1

Protein Structure and Conformational Dynamics

Volkhard Helms

Saarland University, Center for Bioinformatics, Saarland Informatics Campus, Postfach 15 11 50,
66041 Saarbrücken, Germany

1.1 Structural and Hierarchical Aspects

1.1.1 Size of Proteins

The size of proteins ranges from very small proteins, such as the 20-amino acid miniprotein Trp cage, to the largest protein in the human body, titin, which consists of about 27 000 amino acids and has a molecular weight of 3 million Dalton. Generally, when speaking of typical proteins, we refer to compact proteins of about 80 to 500 amino acids (residues) in size. Tiessen et al. reported that archaeal proteins had the smallest average size (283 aa), followed by bacterial proteins (320 aa) and eukaryotic proteins (472 aa) [1]. Among eukaryotes, plant proteins (392 aa) had a smaller size, whereas animal proteins (486 aa) and proteins from fungi (487 aa) were larger.

1.1.2 Protein Domains

The larger a single protein gets, the higher is the chance that it will be composed of multiple structurally distinct "domains." These are typically sequential parts of the protein sequence with a characteristic length between 100 and 200 amino acids [2]. For example, the protein Src kinase consists of an SH3 domain (that binds to proline-rich peptides), an SH2 domain (that binds to phosphorylated tyrosine residues), and the catalytic kinase domain, see Figure 1.1. In the inactive state, the SH3 domain will hold on to the linker connecting SH2 and catalytic domain that contains several prolines, and the SH2 domain will hold on to a phosphorylated tyrosine in the C-terminal tail of the catalytic domain. Thereby, all three domains are locked in a conformationally restricted state. Once activated by dephosphorylation of the tyrosine, these contacts are released, and the catalytic domain can undergo the characteristic Pacman-type opening/closing motion of protein kinases, enabling the binding of adenosine triphosphate (ATP). In the closed conformation, the active site residues catalyze transfer of the terminal γ-phosphate of ATP to a nearby tyrosine of a substrate protein bound on the Src kinase surface. The catalytic

Protein Interactions: The Molecular Basis of Interactomics, First Edition.
Edited by Volkhard Helms and Olga V. Kalinina.

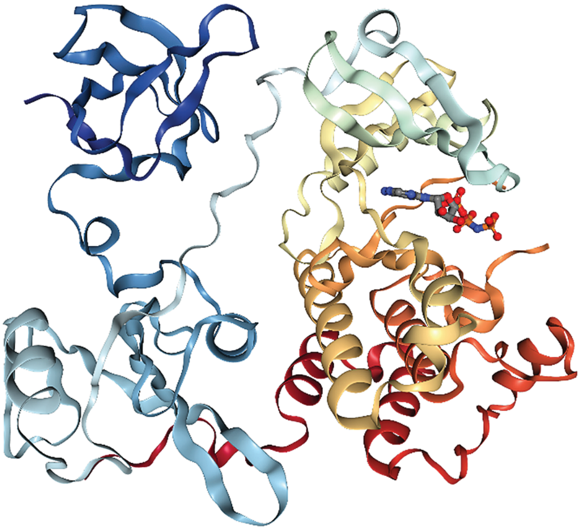


Figure 1.1 X-ray structure (PDB code 1AD5) of human Src kinase. The peptide sequence starts with an SH3 domain (top left), followed by an SH2 domain (bottom left) and then leads to the catalytic kinase domain (right). ATP is bound between small (top) and large lobe (bottom) of the kinase domain. Source: Figure generated with NGL viewer.

domain of kinases itself consists of two domain-like "lobes," a smaller N-terminal lobe (of about 80 aa) and a larger C-terminal lobe (of about 180 aa).

Although multi-domain proteins exist in all life forms, more complex organisms (having a larger number of unique cell types) contain more unique domains and a larger fraction of multi-domain proteins: eukaryotes have more multi-domain proteins than prokaryotes, and animals have more multi-domain proteins than unicellular eukaryotes [3].

1.1.3 Protein Composition

The composition of a protein depends on its environment and its posttranslational modifications, such as phosphorylation and sumoylation. For example, extracellular domains of most cell membrane proteins are often extensively glycosylated. Here, we will focus on the varying mixture of the 20 commonly occurring amino acids that make up most of all existing proteins. Water-soluble proteins possess a rather hydrophobic core and a polar surface that is in contact with the cytoplasm. This clear organizational principle provides the main driving force for the folding of water-soluble domains via the "hydrophobic effect."

Prokaryotic proteins contain more than 10% of leucine and about 9% of alanine residues, but rather few (only 1–2%) cysteine, tryptophan, histidine, and methionine residues [4]. Brüne et al. compared the amino acid composition of prokaryotic and eukaryotic proteins [5]. Eukaryotes have the highest variability for proline, cysteine, and asparagine. Amino acids showing high variability across species are lysine, alanine, and isoleucine, whereas histidine, tryptophan, and methionine vary the least. Cysteine is more common in eukaryotes than in archaea and bacteria, whereas isoleucine is less abundant in eukaryotes. The authors also analyzed the differential usage of amino acids in domains and linkers. Proline and glutamine, but to a smaller extent, polar and charged amino acids, are more common in linkers that are rather exposed to surrounding water. Globular domains contain larger fractions of hydrophobic amino acids, such as leucine and valine, and aromatic ones, such as phenylalanine and tyrosine.

1.1.4 Secondary Structure Elements

Folded proteins contain two types of secondary structure elements, α-helices and β-sheets. α-Helices have lengths between 9 and 37 residues with a peak at 11 amino acids [6]. β-Sheets are considerably shorter, being 2–17 residues long with a peak at 5 residues [7]. The secondary structure content of proteins ranges from purely helical proteins, such as myoglobin, containing six α-helices (see Figure 1.2) over mixed α/β proteins to so-called β-barrels, such as green fluorescent protein (GFP), see Figure 1.3, or Omp membrane pores in the outer membranes of gram-negative bacteria. Secondary structure elements provide stability to the protein structure and serve, e.g to anchor the catalytic residues of the active site at precise positions from each other (see below). α-Helices are also the structural basis of coiled coils, see Figure 1.4, because the helices can nicely pack against each other. α-Helices are frequently used by transcription factors, such as GCN4, at the DNA-binding interface, where the α-helices can intercalate in the major or minor grooves of the DNA double helix.

1.1.5 Active Sites

Active sites of enzymes are locations where bound substrate molecules undergo chemical modifications while being bound to the enzyme. Figure 1.5 shows the active site of the serine protease chymotrypsinogen A with the characteristic catalytic residues serine, histidine, and aspartic acid. In principle, discussing enzymatic mechanisms is out of scope for this book, which mostly deals with interactions that proteins engage in. Some multienzyme complexes having multiple active sites assemble to enable the product of one reaction to be passed from

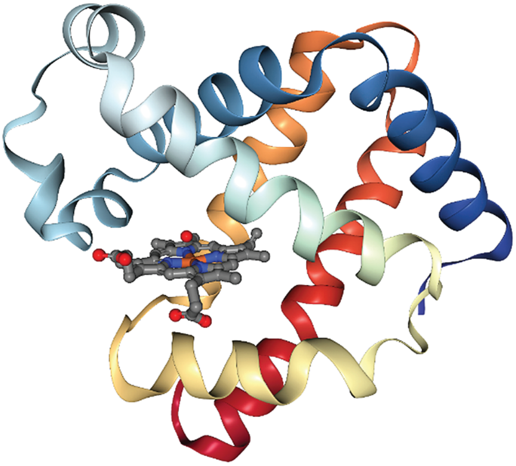


Figure 1.2 X-ray structure (PDB code 1MBN) of myoglobin from *Physeter catodon*. The porphyrin cofactor is anchored between six alpha helices. Source: Figure generated with NGL viewer.

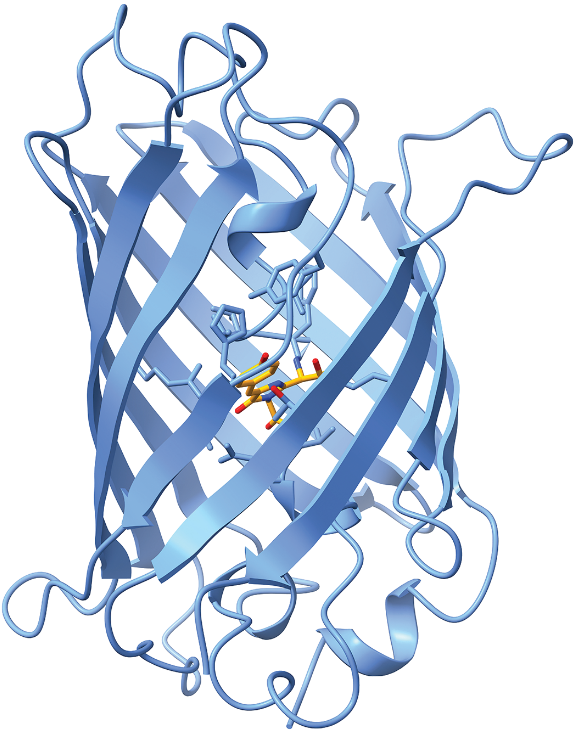


Figure 1.3 X-ray structure of the green fluorescent protein from *Aequorea victoria* (PDB code 1EMA). The barrel-shaped structure is formed by 11 beta-strands surrounding a central alpha-helix holding the chromophore. Source: Figure generated with UCSF Chimera.

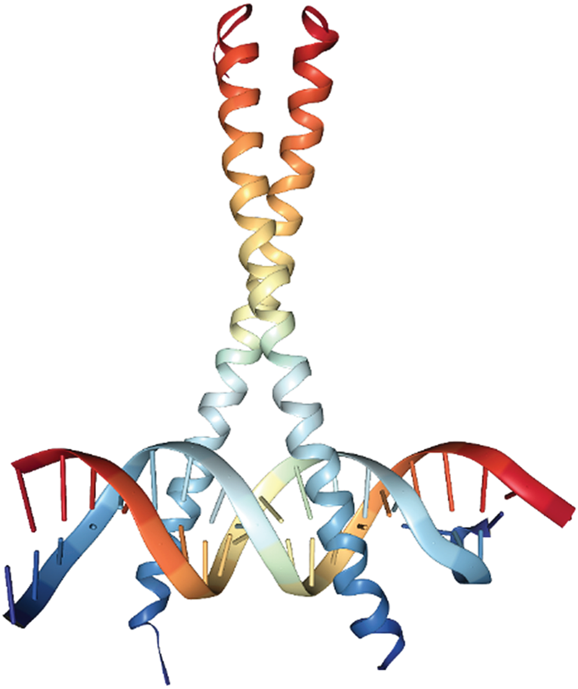


Figure 1.4 X-ray structure of GCN4 dimer from *S. cerevisiae* forming a so-called coiled coil and bound here to DNA (PDB code 1YSA). Source: Figure generated with NGL viewer.

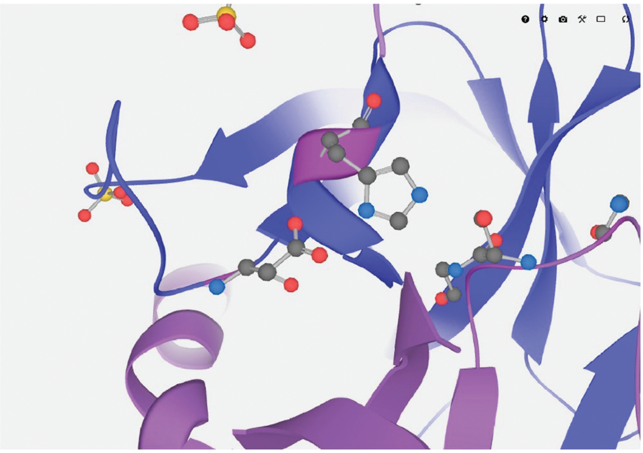
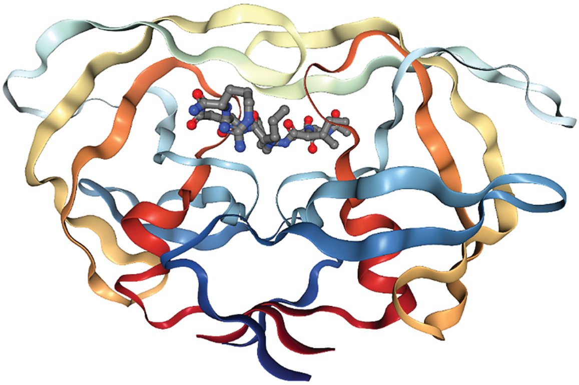


Figure 1.5 Catalytic triad – aspartic acid, histidine, serine – in the active site of a serine protease. Source: European Molecular Biology Laboratory (EMBL).

one active site to other, where it becomes the substrate of a follow-up chemical reaction. Generally, access to active sites should not be precluded by binding to other interaction partners, although, in some cases, binding patches need to be close to the active site, e.g. when a kinase binds its substrate on a patch on the surface of the large lobe so that a phosphate group can be transferred from bound ATP to a serine residue of the bound substrate as mentioned before.

Often, the active sites of enzymes are located on the protein surface, so that substrates can easily bind while remaining partially solvent exposed. A frequent structural motif is a flexible protein loop that reaches over the bound substrate, e.g. in HIV protease, see Figure 1.6. In other cases, the active site is located inside the protein, such as for cytochrome P450 enzymes or acetylcholine esterase. There, substrates need to pass into the protein structure through a channel that may be up to several nanometers long, see Figure 1.7. The main purpose of such an arrangement is to place the substrate in a low-dielectric cavity that enables complicated chemical reactions to take place. Note that the strength of electrostatic interactions is inversely

Figure 1.6 X-ray structure of an HIV protease dimer (PDB code 4HVP). A substrate peptide is bound in the active site. Access to the active site is controlled by opening/closing transitions of two flexible loops above the peptide (flaps). Figure generated with NGL viewer.



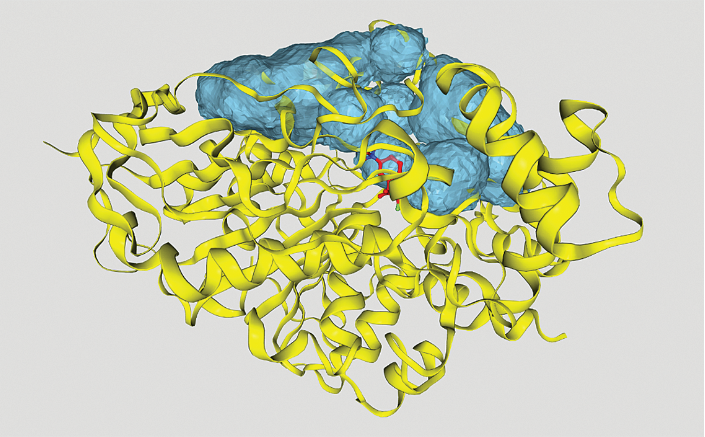


Figure 1.7 Trimethyl ammonio trifluoroacetophenone ligand bound in the active site of acetylcholinesterase from *tetronarce californica* (PDB code 1AMN). The surface contours illustrate several pores and cavities that make up tunnels leading to the internal active site. Source: The figure was generated with the ProPores2 web server (https://service .bioinformatik.uni-saarland.de/propores) [8].

proportional to the dielectric constant of the environment. In a low dielectric environment, charged protein residues can exert stronger electron-pulling or pushing effects on the substrate. Enzyme active sites, ligand binding sites, or translocation pores of ion channels can either reside in individual protein units or in between the interfaces of multimers.

1.1.6 Membrane Proteins

Integral transmembrane proteins are integrated into cellular membranes whereby their amino acid chain crosses the hydrophobic bilayer once or multiple times. While their soluble domains have the same composition as water-soluble proteins, the membrane-spanning parts have a so-called “inside-out” composition. These membrane regions are very hydrophobic on the outside that is in contact with the aliphatic lipid chains of the phospholipid bilayer and have a partially polar interior that often contains a water-filled translocation channel for substrate molecules. When the peptide chain crosses the bilayer, no hydrogen bonding is possible with the aliphatic lipid chains that are in strong contrast to the situation in the water phase. To satisfy the hydrogen bonding capacity of its backbone atoms, the chain thus adopts either an α-helical conformation or a β-sheet conformation in the membrane. Beta barrels consist of 8–22 β-sheets [9] but are only found in the outer membranes of gram-negative bacteria, mitochondria, and chloroplasts. Helical transmembrane proteins possess between 1 and around 20 alpha helices [10] that are between 10 and 30 residues long. The majority of helical membrane proteins possess only 1 transmembrane domain (TMD), followed by those having 2 TMDs and smaller fractions with 3, 4, 7, and 12 TMDs [10]. Oligomerization is frequently found among helical transmembrane proteins, whereby their binding interfaces consist of roughly perpendicular α-helices. Many receptors on cell surfaces form functional dimers. Ion channels form tetra- and hexamers, with the ion-conducting pore between the monomers. Interactions between proteins and membranes are further discussed in Chapter 13.

1.1.7 Folding of Proteins

Predicting the folded structure of a protein from its sequence has long been a holy grail. In the meantime, scientists have been able to put many pieces of this puzzle together. Important contributions to this were, e.g. the phi-value analysis experiments by Fersht and coworkers that quantify the degree of native folded structure around mutated residues in the folding transition state [11] and the theoretical work by Wolynes, Onuchic, and others, who drew an analogy between the folding of biopolymers and relaxation processes in spin glasses [12]. According to this "new view" of protein folding, a polypeptide chain folds on a rugged funnel-shaped energy landscape where the entropy is plotted on the x-axis and the enthalpy on the y-axis. A protein reaches the lowest free energy point, its folded state, by trading entropy for enthalpy. In this model, protein chains are not able to fold properly either above the folding temperature (where adopting a compact folded structure is entropically unfavorable) or below the glass-transition temperature (where the protein dynamics essentially freeze before reaching the folded state). The David Baker group has been leading the protein structure prediction field for many years using their Rosetta simulation method that extensively samples the combinatorial structural manifold made up of small structural fragments [13]. A further important advance was the brute-force molecular dynamics simulations by the D.E. Shaw group, who were able to simulate the repeated folding and unfolding of small globular proteins at the folding temperature [14]. Recently, the company DeepMind successfully applied deep-learning methods to tackle the problem of protein structure prediction [15, 16]. They trained a neural network to make accurate predictions of the distances between pairs of residues. In the latest Critical Assessment of protein Structure Prediction (CASP), their method termed AlphaFold2 created highly accurate structure predictions with a median backbone accuracy of 0.96 Å root mean square deviation (RMSD) and all-atom accuracy of 1.5 Å RMSD.

Proteins are synthesized by ribosomes either in the cytosol, close to the membrane of the endoplasmic reticulum, or close to the bacterial plasma membrane [17]. It is becoming more and more clear that portions of the nascent peptide chains may already start adopting alpha-helical conformations while passing through the ribosomal exit tunnel. All proteins of the secretory pathway and all membrane proteins are passed from the ribosome to the Sec translocon, an integral membrane channel in the endoplasmic reticulum (ER) membrane. The peptide sequences of membrane proteins are able to exit the Sec complex sideways into the membrane via a so-called lateral gate. Proteins targeted for the secretory pathway need to translocate into the ER, and often get glycosylated by a nearby oligosaccharyltransferase enzyme.

1.2 Conformational Dynamics

Thermal motion of atoms implies that proteins are not rigid objects. Yet, they can still be fairly stiff and have a pure scaffolding function. Examples of this are the proteins of virus capsids or the cytoskeleton. Most proteins, however, undergo some type of conformational transition either during their catalytic cycle, when they bind and unbind ligands, or if they are part of a signaling cascade.

1.2.1 Large-Scale Domain Motions

Proteins consisting of multiple domains or lobes (such as kinases) can undergo large-scale conformational transitions by characteristic domain movements. Prototypes for this are kinases and lysozyme. The first normal mode typically describes a Pacman-type opening–closing transition of the two domains relative to each other, see Figure 1.8. The second normal mode would then be a scissor-like motion perpendicular to the first mode. Often, these movements are connected to biological functions and facilitate either ligand binding and unbinding or help in catalyzing the enzymatic reaction. Membrane transporters, such as the leucine transporter LeuT, undergo a conformational transition between an inward-facing conformation and an outward-facing conformation, see Figure 1.9.

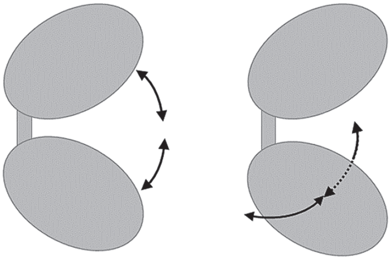


Figure 1.8 Schematic illustration of the first (lowest energy) normal mode of a two-domain protein, such as protein kinases (left), and the second normal mode (right).

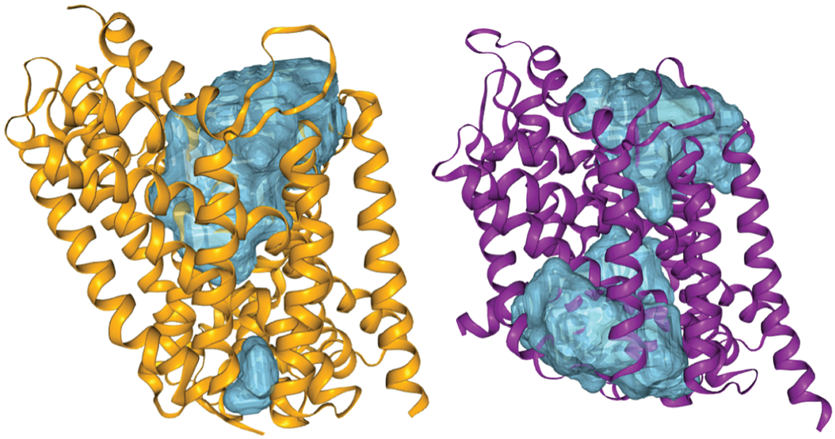


Figure 1.9 X-ray structures of the bacterial leucine transporter LeuT in the outward-facing conformation (left, PDB code 3TT1) and in the inward-facing conformation (right, PDB code 3TT3). The figures were again generated with ProPores2 (cf. Figure 1.7).

Besides such large-scale dynamics, the rest of the protein structure is of course not rigid but undergoes constant thermal motion as well. Since the 1970s, time-resolved IR spectroscopy was used to characterize the dynamics of laser-induced CO dissociation from the internal porphyrin ring of myoglobin [18]. The observed multi-exponential kinetics of the time needed for CO to rebind to the porphyrin was interpreted to reflect the intrinsic dynamics of the myoglobin matrix. Subsequently, Halle and coworkers showed, by NMR, that water molecules buried in the protein bovine pancreatic trypsin inhibitor (BPTI) exchanged with bulk solvent on time scales of milliseconds [19]. This proved that even compact globular protein structures undergo continuous conformational breathing transitions that are large enough to allow the passage of water molecules in and out of a folded protein.

1.2.2 Dynamics of N-Terminal and C-Terminal Tails

N-terminus and C-terminus of a protein chain are typically located on its protein surface, where they often stretch out into solution and have substantial conformational flexibility. Probably, the functionally most important N-terminal tails are those of histone proteins. They undergo posttranslational modifications in many ways, and this strongly affects their interaction with double-stranded DNA that winds around histone proteins. The C-terminal tails of proteins can function, e.g. as recognition sites for PDZ adaptor domains.

1.2.3 Surface Dynamics

Amino acid side chains on the surface of proteins often also show considerable conformational dynamics [20]. Frequently, transient pockets open and close on protein surfaces on a timescale of tens of picoseconds. Thus, the protein surface rather resembles the surface of a sponge. Another type of functionally relevant conformational motions are loop movements on the protein surface, e.g. lipases possess a loop termed “lid” that controls access to the active site beneath. The same is the case for HIV protease as mentioned before. Interestingly, it has been argued that disease-associated mutations in proteins often result in flexibility changes even at positions distal from mutational sites, particularly in the modulation of active-site dynamics [21].

1.2.4 Disordered Proteins

X-ray crystallography and Cryo-EM are perfect structural techniques to resolve precise conformational details of well-ordered portions of proteins. Obviously, N-terminus, C-terminus, and surface loops extend into the solvent, and their conformational dynamics may sometimes not yield precise electron density that can be detected against the background. Furthermore, it came as a surprise when NMR experiments showed in the mid 1990s that there exist numerous “disordered” proteins that do not adopt a well-folded conformation at all. Sometimes, they may refold when they bind to other proteins, or when they undergo a phenotypic order-to-disorder transition, such as the prion protein that is more folded in the non-disease state and is thought to be the origin of mad cow disease. All of us contain prion proteins and we are usually just fine. According to the "protein-only" hypothesis, the key event in the prion disease pathogenesis occurs when the cellular prion protein (PrPC) undergoes a conformational transition from a mainly α-helix-rich folded structure into an infectious and pathogenic β-sheet-rich conformer (PrPSc). PrPSc possesses abnormal physiological properties, such as resistance to proteolytic degradation, relative insolubility, and the propensity to polymerize into scrapie agents [22].

Monzon et al. distinguished short and disordered regions (between 5 and 30 residues long) that are usually associated with flexible linkers or loops in folded proteins and so-called long disorder regions (LDRs) that have at least 30 consecutive disordered residues. These LDRs were found to be enriched in charged and hydrophilic amino acids and depleted in hydrophobic ones [23], such as the linker segments discussed before in the context of protein domains. Disordered regions may also have important roles in mediating protein interactions. For example, so-called eukaryotic linear motifs (ELMs) are located in disordered regions of proteins and mediate interactions between proteins [24].

1.3 From Structure to Function

1.3.1 Evolutionary Conservation

One important principle of evolutionary biology is that functionally important protein regions tend to be conserved between related organisms whereas unimportant regions are subject to considerable variation. Functionally important regions include, of course, active site residues. Mutations of catalytic residues may render enzymes nonfunctional and are, therefore, rarely tolerated. Furthermore, conservation also extends to structural elements, such as disulfide bridges and residues in short turns.

In general, structure is better conserved than sequence. Therefore, functionally related pairs of proteins may sometimes show very low sequence similarity, but fairly high structural similarity. Assuming that both proteins were derived from a distant common ancestor protein, it came about that their structures were conserved during evolution, but their sequences were not, except for a few crucial positions.

1.3.2 Binding Interfaces

Many proteins carry out their function by binding to other proteins, small molecules, membranes, or nucleic acids. This is actually what all of this book is about. Usually, this involves one or more binding patches on the surface of the proteins. Binding interfaces of two proteins have sizes ranging from 500 to 3000 Å^2 [25]. Small interfaces are preferred for transient contacts of small hydrophilic proteins, e.g. those of redox proteins such as the electron carrier cytochrome *c*. In contrast, antibodies bind to their antigens with rather large and hydrophobic interfaces that support permanent or at least long-lasting contacts. Also, permanent dimers tend to have rather hydrophobic interfaces. How much of the protein surface is part of an interface depends on the total size of the complex. An internal protein, e.g. in the ribosome may even be fully shielded from solvent and all of its surfaces are in contact with other biomolecules. Protein–protein interactions and large protein complexes are discussed in Chapters 2–7.

DNA and RNA are strongly negatively charged due to their phosphate backbones. Hence, proteins need to possess complementary, positively charged surface patches, to be able to bind to DNA or RNA. Such patches are typically not suitable for binding to other proteins. However, there are certain proteins that are able to mimic nucleotide polymers. One example is the intracellular inhibitor protein barstar that binds to the RNAse barnase and prevents it from chewing up all mRNA and other RNA molecules inside the cell. Thus, barnase only acts extracellularly. Barstar has a strongly negative binding patch to mimic the natural substrate RNA. Chapters 10–12 give a deeper insight into protein interactions with nucleic acids.

The topology and composition of binding interfaces will be discussed in detail in Chapter 2.

1.3.3 Surface Loops

Surface loops are used, for example by antibodies, to bind to their antigens via complementarity-determining regions (CDRs). As mentioned, surface loops can also regulate the access to the active site of proteins, and they may contain cleavage sites for restriction enzymes. Note that cleavage is almost as frequently observed in α-helices as in regions without secondary structure, such as loops, but less in β-strands [26].

1.3.4 Posttranslational Modifications

Often, the activity of proteins is determined by the proper placement of posttranslational modifications to surface residues. For example, about 75% of all human proteins get phosphorylated, often at multiple positions [27]. Other modifications are glycosylation, farnesylation (e.g. of the Ras protein), etc. Ubiquitination often ends the life of proteins because this modification targets them for transport to the proteasome that shreds peptide sequences into small components. The modification sites are usually located on the protein surface and the modifications are placed by other enzymes, again involving protein interactions. Posttranslational modifications are important markers for binding partners and may also affect protein conformation (see Chapter 17 for further discussion).

1.4 Summary

The characterization of protein structure has become fairly routine these days. For about 70% of all human proteins, there exist structural models either from experimental determination or from homology modeling [28]. In fact, DeepMind, in cooperation with European Bioinformatics Institute (EBI), recently published structural models produced with AlphaFold for all human proteins and proteins of several other model organisms [29]. Some believe that even the protein folding problem has been, at least partially, solved. Despite all the accumulated knowledge, we still do not know the function of a considerable fraction of the human proteins, and it is very hard to rationalize the functional effects of posttranslational modifications or to even predict them. We have a limited understanding of what determines protein interactions, and we are rarely able to correctly predict the structures of protein assemblies from scratch, without additional experimental evidence.

References

1 Tiessen, A., Pérez-Rodríguez, P., and Delaye-Arredondo, L.J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Res. Notes* 5: 85. https://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-5-85.

2 Wheelan, S.J. et al. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics* 16: 613–618.

3 Yu, L., Tanwar, D.K., Penha, E.D.S. et al. (ed.) (2019). Grammar of protein domain architectures. *Proc. Natl. Acad. Sci.* 116: 3636–3645. https://www.pnas.org/content/116/9/3636.

4 Hormoz, S. (2013). Amino acid composition of proteins reduces deleterious impact of mutations. *Sci. Rep.* 3: 2919.

5 Brüne, D., Andrade-Navarro, M.A., and Mier, P. (2018). Proteome-wide comparison between the amino acid composition of domains and linkers. *BMC Res. Notes* 11: 117.

6 Kumar, S. and Bansal, M. (1998). Geometrical and sequence characteristics of α-helices in globular proteins. *Biophys. J.* 75: 1935–1944.

7 Penel, S. et al. (2003). Length preferences and periodicity in β-strands. Antiparallel edge β-sheets are more likely to finish in non-hydrogen bonded rings. *Protein Eng. Des. Sel.* 16: 957–961.

8 Hollander, M., Rasp, D., Aziz, M., and Helms, V. (2021). ProPores2: web service and stand-alone tool for identifying, manipulating and visualizing pores in protein structures. *J. Chem. Inf. Model.* 61: 1555–1559.

9 Tian, W., Lin, M., Tang, K. et al. (2018). High-resolution structure prediction of β-barrel membrane proteins. *Proc. Natl. Acad. Sci.* 115: 1511–1516.

10 Reeb, J., Kloppmann, E., Bernhofer, M., and Rost, B. (2015). Evaluation of transmembrane helix predictions in 2014. *Proteins* 83 (3): 473–484.

11 Matouschek, A., Kellis, J.T. Jr., Serrano, L., and Fersht, A.R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature* 340: 122–126.

12 Onuchic, J.N. and Wolynes, P.G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* 14: 70–75.

13 Yang, J., Anishchenko, I., Park, H. et al. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* 117: 1496–1503.

14 Robustelli, P., Piana, S., and Shaw, D.E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* 115: E4758–E4766.

15 Jumper, J., Evans, R., Pritzel, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.

16 Senior, A.W., Evans, R., Jumper, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* 577: 706–710.

17 Bornemann, T., Jöckel, J., Rodnina, M.V., and Wintermeyer, W. (2008). Signal sequence–independent membrane targeting of ribosomes containing short nascent peptides within the exit tunnel. *Nat. Struct. Mol. Biol.* 15: 494–499.

18 Austin, R.H., Beeson, K.W., Eisenstein, L. et al. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* 14: 5355–5373.

19 Denisov, V.P., Peters, J., Hörlein, H.D., and Halle, B. (1996). Using buried water molecules to explore the energy landscape of proteins. *Nat. Struct. Biol.* 3: 505–509.

20 Helms, V. (2007). Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem* 8: 23–33.

21 Campitelli, P., Modi, T., Kumar, S., and Ozkan, S.B. (2020). The role of conformational dynamics and allostery in modulating protein evolution. *Annu. Rev. Biophys.* 49: 267–288.

22 Baral, P.K., Yin, J., Aguzzi, A., and James, M.N.G. (2019). Transition of the prion protein from a structured cellular form (PrPC) to the infectious scrapie agent (PrPSc). *Protein Sci.* 28: 2055–2063.

23 Monzon, A.M., Necci, M., Quaglia, F. et al. (2020). Experimentally determined long intrinsically disordered protein regions are now abundant in the protein data bank. *Int. J. Mol. Sci.* 21: 4496.

24 Tompa, P., Davey, N.E., Gibson, T.J., and Babu, M.M. (2014). A million peptide motifs for the molecular biologist. *Mol. Cell* 55: 161–169.

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

26 Timmer, J.C., Zhu, W., Pop, C. et al. (2009). Structural and kinetic determinants of protease substrates. *Nat. Struct. Mol. Biol.* 16: 1101–1108.

27 Sharma, K., D’Souza, R.C.J., Tyanova, S. et al. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* 8: 1583–1594.

28 Somody, J.C., MacKinnon, S.S., and Windemuth, A. (2017). Structural coverage of the proteome for pharmaceutical applications. *Drug Discovery Today* 22: 1792–1799.

29 Varadi, M., Anyango, S., Deshpande, M. et al. (2022). AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50: D439–D444.