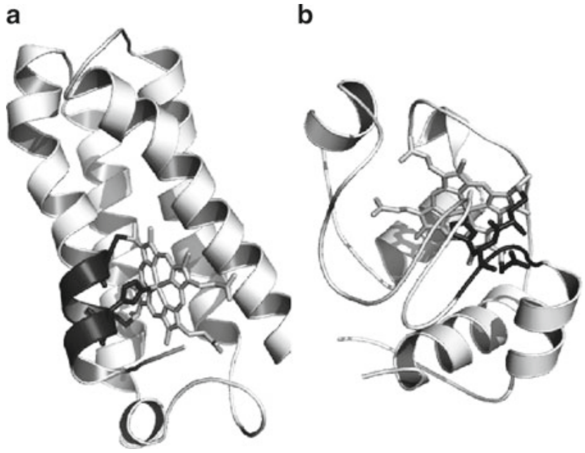

Fig. 2. The structures of (**a**) cytochrome c′ (pdb 1a7v) and (**b**) cytochrome c (pdb 1fhb). The sequence motif common to both proteins is shown in *black*.

oxyanion hole" structural motif that encompasses a discontinuous β/βα motif and harbours the nucleophilic and the oxyanion-hole amino acid residues that constitute the catalytic site in different enzymes. The nucleophile (Ser, Asp, or Cys) is located in a sharp turn between a β-strand and an α-helix, the so-called nucleophile elbow. The oxyanion-hole is usually formed by mainchain NH groups of two Gly, one of which frequently follows the nucleophile. The conserved β/βα structural motif is found in a number of α/β catalytic domains with different β-sheet topologies (Fig. 3).

The presence of common sequence motifs in proteins with dissimilar structures can create challenges for protein structure prediction (see Note 6). Knowledge of the occurrence of these motifs and the structural context in which they are observed is essential for protein modeling.

Sequence motifs can be easily identified within a multiple sequence alignment or by sequence comparisons. One widely used

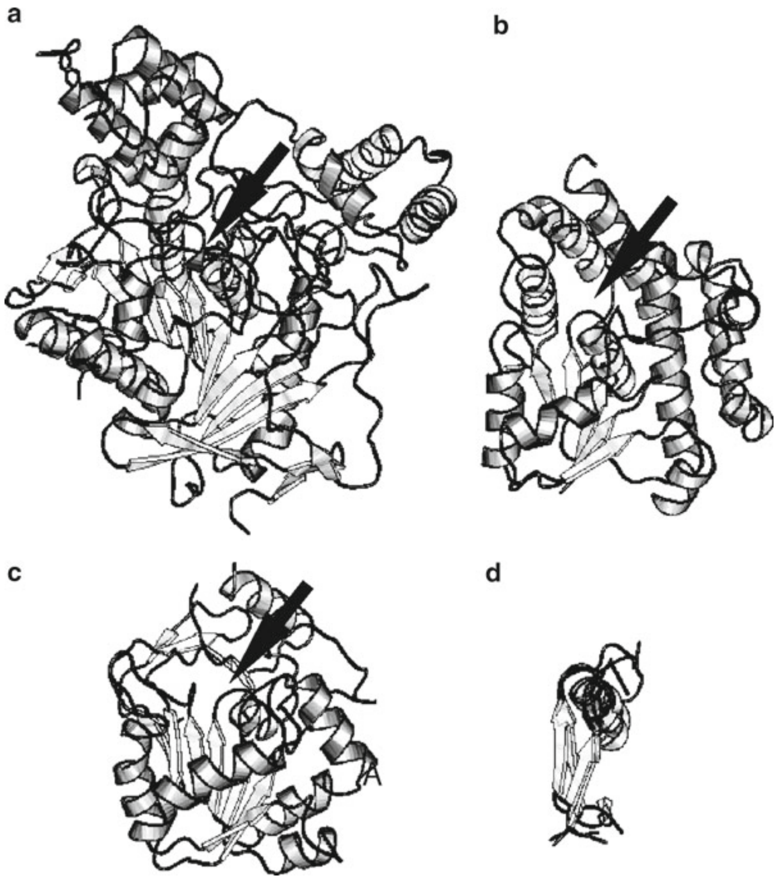

Fig. 3. The structures of (**a**) acetylcholinesterase (pdb 2ack), (**b**) malonyl-CoA:acyl carrier protein transacylase (pdb 1mla), (**c**) aspartyl dipeptidase (pdb 1fye), and (**d**) the "Nucleophile elbow and oxyanion hole" structural motif. *Arrows* indicate the location of the motif in the structures.

resource is PROSITE that contains a collection of protein sequence motifs along with tools for protein sequence analysis and motif detection (26). Programmes are available for automatic generation of sequence patterns (38–41). Detection of structural motifs, particularly in the absence of sequence similarity, is not straightforward. SPASM/RIGOR are programmes that can be used for the detection of small structural motifs (42). Spatial arrangements of side chain and main chain (SPASM) uses a user-defined motif and compares it against a database of protein structures. RIGOR allows searches with entire protein structure using a database of predefined structural motifs.

3.2.2. Protein Repeats

Symmetry and structural duplication are widespread features of natural proteins. A vast number of protein structures with internal symmetry and/or regularly repeating structural units are known to date. These units, also called protein repeats, are usually arranged tandemly in a sequence and/or structure. They exist in multiplicity and thus differ from domains that can exist on their own. Two types of repeats can be distinguish: sequence and structural repeats. Sequence repeat can be defined as any sequence of the same amino acid residue or group of similar amino acid residues repeated in a protein. Frequently, the sequence identity and the number of sequence repeats vary across protein homologs. Structural repeat is regarded as any arrangement of secondary structural elements repeated in a protein structure. The boundaries of sequence repeats frequently correlate with those of structural repeats but in some proteins, e.g., potII family of proteinase inhibitors (43) and WD40-containing proteins (44), the sequence and structural repeats do not coincide.

Protein repeats can fold into compact domains that have a different degree of complexity and shape; and are often symmetrical. Some homologous repetitive structures can bent and coil in different ways so that their global structural similarity can become negligible. These considerable structural variations are usually a result of distinct packing interactions between neighbouring repeats. Protein repeats can form fibrous domains, globular domains, solenoids, and toroids. Repeats in fibrous domains are usually small, comprising only a few residues [collagen, coiled coil (Fig. 4a)]. Some globular proteins contain interlocking repeats that are formed by supersecondary structural elements (Fig. 4b). Solenoids are formed by more simple secondary structural elements such as αα-hairpins [heat, armadillo, and tetratricopeptide repeats (Fig. 4c)], ββ-hairpins and β-arches [β-superhelix (Fig. 4d)], αβ-hairpins [leucine-rich repeat (Fig. 4e)] and fold into open sometimes elongated repetitive structures. Similarly, toroids are built by simple secondary structural elements but in contrast to solenoids form closed structures [αα-toroids (Fig. 4f), β-propellers (Fig. 4g), (βα)8-barrels (Fig. 4h)].

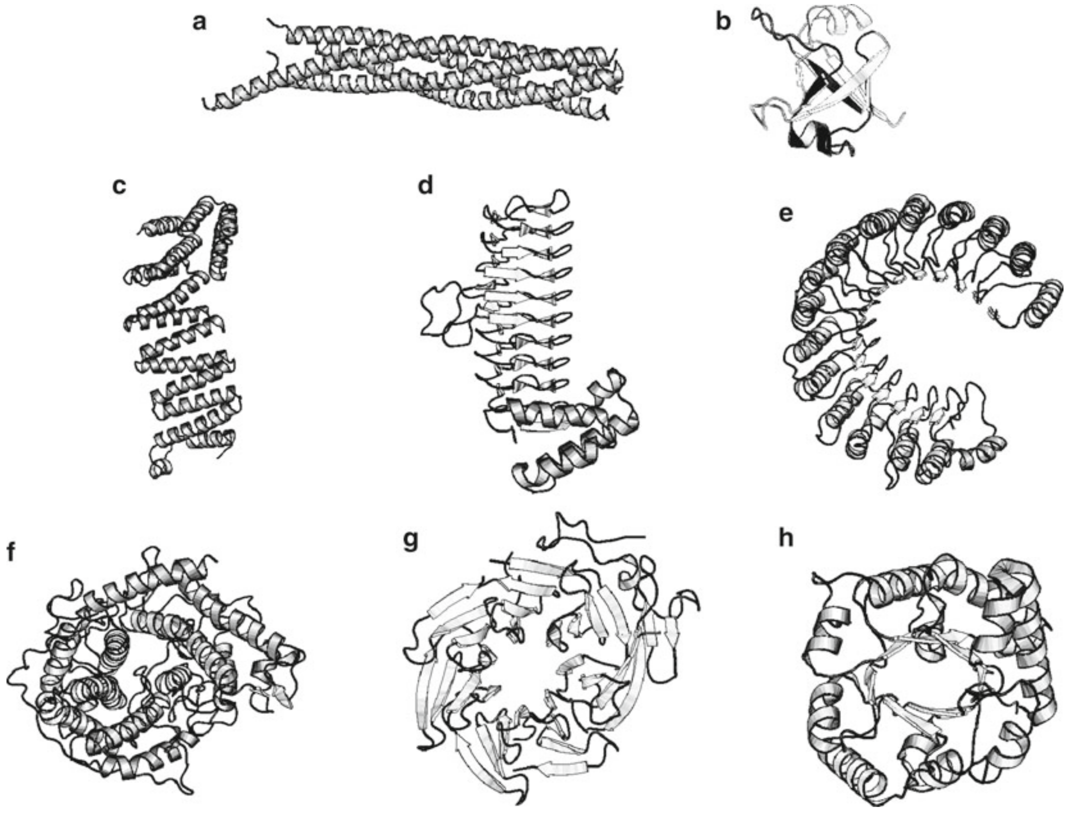

Fig. 4. Representative repetitive structures. (**a**) Coiled coil (pdb 1n7s), (**b**) structural repeats in globular domain (pdb 1cz4), (**c**) α-solenoid (pdb 1qqe), (**d**) β-solenoid (pdb 2jf2), (**e**) $\beta\alpha$-solenoid (pdb 2bnh), (**f**) α-toroid (pdb 1gai), (**g**) β-toroid (pdb 1erj), and (**h**) $\beta\alpha$-toroid (pdb 2jk2).

Methods for detecting repeats are available (45–48). Most of the methods for identification of sequence repeats utilize standard sequence comparison algorithms that are adapted for repeats. They usually perform well when the sequence similarity between repeats is substantial but fail to detect repeats with low sequence similarity or containing large insertions or deletions.

3.2.3. Protein Complexes

Majority of globular and membrane proteins assemble into oligomeric complexes consisting of two or more polypeptide chains. Within these oligomeric complexes two types can be distinguished, homomeric and heteromeric, that are composed of identical and non-identical chains, respectively. A large portion of protein complexes are homomeric with about 50–70% of proteins known to assemble into such structures (49). There are two different types of interfaces in oligomeric complexes: isologous (homologous) and heterologous. Isologous interface is formed by identical surfaces of the two subunits, whereas in heterologous interface, these surfaces are non-identical. Several studies in the past have addressed the structural properties of the oligomeric interfaces such as shape, size, packing, complementarity, etc. (50, 51) but these are beyond the scope of this chapter. Most of oligomeric structures posses symmetry. Dimers and trimers usually adopt cyclic symmetry, whereas dihedral symmetry is more common to tetramers (27, 52). Cubic symmetry is used in protein complexes such as ferritin and viral capsids to enclose vast cavities. Most oligomers adopt either cyclic or dihedral symmetry and only a small fraction of protein complexes have a cubic symmetry (53). Each of the features described above can be used as a criteria to organize and classify protein oligomeric complexes.

4. Classification Based on Protein Types

Proteins fall into four main groups each of which to large extent correlates with characteristic sequence and structural features. Given the striking differences between these groups, their organization and classification will be discussed separately.

4.1. Globular Proteins

Globular proteins are soluble in aqueous solutions. They tend to fold into compact units and their three-dimensional structure reflects their interaction with the solvent. Globular proteins are comparatively easy to analyse and crystallize and therefore, not surprisingly, this group of proteins is the best structurally characterized and comprises the largest fraction of protein structural space available for modeling. Their classification will be described in the next section of this chapter.

4.2. Fibrous Proteins

This group includes a number of structural proteins such as collagen, keratin, elastin, etc., most of which are insoluble. Depending on the secondary structure, fibrous proteins can be subdivided into three groups: triple helix, β-sheet fibres, and α-fibrous proteins. The former group is exemplified by collagen in which each individual polypeptide chain is folded into an extended polyproline type II helix. Three collagen chains coil around a central axis to form a right-handed triple helix. The second group of fibrous proteins tend to form β-sheet structures in which array of extended chains are stacked along the fibril axis. Besides β-keratin and silk proteins, this group includes amyloid fibres. The third group, also known as coiled-coil proteins, is becoming increasingly better understood in terms of sequence and structure. Typically, coiled coils are bundles of two, three, or more helices in which each helix is oriented parallel or antiparallel with respect to the adjacent one. These helices wrap around each other to form a supercoil which is usually left-handed. Although the formation of right-handed coiled-coils is less favourable, these are also observed in nature, e.g. in the structures of tetrabrachion (54), tetramerization domain of VASP

(55), IF regulatory subunitt of F-ATPase (56), and tetramerization domain of MNT repressor (57). Coiled-coil proteins can be homooligomeric or heterooligomeric.

A characteristic feature of the fibrous protein sequences is the presence of repetitive sequence motifs. Collagen, for instance, contains a short Gly-X-Y sequence motif where X is usually Pro and Y is Hyp. Characteristic for the canonical (left-handed, parallel) coiled-coil proteins are heptad repeats denoted as **a-b-c-d-e-f-g**, where **a** and **d** are hydrophobic residues located at the interface of the coiled-coil helices and **e** and **g** are polar residues exposed to the solvent. Nonheptad repeats result in non-canonical coiled-coils that lack left-handness or regular geometry. Right-handed coiled coils, for instance, contain an 11 residue repeat (undecatad repeat). The hydrophobic packing in these proteins substantially differs from the packing of the canonical coiled coils (54). Programmes for analysis of coils are Socket (58) and Twister (59). Socket identifies knobs-into-holes packing in coiled coils, whereas Twister determines the local structural parameters and detects local fluctuations in coiled-coil structures.

The first two subgroups of fibrous proteins are very poorly characterized and only few low resolution structures are available, e.g. the structure of collagen type I that has been recently determined by X-ray fibre diffraction (60). Coiled-coil proteins are difficult to crystallize due to aggregation problems and structures of fragments or relatively short coils are available. Classification of these proteins is usually based on the number of helices, their direction (parallel or antiparallel) and the handedness of the supercoil (left or right).

4.3. Membrane Proteins

Since the first low resolution structure of bacteriorhodopsin was determined by Henderson and Unwin in 1975 (61), much progress has been made in membrane crystallography. Currently, there are more than 200 high-resolution structures of unique membrane proteins. The majority of integral membrane proteins consist of transmembrane α-helices usually organized in bundles. Their topology can be defined on the basis of the number of transmembrane helices and their relative orientation with respect to the plane of the membrane bilayer. The geometry of the side-chains packing at the helix interfaces is reminiscent to knobs-into-holes packing observed in coiled coils (62). The transmembrane helices of proteins involved in proton and electron transport are highly hydrophobic, whereas transporter proteins such as lactose permease (63) have large hydrophilic cavities spanning along the membrane and their helices contain a number of polar and charged residues that are buried in the interior of the transmembrane domain. The transmembrane helices can have different length, different tilt with respect to the bilayer, and different type of distortions, e.g. kinks. Large dynamic changes in the helix orientation and packing interactions or local helix to coil transitions can occur in transmembrane proteins. This intrinsic dynamics of α-helical membrane proteins is a well-documented phenomenon and should be taken into account during structural analysis and classification (64–68).

Another architectural type observed mainly in outer membrane proteins is the β-sheet barrel. All known transmembrane β-barrels form closed structures in which their first strand is hydrogen bonded to the last. The number of strands in the barrel is even and all β-strands are antiparallel. Many barrels contain water filled channels and thus the interior residues are predominantly polar, whereas hydrophobic residues are exposed on the barrel surface. In some proteins, the barrel interior is occupied by additional secondary structural elements or domains. The barrel of autotransporter Nalp, for instance, is filled with an N-terminal helix (69), whereas the barrel of FhuA receptor is plugged by α/β domain (70).

Classification of membrane proteins is primary based on their typical architectural and topological features. Since some membrane proteins have evolved via duplication and fusion, it is important to examine the structure for the presence of internal repeats before it is compared to structures of other proteins. Structure comparison search with a repeat of this kind could reveal a similarity that can be missed if the entire structure is used.

4.4. Intrinsically Unstructured Proteins

Regions of proteins or even entire proteins at native conditions may lack ordered structure but in their functional state they can undergo disorder-to-order transition. These are known as natively unfolded, intrinsically disordered or intrinsically unstructured proteins (IUPs) (71–75). IUPs gained much interest over the last years particularly because they reside in functionally important regions in proteins and comprise a substantial fraction of eukaryotic proteome. Most importantly, these proteins or regions of proteins violate the classical sequence–structure–function paradigm of structural biology, that is, the protein sequence determines a unique 3D structure that in turn determines the proteins’ function.

Intrinsic disorder offers several advantages such as binding of diverse ligands (functional promiscuity), provides a large interaction interface, rapid turnover in the cell, and allows high-specificity coupled with low-affinity interaction. IUPs exist in dynamic ensembles in which the backbone conformation varies over the time and which undergo non-cooperative conformational changes. Typically, the binding to their target (nucleic acid or protein) is accompanied with a shift in the conformational ensemble and a selection of “bound” conformation which is complementary to the binding partner. For example, a number of proteins such as VP16 and p53 contain acidic activation domains that are unstructured in a free state. Upon binding to different target proteins, they undergo disorder-to-order conformational change (76–79). Both electrostatic and hydrophobic interactions are attributed to this phenomenon.

While electrostatics is essential for the mutual attraction to the partner domain, the hydrophobic interactions are essential for the folding of the activation domain (78). Remarkably, although these activation domains bind to structurally distinct protein domains, in all instances they adopt α-helical conformation. Other IUPs, e.g. α-synuclein (80), the C-terminal regulatory domain of p53 (76), exhibit chameleon behaviour and can adopt different conformations (α-helical or β-structures) depending on the environment and the nature of their target domain.

When compared with globular proteins, sequences of IUPs are less conserved. In the absence of strong structural constraints, their sequences have change rapidly during the evolution. In general, IUPs lack the typical patterns of hydrophobic residues observed in globular proteins. Most of them have unusual sequences exhibiting low sequence complexity or high content of charged and low content of hydrophobic residues. This strong bias in their amino acid composition allows successful prediction of protein disorder from the sequence. Several programmes have been developed over the past years (81–83). Structures of quite a few intrinsically disordered regions of proteins bound to their partner proteins have been determined by X-ray crystallography and NMR. None of these, however, have been included in the scope of any of the current protein classifications. A recently developed database, DisProt, provides structural and functional information about disordered proteins (84).

5. Classification of Globular Proteins

The strategy for classifying protein structures, described here, concerns classification of globular proteins but it can be employed for other protein types such as membrane proteins. Steps in the classification procedure of protein domains will be outlined.

Classification of a new protein structure usually begins with analysis of the structure itself. This includes a search for any internal sequence and structural similarity; analysis of the proteins' oligomeric state (biological unit) and domain assignment. Detection of internal similarity can indicate duplication of domains in multidomain proteins or repeats in single domains. The constituent subunits of homooligomeric complexes can exchange equivalent core secondary structural elements (segment-swapping) and domains in these swapped structures should be defined by including corresponding parts of both polypeptide chains. Protein domains are usually consecutive in sequence, but in some proteins one domain can be inserted into another or in a more complex scenario, equivalent structural elements can be swapped between both domains. Because of the ambiguity in identifying domains on the basis of a single structure, it is usually best to start with preliminary domain assignment and tentatively to refine it during the classification process.

Classification of new protein structure depends on its relationship to other proteins with known 3D structure. This relationship can be structural arising from physics and chemistry of proteins favouring particular packing arrangements and topologies or evolutionary due to a descent from a common ancestral protein. Steps of classification aiming identification of these relationships are described below.

5.1. Assignment of Probable Evolutionary Relationships

Protein domains that have evolved from a common ancestor usually share common sequence, structural, and/or functional features. Significant global sequence similarity is considered to be a sufficient evidence for a common ancestry and usually defines close evolutionary relationships. Close evolutionary relationships are detectable with simple BLAST searches (85). More distant (remote) evolutionary relationships can be detected using PSI-BLAST or HMM-profile (86) searches or more sensitive profile–profile approaches such as PRC (87) and COMPASS (88). In the absence of sequence similarity, structural similarity along with commonality in function can also indicate a distant homology. In addition, conserved features such as rare or unusual topological details, conserved packing interactions, common binding/active sites can be used to support a confident conclusion for a common ancestry.

5.2. Assignment of Protein Fold

Assignment of fold is not trivial since there is no single universal definition of protein fold. The term “fold” was originally introduced to outline three major aspects of protein structure: the secondary structural elements of which it is composed, their spatial arrangement and their connectivity. The term “common fold” is used to describe the consensus subset of structural elements shared by a group of proteins. Proteins with the same common fold usually differ in their peripheral structural elements that may have distinct conformation or size. In extreme cases, particularly when homologous proteins are more divergent or have underwent events, such as deletions, insertions, etc (described in the next section), these differences may comprise more than a half of the domain.

Some folds are easy to recognize by eye, e.g. (βα)8-barrel, β-propeller, and many others. For identification of a common fold, it is usually best to perform a structure comparison search against a database of proteins with known structures. Various structure comparison tools can be used to detect structural similarities and some of these are shown in Table 1. Frequently, different methods give different results. For interpretation of the structural similarities is recommended to use the results of several structure comparison algorithms (see Note 4).

5.3. Assignment of Protein Class

Depending on the secondary structure composition, globular protein domains can be divided into four major classes: all-α (predominantly α-helices), all-β (predominantly β-strands), α/β (alternating α-helices and β-strands, and α+β (segregated α-helices and β-strands) (see Note 5). A fifth class includes small proteins with little or no secondary structures. These are usually small proteins that are stabilized either by disulphide bonds or by metal coordination. The division into five classes is adopted by the SCOP classification scheme. Usually, the assignment of all-α and all-β protein classes is straightforward. The borderline between α/β and α+β classes is not always clear. For this reason, the authors of the CATH database, for instance, have merged these two classes into one, namely mixed αβ structures.

6. Dogmas, Principles and Rules, and Their Exceptions

The plethora of structural data accumulated over the past decade revealed numerous examples of atypical structural features and large structural variations that have challenged many longstanding tenets in protein science (33, 89–92). The central dogma of protein folding "one sequence–one structure" is increasingly being challenged as many structural variations are observed in protein families and their individual members. Many exceptions to the topological rules established by earlier protein structure analyses also become apparent. Knowledge of these is essential for both protein structure classification and modeling. Some examples are discussed in this section.

6.1. Sequence–Structure Relationships

In the early 1960s, Anfinsen proposed what he called a "thermodynamic hypothesis" of protein folding to explain the biologically active conformation of protein structure (93, 94). He theorized that the native structure of protein is thermodynamically the most stable under in vivo conditions. Anfinsen postulated that in a given environment, the protein structure is determined by the sum of interatomic interactions and hence by the amino acid sequence. While to a large extent this theory holds true for most proteins, there is a new growing phenomenon of proteins existing in multiple conformational states or adopting conformation that is not at the thermodynamic minimum. In addition, regions of some proteins exhibit chameleon behaviour and can fold into alternative secondary structures.

6.1.1. One Sequence: Many Folds

The most remarkable examples of proteins existing in equilibrium between two entirely different conformational states are Mad2 (95) and lymphotactin (96) (Fig. 5). The transition between the two conformations in both proteins involves a large rear-

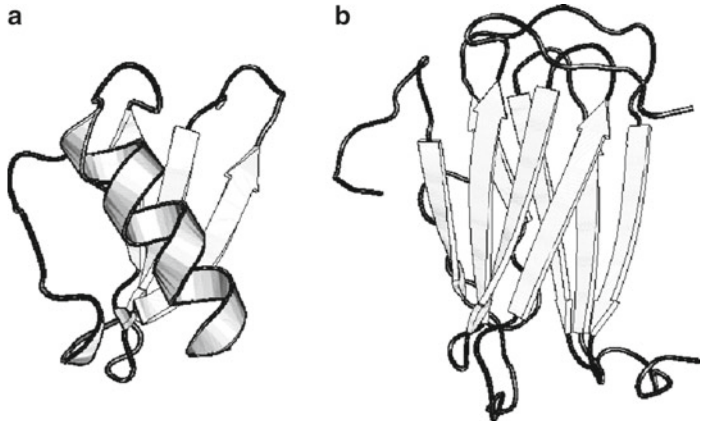

Fig. 5. The structures of two alternative folds of lymphotactin (Ltn10). (**a**) Monomeric Ltn10 (pdb 1j8i) and (**b**) dimeric Ltn10 (pdb 2jp1).

rangement of the hydrogen bonding network and many of the packing interactions.

Several proteins that assume multiple conformational states can adopt biologically active conformation that is not the thermodynamically most stable. This has been shown to play an important role for function. α-Lytic protease and α_1-antitrypsin, for instance, fold into metastable native state, while avoiding the stable but inactive conformation (reviewed in ref. 97). The formation of a metastable native state structure has been described for a number of proteins such as hemaglutinin (98), gp120 and gp41 from HIV (99), protein E from TBEV (100), and some heat shock transcription factors (101).

Depending on the environment some proteins can undergo dramatic conformational changes. The death domain of protein kinase Pelle (Pelle-DD), for example, adopts a six helical bundle characteristic for the death domain family. In the presence of MPD (2-methyl-2,4-pentanediol), the structure of Pelle-DD refolds into a single helix (102) (Fig. 6). Other factors such as pH, salt concentration, temperature are also known to induce conformational transitions. Lymphotactin, for instance, undergoes large structural rearrangement depending on temperature and salt concentration (103). In certain proteins, conformational transitions can be induced by changes in pH, as observed in influenza virus hemagglutinin (98) or pheromone-binding protein (104). Conformational switches can also be a result of experimental design. The design of truncated proteins, in which parts of the polypeptide chain is omitted, may result in dramatic changes of their fold or oligomeric state as observed in p73 (105), MinC (106), Kv7.1 (107), and more recently in human splicing protein PRP8 D4 domain (108).

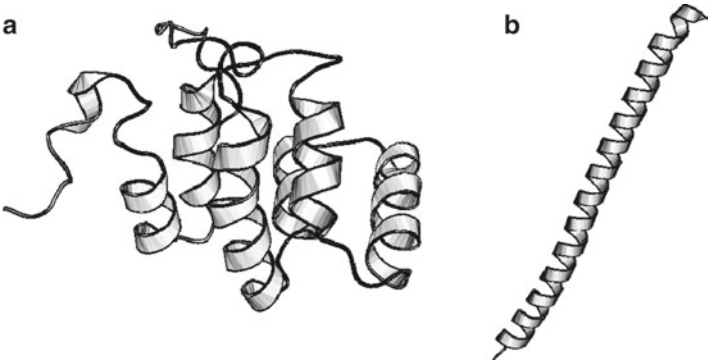

Fig. 6. The death domain of protein kinase Pelle (Pelle-DD) (**a**) solution structure, (**b**) crystal structure in MPD.

6.1.2. Chameleon Sequences

Strings of identical amino acid residues, the so-called chameleon sequences, can adopt alternative secondary structures (α-helix, β-strand, coil). Some chameleon sequences are found in structurally distinct proteins (109, 110). Others are present in individual proteins such as MAD2 (95), matα2 (111), elongation factor Tu (112, 113), p53 (76), Axh (114, 115), Radixin (116, 117), SecA (118), Lekti (119), etc. Most of these chameleon sequences undergo transitions from α-helix to β-strand. The conformational transitions in MAD2 and matα2 are particularly interesting since they are observed under identical conditions. In some proteins, these transitions occur upon oligomer formation. In isolated α-apical domain of thermosome, for instance, the crystal contacts involve a short helical segment resulting in the formation of a four helical bundle between symmetry-related molecules (Fig. 7a) (120, 121). In the closed thermosome, the same region participates in the formation of a β-barrel ring (Fig. 7b). Its conformation is stabilized by interactions provided by the equivalent regions of the adjacent subunits.

6.2. Topological Principles That Determine the Protein Structure

Several topological rules have been established during early analyses aiming to underline the basic principles that govern the protein structure (122–125). One of these postulates that secondary structures, α-helices, and β-sheets, closely pack to enclose hydrophobic core. Others describe preferences such as secondary structures adjacent in sequence are adjacent in structure, right-handedness of connections in β-X-β units, etc. Some topological features as knots and crossing connections were considered improbable and even prohibited. Nowadays, many exceptions of these rules have been found in protein structures. Some of these are shown in Fig. 8.

6.3. Evolution of Protein Structures

A common tenet of protein evolution is that the structure is more conserved than the protein sequence. While for many proteins that's true, steadily growing is the number of evolutionarily related proteins that revealed dramatic changes in their fold. These changes

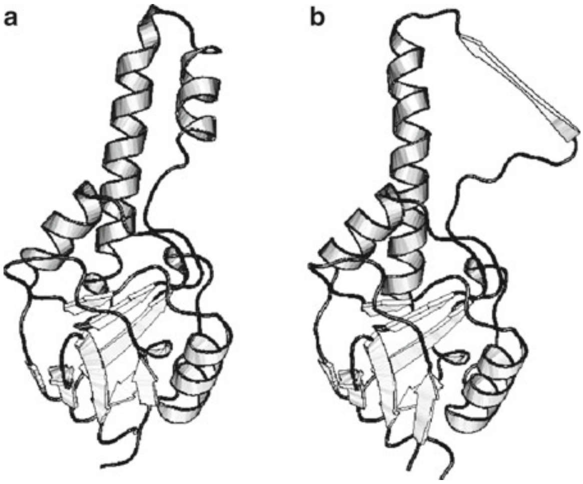

Fig. 7. α-Apical domain of thermosome. (**a**) Structure of isolated domain, (**b**) structure of a subunit in the closed thermosome.

affect not only the peripheral elements but the structural core as well (reviewed in refs. 33, 90, 92). Some examples are given below.

6.3.1. Fold Decay

Fold decay is a deletion event that affects the protein common fold. Fold decay is observed, for instance, in the family B of DNA polymerases. The exonuclease domain of prokaryotic DNA polymerases contains an additional five-stranded β-barrel subdomain with a canonical OB-fold. In the structures of archaeal polymerases, this domain has deletions of different size resulting in the formation of either a three-stranded curved β-sheet or an open β-barrel (Fig. 9).

6.3.2. Fold Transitions

Perhaps the most remarkable example of fold transition is observed in the structures of NusG and RfaH (126). The C-terminal domain of NusG is a SH3-like barrel that contains the so-called KOW motif. Despite the significant sequence similarity between this domain and the C-terminal domain of its homolog RfaH, the latter folds into α-helical domain instead of β-barrel (Fig. 10). Homology modeling of RfaH using the structure of NusG showed that the RfaH sequence can be easily tread on the NusG β-barrel while maintaining the hydrophobic core and avoiding steric clashes (126).

6.3.3. Architecture Transitions

Insertion of additional secondary structures to a common fold core can result in a novel architecture. YaeQ, for example, resembles the restriction endonucleases fold but it contains additional N- and C-terminal β-structures forming a five-stranded β-sheet (127) (Fig. 11). These extra secondary structural elements contribute to the formation of a distinct barrel-like architecture. Despite these

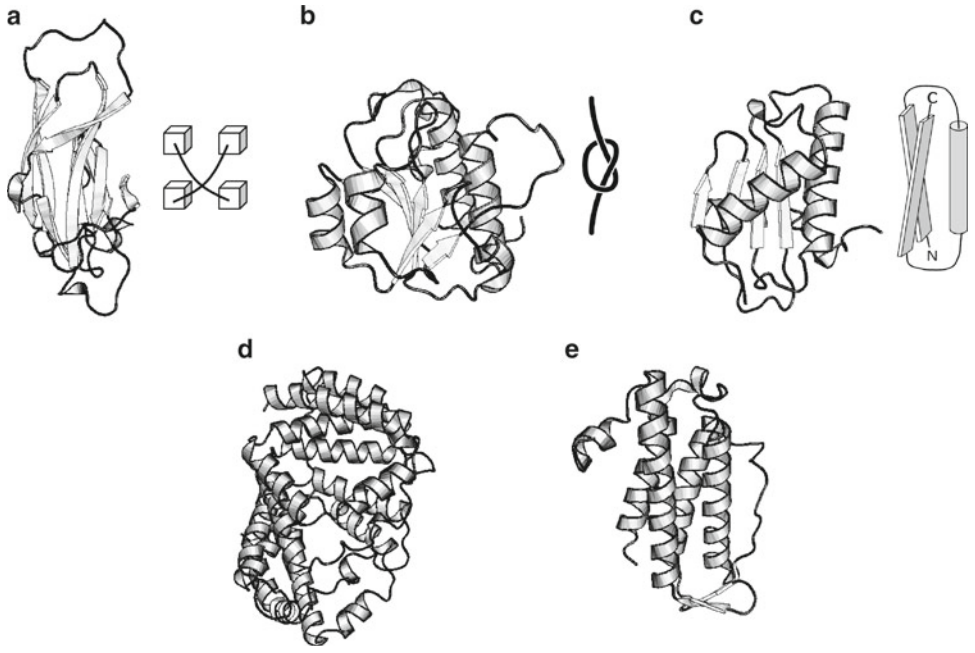

Fig. 8. Examples of exceptions to topological rules. **Rule**: connections between secondary structures neither cross each other nor make knots in the chain. **Exceptions**: (**a**) crossing connections in ecotin (pdb 1ifg) and (**b**) deep trefoil knot in the structure of YibK methyltransferase (pdb 1mxi); **Rule**: connections of β-X-β are right handed. Exception: (**c**) left-handed connection in the structure of Ribonuclease P (pdb 1a6f); **Rule**: the association of secondary structures, α-helices and β-sheets, close pack to form a hydrophobic core. **Exception**: (**d**) the structure of peridinin–chlorophyll–protein (pdb 1ppr) that does not have a core but instead enclosing ligand binding cavity; **Rule**: pieces of secondary structures that are adjacent in sequence are often in contact in three dimensions. **Exception**: (**e**) high contact order structure of representative of DinB-like family (pdb 2f22).

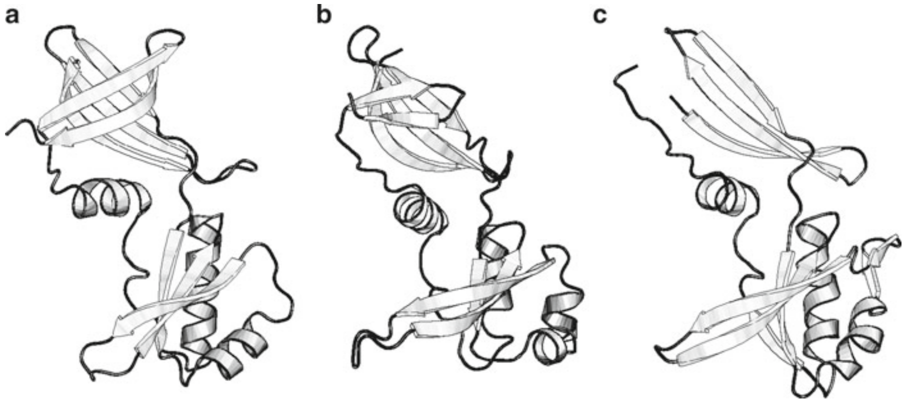

Fig. 9. Fold decay. Structures of exonuclease domains of (**a**) *Escherichia coli* DNA polymerase (pdb 1q8i), (**b**) *Sulfolobus solfataricus* DNA polymerase (pdb 1s5j), (**c**) *Thermococcus gorgonarius* DNA polymerase (pdb 1tgo).

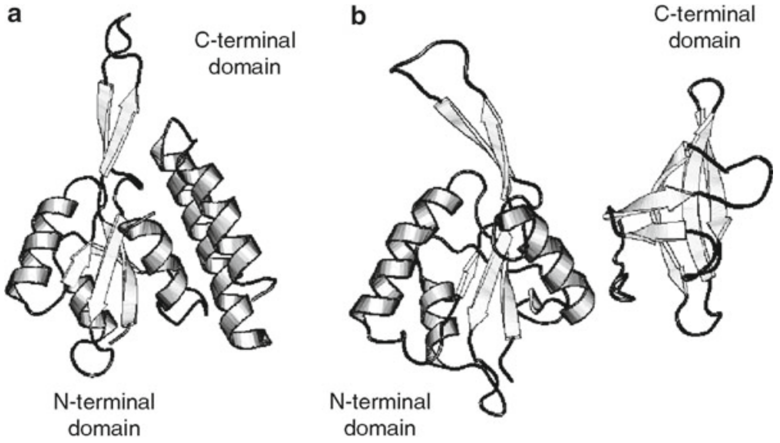

Fig. 10. Fold transition. Structures of (**a**) RfaH and (**b**) NusG.

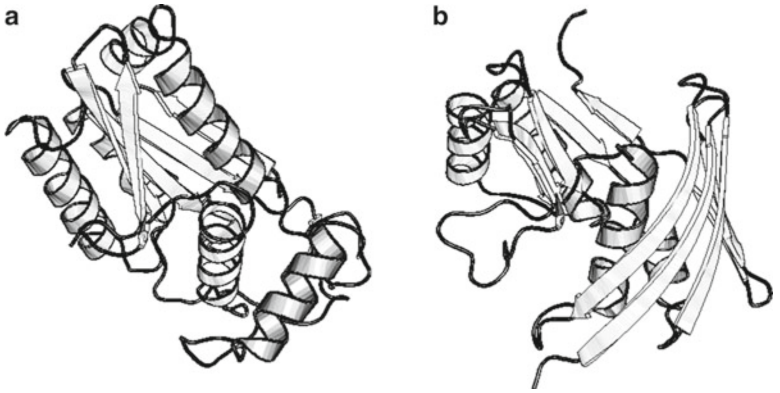

Fig. 11. Architecture transition. Structures of (**a**) restriction endonuclease *Bam*HI (pdb 1bam) and (**b**) *Yae*Q (pdb 2g3w).

differences, residues essential for catalysis in restriction endonucleases, are conserved in the YaeQ structure.

6.3.4. Circular Permutations

Circular permutation can be regarded as a change of the sequential order of the N- and C-terminal parts in protein structures. As such, it does not affect the relative spatial arrangement or packing interactions of the secondary structural elements. Numerous examples of circular permutations are known to date. One example is the structure of phospholipase CΔ C2-domain that has a circularly permuted topology of synaptotagmin I C2-domain (128, 129). The difference between the two topologies is in the first strand of synaptotagmin C2-domain that occupies the same spatial position as the last strand of the phospholipase CΔ C2-domain (Fig. 12).

6.3.5. Strand Flip and Swap

Strand flip is regarded as change of the orientation of the strand with respect to the core elements, whereas strand swap is an internal

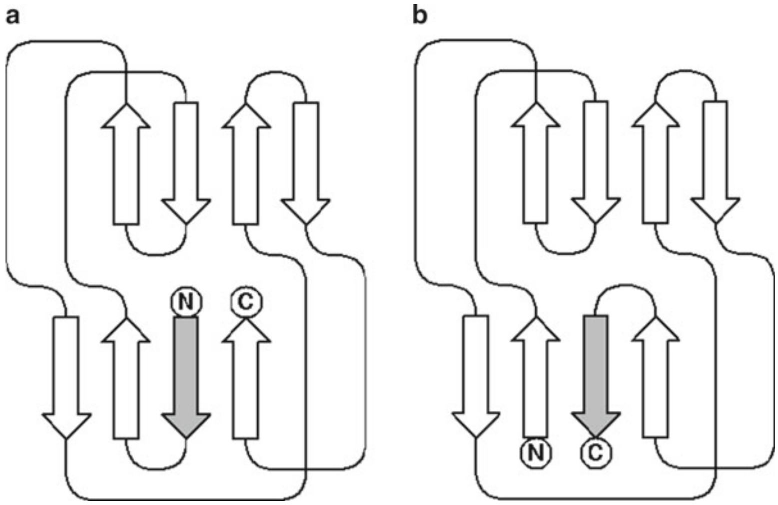

Fig. 12. Circular permutation. Topology diagram of (**a**) synaptotagmin C2-domain, (**b**) phospholipase CΔ C2-domain. Circularly permuted strand is shown in *grey.*

exchange of β-strands that occupy positions with similar environment. One well-known example of strand swap is triabin. The sequence similarity between triabin and nitrophorin is detectable with BLAST. The nitrophorin structure comprises an eight-stranded β-barrel in which all strands are antiparallel. The N-terminal region of triabin differs by swap of a β-hairpin, which results in a parallel arrangement of two pairs of β-strands (Fig. 13).

7. Protein Structure Classification Schemes

Two major manually curated classifications of protein structures are currently available, SCOP (10, 130, 131) and CATH (11, 19, 132). Both classifications have a hierarchical tree-like structure in which protein domains are arranged according to their structural and evolutionary relationships. While these classifications share some common philosophical underpinnings, they differ in several aspects such as domain definitions and classification assignments (133, 134). An overview of these classifications is given below.

A number of other resources that automatically cluster protein structures to build structural neighbourhoods are also available (8, 135–137) (see Table 1). The clustering in these databases depends on the structure comparison method that is employed and algorithm settings that are used. Since comparison methods differ in their results, particularly when the structural similarity between proteins is not significant, the resulting clusters are frequently very different.

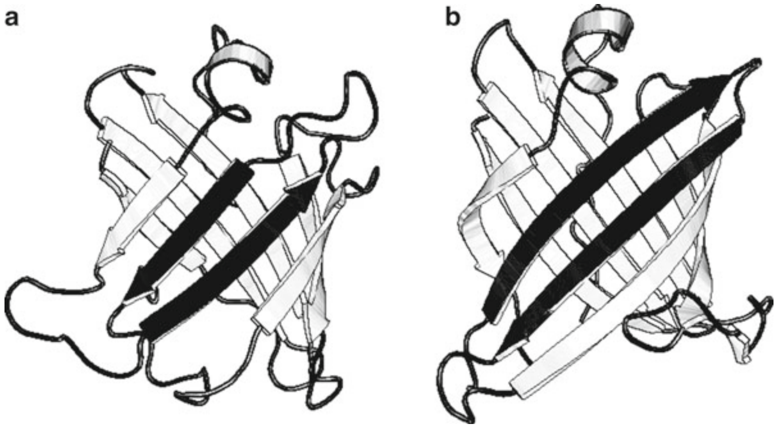

Fig. 13. Strand swap. Structures of (**a**) triabin (pdb 1avg) and (**b**) nitrophorin (pdb 1pee). Swapped β-hairpin is shown in *black.*

7.1. SCOP

SCOP is a database, in which the main focus is to place the proteins in a coherent evolutionary framework, based on their conserved sequence and structural features. It has been created as a hierarchy in which protein domains are arranged in different levels according to their structure and evolution. The SCOP hierarchy comprises the following seven levels: protein **Species**, representing a distinct protein sequence and its naturally occurring or artificially created variants; **Protein**, grouping together similar sequences of essentially the same functions that either originate from different biological species or present different isoforms within the same organism; **Family**, organizing proteins of related sequences but distinct functions; **Superfamily**, bringing together protein families with a common functional and structural features. Near the root of the SCOP hierarchy, structurally similar superfamilies are grouped into **Folds**, which are further arranged into **Classes** based on their secondary structural content.

The classification of proteins in SCOP is a bona fide research. During the classification process, the sequence and structural similarities between proteins are very carefully analysed and interpreted to achieve an optimal prediction of the proteins' evolutionary history. Thus, SCOP is an excellent resource to study the sequence and structural divergence of homologous proteins and the type of structural changes they underwent in the course of evolution.

Structural variations amongst homologous and individual proteins, and the existence of motifs common to structurally distinct proteins add extra complexity and create difficulties in their presentation on the SCOP hierarchy. A comprehensive annotation of these proteins is provided in SISYPHUS, a compendium of

SCOP database (28). The SISYPHUS design conceptually differs from the established classification schemes. In contrast to the latter that are domain-based, the database contains protein structural regions of different size that range from short fragments (motifs or repeats), domains to oligomeric biological units. These protein structural regions are organized in categories that are connected by complex non-hierarchical interrelationships. The relationships between these structural regions are evidenced by multiple alignments and annotated using controlled vocabulary (keywords) and Gene Ontology terms.

7.2. CATH

CATH is a hierarchical protein structure classification in which the protein domains are organized in nine levels. Lower levels of CATH comprise subfamilies of domains that are clustered based on their sequence similarity. Protein domains are merged in **Homologous superfamily** *(H-level)* if they share significant sequence, structure, and/or functional similarity. **Topology** *(T-level)* groups together proteins with a similar arrangement of their secondary structures and topology. Next level, **Architecture** *(A-level)* refers to the overall arrangement of the secondary structures regardless their connectivity. At the root of the hierarchy, **Class** *(C-level)* is defined according to the secondary structure composition. With the exception of A-level that is unique to CATH, the other levels have their equivalent in the SCOP database. The CATH classification protocol uses a highly automated system combined with manual curation (19). Supplementary resource to CATH is CATH-DHS (Dictionary of Homologous Structures) which contains multiple structural alignments, consensus information and functional annotations for proteins grouped at H-level in the classification (138).

7.3. 3D Complex

3D complex is a classification of protein complexes of known three-dimensional structure, representing their fundamental structural features as a graph (27, 52). Proteins are organized in 12 hierarchical levels by using one or more of the following criteria for comparison of the protein complexes: (1) topology of the complex, represented by the number of chains and their pattern of contacts; (2) domain architecture of each constituent chain in the complex according to SCOP classification; (3) number of non-identical chains per domain architecture within each complex; (4) sequence similarity between the constituent chains in the complex; (5) symmetry of the complex. The database allows browsing and analysis of both homomeric and heteromeric complexes and their evolutionary relationships.

8. Notes

1. Because of many structural variations observed amongst homologous proteins and exceptions to rules and definitions, any classification of protein structures will be approximate. The choice of classification scheme should depend on the applications for which it will be used.
2. Every group of related proteins has its own evolutionary history and may underwent events that may not be observed in other proteins. Case by case analysis of protein sequence and structural similarities is, therefore, recommended as it is more powerful way for the detection of protein evolutionary relationships.
3. Given a protein structure, perform sequence analysis of its close homologs with unknown structure. This is best done by search against a sequence database (see Table 1). The sequences of close homologs can be used to generate a multiple sequence alignment and project the sequence conservation on the structure. Best tools to use are Jalview (139) and Consurf (140). Analysis of this type can reveal strictly conserved structural features within the protein family some of which may be related to function.
4. Seek for peculiarities in protein structures such as unusual packing or topological details (knots, left-handed connections, crossing connections). These are characteristic features of folds and can assist in the decision making process during fold assignment.
5. During assignment of protein class, only the core elements of protein domain should be considered. The peripheral elements are usually less conserved and may contain additional structural elements.
6. A significant local sequence similarity between proteins does not necessarily indicate that their structures are globally similar. If a common sequence motif is identified in proteins with known structure, always analyse and compare their structures in order to classify them. If a local sequence match to a protein template structure is found, this not always means that the structure is a suitable template for homology modeling.

References

1. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis, *Nature 181*, 662–666.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res 28*, 235–242.
3. Chothia, C. (1984) Principles that determine the structure of proteins, *Annu. Rev. Biochem. 53*, 537–572.
4. 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.
5. Levitt, M., and Chothia, C. (1976) Structural patterns in globular proteins, *Nature 261*, 552–558.
6. Richardson, J. S. (1977) beta-Sheet topology and the relatedness of proteins, *Nature 268*, 495–500.
7. Richardson, J. S. (1981) The anatomy and taxonomy of protein structure, *Adv. Protein Chem. 34*, 167–339.
8. Holm, L., and Sander, C. (1994) The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res 22*, 3600–3609.
9. Ohkawa, H., Ostell, J., and Bryant, S. (1995) MMDB: an ASN.1 specification for macromolecular structure, *Proc Int Conf Intell Syst Mol Biol 3*, 259–267.
10. 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.
11. Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., and Thornton, J. M. (1999) The CATH Database provides insights into protein structure/function relationships, *Nucleic Acids Res 27*, 275–279.
12. 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.
13. Wetlaufer, D. B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc Natl Acad Sci USA 70*, 697–701.
14. Rossmann, M. G., Moras, D., and Olsen, K. W. (1974) Chemical and biological evolution of nucleotide-binding protein, *Nature 250*, 194–199.
15. Remaut, H., Bompard-Gilles, C., Goffin, C., Frere, J. M., and Van Beeumen, J. (2001) Structure of the Bacillus subtilis D-aminopeptidase DppA reveals a novel self-compartmentalizing protease, *Nat Struct Biol 8*, 674–678.
16. Alden, K., Veretnik, S., and Bourne, P. E. (2010) dConsensus: a tool for displaying domain assignments by multiple structure-based algorithms and for construction of a consensus assignment, *BMC Bioinformatics 11*, 310.
17. Alexandrov, N., and Shindyalov, I. (2003) PDP: protein domain parser, *Bioinformatics 19*, 429–430.
18. Holm, L., and Sander, C. (1994) Parser for protein folding units, *Proteins 19*, 256-268.
19. Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M., and Orengo, C. A. (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures, *PLoS Comput Biol 3*, e232.
20. Siddiqui, A. S., and Barton, G. J. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions, *Protein Sci 4*, 872–884.
21. Sowdhamini, R., and Blundell, T. L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins, *Protein Sci 4*, 506–520.
22. Swindells, M. B. (1995) A procedure for detecting structural domains in proteins, *Protein Sci 4*, 103–112.
23. Taylor, W. R. (1999) Protein structural domain identification, *Protein Eng 12*, 203–216.
24. Veretnik, S., Bourne, P. E., Alexandrov, N. N., and Shindyalov, I. N. (2004) Toward consistent assignment of structural domains in proteins, *J Mol Biol 339*, 647–678.
25. Zhou, H., Xue, B., and Zhou, Y. (2007) DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile, *Protein Sci 16*, 947–955.
26. Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation, *Nucleic Acids Res 38*, D161–166.
27. Levy, E. D., Pereira-Leal, J. B., Chothia, C., and Teichmann, S. A. (2006) 3D complex: a structural classification of protein complexes, *PLoS Comput Biol 2*, e155.

28. Andreeva, A., Prlic, A., Hubbard, T. J., and Murzin, A. G. (2007) SISYPHUS – structural alignments for proteins with non-trivial relationships, *Nucleic Acids Res 35*, D253–259.

29. Hemmingsen, J. M., Gernert, K. M., Richardson, J. S., and Richardson, D. C. (1994) The tyrosine corner: a feature of most Greek key beta-barrel proteins, *Protein Sci 3*, 1927–1937.

30. Brennan, R. G., and Matthews, B. W. (1989) The helix-turn-helix DNA binding motif, *J Biol Chem 264*, 1903–1906.

31. Doherty, A. J., Serpell, L. C., and Ponting, C. P. (1996) The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA, *Nucleic Acids Res 24*, 2488–2497.

32. Religa, T. L., Johnson, C. M., Vu, D. M., Brewer, S. H., Dyer, R. B., and Fersht, A. R. (2007) The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain, *Proc Natl Acad Sci USA 104*, 9272–9277.

33. Andreeva, A., and Murzin, A. G. (2006) Evolution of protein fold in the presence of functional constraints, *Current Opinion in Structural Biology 16*, 399–408.

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

35. Bellamacina, C. R. (1996) The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins, *FASEB J 10*, 1257–1269.

36. Rigden, D. J., and Galperin, M. Y. (2004) The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution, *J Mol Biol 343*, 971–984.

37. 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.

38. Jonassen, I. (1997) Efficient discovery of conserved patterns using a pattern graph, *Comput Appl Biosci 13*, 509–522.

39. Jonassen, I., Collins, J. F., and Higgins, D. G. (1995) Finding flexible patterns in unaligned protein sequences, *Protein Sci 4*, 1587–1595.

40. Rigoutsos, I., and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm, *Bioinformatics 14*, 55–67.

41. Ye, K., Kosters, W. A., and Ijzerman, A. P. (2007) An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences, *Bioinformatics 23*, 687–693.

42. Kleywegt, G. J. (1999) Recognition of spatial motifs in protein structures, *J Mol Biol 285*, 1887–1897.

43. Lee, M. C., Scanlon, M. J., Craik, D. J., and Anderson, M. A. (1999) A novel two-chain proteinase inhibitor generated by circularization of a multidomain precursor protein, *Nat Struct Biol 6*, 526–530.

44. Neer, E. J., Schmidt, C. J., Nambudripad, R., and Smith, T. F. (1994) The ancient regulatory-protein family of WD-repeat proteins, *Nature 371*, 297–300.

45. Murray, K. B., Gorse, D., and Thornton, J. M. (2002) Wavelet transforms for the characterization and detection of repeating motifs, *J Mol Biol 316*, 341–363.

46. Heger, A., and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences, *Proteins 41*, 224–237.

47. Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates, *J Mol Biol 298*, 521–537.

48. Murray, K. B., Taylor, W. R., and Thornton, J. M. (2004) Toward the detection and validation of repeats in protein structure, *Proteins 57*, 365–380.

49. Levy, E. D., Boeri Erba, E., Robinson, C. V., and Teichmann, S. A. (2008) Assembly reflects evolution of protein complexes, *Nature 453*, 1262–1265.

50. Chothia, C., and Janin, J. (1975) Principles of protein-protein recognition, *Nature 256*, 705–708.

51. Jones, S., and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches, *J Mol Biol 272*, 121–132.

52. Levy, E. D. (2007) PiQSi: protein quaternary structure investigation, *Structure 15*, 1364–1367.

53. Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008) Protein-protein interaction and quaternary structure, *Q Rev Biophys 41*, 133–180.

54. Stetefeld, J., Jenny, M., Schulthess, T., Landwehr, R., Engel, J., and Kammerer, R. A. (2000) Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer, *Nat Struct Biol 7*, 772–776.

55. Kuhnel, K., Jarchau, T., Wolf, E., Schlichting, I., Walter, U., Wittinghofer, A., and Strelkov, S. V. (2004) The VASP tetramerization domain is a right-handed coiled coil based on a 15-residue repeat, *Proc Natl Acad Sci USA 101*, 17027–17032.

56. Cabezon, E., Runswick, M. J., Leslie, A. G., and Walker, J. E. (2001) The structure of bovine IF(1), the regulatory subunit of mitochondrial F-ATPase, *EMBO J 20*, 6990–6996.

57. Nooren, I. M., Kaptein, R., Sauer, R. T., and Boelens, R. (1999) The tetramerization

domain of the Mnt repressor consists of two right-handed coiled coils, *Nat Struct Biol 6*, 755–759.

58. Walshaw, J., and Woolfson, D. N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures, *J Mol Biol 307*, 1427–1450.
59. Strelkov, S. V., and Burkhard, P. (2002) Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation, *J Struct Biol 137*, 54–64.
60. Orgel, J. P., Irving, T. C., Miller, A., and Wess, T. J. (2006) Microfibrillar structure of type I collagen in situ, *Proc Natl Acad Sci USA 103*, 9001–9005.
61. Henderson, R., and Unwin, P. N. (1975) Three-dimensional model of purple membrane obtained by electron microscopy, *Nature 257*, 28–32.
62. Walters, R. F., and DeGrado, W. F. (2006) Helix-packing motifs in membrane proteins, *Proc Natl Acad Sci USA 103*, 13658–13663.
63. Guan, L., Mirza, O., Verner, G., Iwata, S., and Kaback, H. R. (2007) Structural determination of wild-type lactose permease, *Proc Natl Acad Sci USA 104*, 15294–15298.
64. Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., and Iwata, S. (2003) Structure and mechanism of the lactose permease of Escherichia coli, *Science 301*, 610–615.
65. Gupta, S., Bavro, V. N., D'Mello, R., Tucker, S. J., Venien-Bryan, C., and Chance, M. R. (2010) Conformational changes during the gating of a potassium channel revealed by structural mass spectrometry, *Structure 18*, 839–846.
66. Toyoshima, C., and Nomura, H. (2002) Structural changes in the calcium pump accompanying the dissociation of calcium, *Nature 418*, 605-611.
67. Olesen, C., Sorensen, T. L., Nielsen, R. C., Moller, J. V., and Nissen, P. (2004) Dephosphorylation of the calcium pump coupled to counterion occlusion, *Science 306*, 2251–2255.
68. Huang, Y., Lemieux, M. J., Song, J., Auer, M., and Wang, D. N. (2003) Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli, *Science 301*, 616–620.
69. Oomen, C. J., van Ulsen, P., van Gelder, P., Feijen, M., Tommassen, J., and Gros, P. (2004) Structure of the translocator domain of a bacterial autotransporter, *EMBO J 23*, 1257–1266.
70. Locher, K. P., Rees, B., Koebnik, R., Mitschler, A., Moulinier, L., Rosenbusch, J. P., and Moras, D. (1998) Transmembrane signaling across the ligand-gated FhuA receptor: crystal structures of free and ferrichrome-bound states reveal allosteric changes, *Cell 95*, 771–778.
71. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat Rev Mol Cell Biol 6*, 197–208.
72. Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008) Function and structure of inherently disordered proteins, *Curr Opin Struct Biol 18*, 756–764.
73. Uversky, V. N., and Dunker, A. K. (2010) Understanding protein non-folding, *Biochim Biophys Acta 1804*, 1231–1264.
74. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics, *Protein Sci 11*, 739–756.
75. Tompa, P. (2002) Intrinsically unstructured proteins, *Trends Biochem Sci 27*, 527–533.
76. Joerger, A. C., and Fersht, A. R. (2010) The tumor suppressor p53: from structures to drug discovery, *Cold Spring Harb Perspect Biol 2*, a000919.
77. Rajagopalan, S., Andreeva, A., Rutherford, T. J., and Fersht, A. R. (2010) Mapping the physical and functional interactions between the tumor suppressors p53 and BRCA2, *Proc Natl Acad Sci USA 107*, 8587–8592.
78. Rajagopalan, S., Andreeva, A., Teufel, D. P., Freund, S. M., and Fersht, A. R. (2009) Interaction between the transactivation domain of p53 and PC4 exemplifies acidic activation domains as single-stranded DNA mimics, *J Biol Chem 284*, 21728–21737.
79. Jonker, H. R., Wechselberger, R. W., Boelens, R., Folkers, G. E., and Kaptein, R. (2005) Structural properties of the promiscuous VP16 activation domain, *Biochemistry 44*, 827–839.
80. Uversky, V. N. (2003) A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders, *J Biomol Struct Dyn 21*, 211–234.
81. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics, *Structure 11*, 1453–1459.
82. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38–48.
83. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004)

Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J Mol Biol 337*, 635–645.

84. Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., and Dunker, A. K. (2007) DisProt: the Database of Disordered Proteins, *Nucleic Acids Res 35*, D786–793.

85. 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.

86. Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure, *BMC Bioinformatics 11*, 431.

87. Madera, M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models, *Bioinformatics 24*, 2630–2631.

88. Sadreyev, R. I., Tang, M., Kim, B. H., and Grishin, N. V. (2009) COMPASS server for homology detection: improved statistical accuracy, speed and functionality, *Nucleic Acids Res 37*, W90–94.

89. Andreeva, A., Prlic, A., Hubbard, T. J., and Murzin, A. G. (2007) SISYPHUS – structural alignments for proteins with non-trivial relationships, *Nucleic Acids Res. 35*, D253–259.

90. Grishin, N. V. (2001) Fold change in evolution of protein structures, *J Struct Biol 134*, 167–185.

91. Kinch, L. N., and Grishin, N. V. (2002) Evolution of protein structures and functions, *Curr Opin Struct Biol 12*, 400–408.

92. Alva, V., Koretke, K. K., Coles, M., and Lupas, A. N. (2008) Cradle-loop barrels and the concept of metafolds in protein classification by natural descent, *Curr Opin Struct Biol 18*, 358–365.

93. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains, *Science 181*, 223–230.

94. Anfinsen, C. B., Haber, E., Sela, M., and White, F. H., Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, *Proc Natl Acad Sci USA 47*, 1309–1314.

95. Luo, X., Tang, Z., Xia, G., Wassmann, K., Matsumoto, T., Rizo, J., and Yu, H. (2004) The Mad2 spindle checkpoint protein has two distinct natively folded states, *Nat Struct Mol Biol 11*, 338–345.

96. Tuinstra, R. L., Peterson, F. C., Kutlesa, S., Elgin, E. S., Kron, M. A., and Volkman, B. F. (2008) Interconversion between two unrelated protein folds in the lymphotactin native state, *Proc Natl Acad Sci USA 105*, 5057–5062.

97. Cabrita, L. D., and Bottomley, S. P. (2004) How do proteins avoid becoming too stable? Biophysical studies into metastable proteins, *Eur Biophys J 33*, 83–88.

98. Bullough, P. A., Hughson, F. M., Skehel, J. J., and Wiley, D. C. (1994) Structure of influenza haemagglutinin at the pH of membrane fusion, *Nature 371*, 37–43.

99. Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. (1997) Core structure of gp41 from the HIV envelope glycoprotein, *Cell 89*, 263–273.

100. Stiasny, K., Allison, S. L., Mandl, C. W., and Heinz, F. X. (2001) Role of metastability and acidic pH in membrane fusion by tick-borne encephalitis virus, *J Virol 75*, 7392–7398.

101. Orosz, A., Wisniewski, J., and Wu, C. (1996) Regulation of Drosophila heat shock factor trimerization: global sequence requirements and independence of nuclear localization, *Mol Cell Biol 16*, 7018–7030.

102. Xiao, T., Gardner, K. H., and Sprang, S. R. (2002) Cosolvent-induced transformation of a death domain tertiary structure, *Proc Natl Acad Sci USA 99*, 11151–11156.

103. Kuloglu, E. S., McCaslin, D. R., Markley, J. L., and Volkman, B. F. (2002) Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions, *J Biol Chem 277*, 17863–17870.

104. Zubkov, S., Gronenborn, A. M., Byeon, I. J., and Mohanty, S. (2005) Structural consequences of the pH-induced conformational switch in A. polyphemus pheromone-binding protein: mechanisms of ligand release, *J Mol Biol 354*, 1081–1090.

105. Joerger, A. C., Rajagopalan, S., Natan, E., Veprintsev, D. B., Robinson, C. V., and Fersht, A. R. (2009) Structural evolution of p53, p63, and p73: implication for heterotetramer formation, *Proc Natl Acad Sci USA 106*, 17705–17710.

106. Cordell, S. C., Anderson, R. E., and Lowe, J. (2001) Crystal structure of the bacterial cell division inhibitor MinC, *EMBO J 20*, 2454–2461.

107. Xu, Q., and Minor, D. L., Jr. (2009) Crystal structure of a trimeric form of the K(V)7.1 (KCNQ1) A-domain tail coiled-coil reveals structural plasticity and context dependent changes in a putative coiled-coil trimerization motif, *Protein Sci 18*, 2100–2114.

108. Schellenberg, M. J., Ritchie, D. B., Wu, T., Markin, C. J., Spyracopoulos, L., and Macmillan,

A. M. (2010) Context-Dependent Remodeling of Structure in Two Large Protein Fragments, *J Mol Biol 402*, 720–730.

109. Guo, J. T., Jaromczyk, J. W., and Xu, Y. (2007) Analysis of chameleon sequences and their implications in biological processes, *Proteins 67*, 548–558.

110. Mezei, M. (1998) Chameleon sequences in the PDB, *Protein Eng 11*, 411–414.

111. Tan, S., and Richmond, T. J. (1998) Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex, *Nature 391*, 660–666.

112. Abel, K., Yoder, M. D., Hilgenfeld, R., and Jurnak, F. (1996) An alpha to beta conformational switch in EF-Tu, *Structure 4*, 1153–1159.

113. Polekhina, G., Thirup, S., Kjeldgaard, M., Nissen, P., Lippmann, C., and Nyborg, J. (1996) Helix unwinding in the effector region of elongation factor EF-Tu-GDP, *Structure 4*, 1141–1151.

114. Chen, Y. W., Allen, M. D., Veprintsev, D. B., Lowe, J., and Bycroft, M. (2004) The structure of the AXH domain of spinocerebellar ataxin-1, *J Biol Chem 279*, 3758–3765.

115. de Chiara, C., Menon, R. P., Adinolfi, S., de Boer, J., Ktistaki, E., Kelly, G., Calder, L., Kioussis, D., and Pastore, A. (2005) The AXH domain adopts alternative folds the solution structure of HBP1 AXH, *Structure 13*, 743–753.

116. Hamada, K., Shimizu, T., Yonemura, S., Tsukita, S., and Hakoshima, T. (2003) Structural basis of adhesion-molecule recognition by ERM proteins revealed by the crystal structure of the radixin-ICAM-2 complex, *EMBO J 22*, 502–514.

117. Kitano, K., Yusa, F., and Hakoshima, T. (2006) Structure of dimerized radixin FERM domain suggests a novel masking motif in C-terminal residues 295-304, *Acta Crystallogr Sect F Struct Biol Cryst Commun 62*, 340–345.

118. Zimmer, J., Li, W., and Rapoport, T. A. (2006) A novel dimer interface and conformational changes revealed by an X-ray structure of B. subtilis SecA, *J Mol Biol 364*, 259–265.

119. Tidow, H., Lauber, T., Vitzithum, K., Sommerhoff, C. P., Rosch, P., and Marx, U. C. (2004) The solution structure of a chimeric LEKTI domain reveals a chameleon sequence, *Biochemistry 43*, 11238–11247.

120. Ditzel, L., Lowe, J., Stock, D., Stetter, K. O., Huber, H., Huber, R., and Steinbacher, S. (1998) Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT, *Cell 93*, 125–138.

121. Klumpp, M., Baumeister, W., and Essen, L. O. (1997) Structure of the substrate binding domain of the thermosome, an archaeal group II chaperonin, *Cell 91*, 263–270.

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

123. Chothia, C., and Finkelstein, A. V. (1990) The classification and origins of protein folding patterns, *Annu Rev Biochem 59*, 1007–1039.

124. Sternberg, M. J., and Thornton, J. M. (1976) On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit, *J Mol Biol 105*, 367–382.

125. Sternberg, M. J., and Thornton, J. M. (1977) On the conformation of proteins: the handedness of the connection between parallel beta-strands, *J Mol Biol 110*, 269–283.

126. Belogurov, G. A., Vassylyeva, M. N., Svetlov, V., Klyuyev, S., Grishin, N. V., Vassylyev, D. G., and Artsimovitch, I. (2007) Structural basis for converting a general transcription factor into an operon-specific virulence regulator, *Mol Cell 26*, 117–129.

127. Guzzo, C. R., Nagem, R. A., Barbosa, J. A., and Farah, C. S. (2007) Structure of Xanthomonas axonopodis pv. citri YaeQ reveals a new compact protein fold built around a variation of the PD-(D/E)XK nuclease motif, *Proteins 69*, 644–651.

128. 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.

129. 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 Ca2+/phospholipid-binding fold, *Cell 80*, 929–938.

130. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data, *Nucleic Acids Res 32*, D226–229.

131. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: new developments, *Nucleic Acids Res 36*, D419–425.

132. Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A., Garratt, R., Thornton, J., and Orengo, C. (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space, *Structure 17*, 1051–1062.

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

134. Day, R., Beck, D. A., Armen, R. S., and Daggett, V. (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary, *Protein Sci 12*, 2150–2160.

135. Holm, L., and Park, J. (2000) DaliLite workbench for protein structure comparison, *Bioinformatics 16*, 566–567.

136. Suhrer, S. J., Wiederstein, M., Gruber, M., and Sippl, M. J. (2009) COPS – a novel workbench for explorations in fold space, *Nucleic Acids Res 37*, W539–544.

137. Li, Z., Ye, Y., and Godzik, A. (2006) Flexible Structural Neighborhood – a database of protein structural similarities and alignments, *Nucleic Acids Res 34*, D277–280.

138. Bray, J. E., Todd, A. E., Pearl, F. M., Thornton, J. M., and Orengo, C. A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues, *Protein Eng 13*, 153–165.

139. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench, *Bioinformatics 25*, 1189–1191.

140. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Res 38 Suppl*, W529–533.

141. (2010) The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res 38*, D142–148.

142. Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res 37*, D5–15.

143. Holm, L., and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D, *Nucleic Acids Res 38 Suppl*, W545–549.

144. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci USA 85*, 2444–2448.

145. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996) Surprising similarities in structure comparison, *Curr Opin Struct Biol 6*, 377–385.

146. Orengo, C. A., and Taylor, W. R. (1996) SSAP: sequential structure alignment program for protein structure comparison, *Methods Enzymol 266*, 617–635.

147. Ye, Y., and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists, *Bioinformatics 19 Suppl 2*, ii246–255.

148. Shindyalov, I. N., and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng 11*, 739–747.

149. Ortiz, A. R., Strauss, C. E., and Olmea, O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison, *Protein Sci 11*, 2606–2621.

150. Sippl, M. J., and Wiederstein, M. (2008) A note on difficult structure alignment problems, *Bioinformatics 24*, 426–427.

151. Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score, *Nucleic Acids Res 33*, 2302–2309.

152. Jayasinghe, S., Hristova, K., and White, S. H. (2001) MPtopo: A database of membrane protein topology, *Protein Sci 10*, 455–458.