1

An Introduction to the Protein Molecule

1.1 Why Study Protein Moonlighting?

It is a mitochondrial protein essential for energy production. It is also a key controller of the essential process of apoptosis. It is the second enzyme of the glycolytic pathway and a secreted pro-cancer signal important in breast cancer. It is the central enzyme of glycolysis, which also performs the functions of the major bacterial virulence factors.

These seemingly paradoxical statements encapsulate what is an emerging story in the biology of the protein molecule. A growing number of prokaryotic and eukaryotic proteins have been found to exhibit more than one unique biological function. The number of such multifunctional, or moonlighting, proteins being discovered is increasing, and reviews of the literature, such as this book, are also identifying historical reports of protein moonlighting. A number of databases that encapsulate the data on the known moonlighting proteins are now available online (Hernandez et al. 2014; Mani et al. 2015). It is estimated that up to 300 proteins have protein moonlighting behavior. As will be discussed in later chapters of this book, this is likely to be only a small proportion of the total number of proteins that can moonlight. Indeed, this is one of the key questions that need to be addressed in the field of protein biology. It is recognized that multicellular eukaryotes have low numbers of protein-coding genes. For example, *Homo sapiens* seems to be able to control its 10^{13} cells with only 19 000 protein-coding genes (Ezkurdia et al. 2014). This seems a very low number of genes to generate the human functional proteome. Protein moonlighting might be one phenomenon that could account for the needs for such small numbers of proteins to be able to "run a human."

The three examples of moonlighting proteins that began this discussion are the very well-known proteins: cytochrome C (Cyt C), phosphoglucoisomerase (PGI), and glyceraldehyde 3-phosphate dehydrogenase (GAPDH). In addition to their established enzymatic functions, these three proteins have completely distinct and novel functions important in both physiological and pathological processes. At the current time, Cyt C appears only to have actions revolving around the control of apoptosis. The PGI protein has five distinct biological functions (see Chapter 3), and the family of GAPDH proteins has a bewilderingly large number of biological functions in both prokaryotes and eukaryotes (Sirover 2014). Surprisingly, as will be discussed in Chapter 8, GAPDH proteins from a number of pathogenic bacteria can function as so-called virulence factors mimicking the actions of bacterial toxins, adhesins, invasins, evasins, and iron-binding proteins. Indeed, one of the many surprises in the protein moonlighting literature is

Protein Moonlighting in Biology and Medicine, First Edition. Brian Henderson, Mario A. Fares and Andrew C. R. Martin.

that both human GAPDH (Sheokand et al. 2013) and the GAPDH from some bacteria like the major human pathogen, *Mycobacterium tuberculosis* (Boradia et al. 2014), function as cell surface and secreted binding proteins for the iron-carrying protein, transferrin. It would appear that the same moonlighting protein is important in iron sequestration in humans and mycobacteria and are likely to be pitted against each other in the ancient scourge, tuberculosis. This example of GAPDH exemplifies the finding that a proportion of moonlighting proteins can exhibit multiple functions. For example, the molecular chaperone, chaperonin (Cpn) or heat shock protein (Hsp)60 family of proteins, exhibits over 40 different biological functions (Henderson et al. 2013). It is not known if all moonlighting proteins have this capacity for multiple functionality.

Moonlighting proteins are now firmly established as participants in normal cellular, tissue, and organismal homeostasis as well as being parts of the mechanisms of tissue pathology and infectious disease. This book, written by a cellular biologist (Henderson), a protein bioinformaticist (Martin), and an evolutionary biologist (Fares), brings together the literature on protein moonlighting to provide a current overview of this new area of biology. To get the story started, this first chapter will introduce the reader to the world of the protein molecule.

1.2 A Brief History of Proteins

The concept of proteins first entered science in the eighteenth century. The French chemist, Antoine Fourcroy, in 1789, identified three different categories of what we now know are "proteins" from animal sources—albumin, fibrin, and gelatin—in addition to at least two classes in plants. Indeed, the name "albumins" was used as a generic term to describe all proteins at this time. The term "protein" emerges from the studies of two chemists, the world-renowned Swedish chemist, Jacob Berzelius, and the less well-known Dutch physician and chemist, Gerrit Mulder. Mulder was exploring the composition of natural products using newly developed methods of compositional analysis. Analyzing various "albumins," he was surprised to find that they all had virtually the same atomic composition (Mulder 1838). This led Mulder to speculate that all the albumins he had been studying might be composed of the same substance that he termed "Grundstoff." Mulder was in correspondence with Berzelius, who thought that this result should be noted with a specific name for the generic material composing all the albumins examined. The name he suggested was "protein," derived from the Greek word *proteos*, meaning "standing in front" or "in the lead" (Tanford and Reynolds 2003).

Soon after Mulder's paper was published, the influential scientist, Justus Liebig, entered the story. In 1841, he praised the work of Mulder and concluded that only four proteins existed in plants, while in animals he concluded that albumin and fibrin could be converted into blood. While not directly true, of course, we now know that these proteins are formed of the same 20 amino acids, which can be assembled in different ways. Gradually, the truth started to unfold. While "Grundstoff" was thought only to contain carbon, hydrogen, oxygen, and nitrogen in a fixed ratio, and sometimes was associated with sulfur, Liebig found that the sulfur could not always be separated; we now know that two amino acids (cysteine and methionine) contain sulfur. J.B. Dumas showed in 1842 that the ratio of carbon, hydrogen, oxygen, and nitrogen was not fixed, as thought by Mulder, showing that "Grundstoff" was much more varied than previously thought.

By 1900, it was realized that proteins are in fact made up of amino acid building blocks; and in 1902, the concept of the peptide bond linking amino acids was independently presented by both Emil Fischer and Franz Hofmeister at the same scientific meeting. At the time of Mulder's work in 1840, the only amino acids known were glycine and leucine—the rest were discovered over a period of more than 60 years by scientists isolating and characterizing the products of hydrolysis of proteins (Vickery and Schmidt 1931). All but three had been identified by 1901, with the last, threonine, being identified in 1936 (see Simoni et al. 2002).

The understanding of proteins made a leap forward with the discovery of the ability to crystallize them and solve their structures by X-ray crystallography. As far back as 1840, it had been discovered that components of earthworm blood (what we now know to be hemoglobin) could be persuaded to crystallize on glass plates and later, blood components from nearly 50 animal species were crystallized in a similar way. The discovery of X-rays in 1895 and the realization that they have wavelengths of the same order as interatomic separation set the scene for X-ray crystallography. In 1913, father and son team, William Henry and William Lawrence Bragg (1913), introduced their famous equation:

$$n\lambda = 2d \sin\theta$$

Here, λ is the wavelength, d is the repeat distance in the crystal, θ is the diffraction angle, and n is an integer leading to first-, second-, third- (etc.) order reflections. A typical diffraction pattern is shown in Figure 1.1, in which the spots are often known as "reflections." The structures of simple crystallized compounds, such as diamond or simple salts, could readily be elucidated from the values of d that could be calculated from the diffraction patterns. More complex chemical fibers such as cellulose were found to diffract too, but interpreting the results was more complex. In the 1930s, William Astbury started to look at protein fibers—mostly keratin from hair and feathers (see Hall 2014 for the story of this pioneer), but it was John Desmond Bernal and Dorothy Crowfoot (later Hodgkin) who collected the first X-ray data on a globular protein—the enzyme, pepsin (Bernal and Crowfoot 1934). Astbury's great contribution was to suggest that globular proteins, like pepsin, might be folded from structural elements essentially the same as fibrous proteins like keratin. Bernal and Hodgkin's work revealed the importance of water in stabilizing protein crystals and confirmed the globular shape of non-fibrous proteins. The reflections obtained indicated separations as small as 2 Å (0.2 nm), similar to the typical 1.54 Å (0.154 nm) bond length between two carbon atoms, but it took another 25 years before the structure of proteins was revealed in atomic detail.

As shown in Figure 1.1, the collection of data from X-ray diffraction results in a set of spots on a film or other recording device. For an electromagnetic wave, with a sine-wave oscillation, the recorded intensity depends on both the amplitude of the oscillation and the position within the wave (the phase) at which it strikes the recording device. The intensity of these spots can be measured, but to calculate the atomic coordinates that have led to the observed diffraction pattern, the phase is also needed. Max Perutz, working in Cambridge, realized that replacing one metal ion bound to a protein with another could change the intensities of specific spots in the diffraction pattern without disrupting the pattern as a whole, because the overall structure would not change. This discovery of "isomorphous replacement," together with the availability of

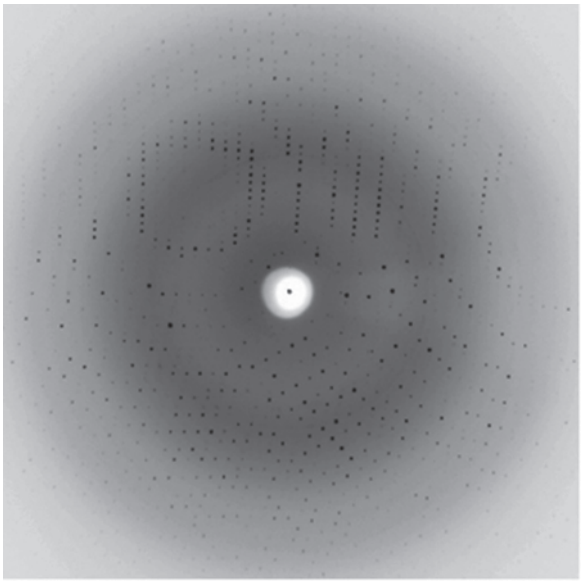

Figure 1.1 A typical diffraction pattern from X-ray crystallography. Source: http://www.chem.ucla.edu/harding/IGOC/D/diffraction_pattern.html. © University of California.

computers to crunch the data, allowed the phase problem to be circumvented. Perutz worked on hemoglobin, while his colleague John Kendrew worked on the smaller (but related) protein, myoglobin. The structure of myoglobin was solved, first at a resolution of around 6–7 Å and published in 1958 using data from around 400 diffraction spots (Kendrew et al. 1958). By 1960, a 2 Å resolution structure, using around 9600 spots, was published the first time a protein was seen at full atomic resolution (Kendrew et al. 1960).

1.3 Protein Biology

As we have seen, proteins are of course molecules and have a unique chemical composition. Their large size leads to the term "macromolecules," and like the other biological macromolecules, DNA and RNA, they are polymers built from small building blocks. In the case of proteins, those building blocks are the amino acids. There are 20 naturally occurring amino acids encoded by the DNA, all of which have a common structure as shown in Figure 1.2. With the exception of glycine, the alpha carbon has four different groups attached to it and, consequently, is optically active having left- and right-handed stereoisomers (enantiomers), known as the L- and D-forms. In proteins, the L-form is used virtually exclusively (the exceptions being some bacterial envelopes and natural antibiotics). In addition to the DNA-encoded 20 amino acids (Table 1.1), some other types are seen occasionally as a result of post-translational modifications or incorporation via variant use of the genetic code.

DNA encodes the amino acids that will form a protein in the form of triplets of DNA bases known as codons. There are four types of DNA base: A (adenine), T (thymine), C (cytosine), and G (guanine). DNA can therefore have $4^3 = 64$ different triplets; and

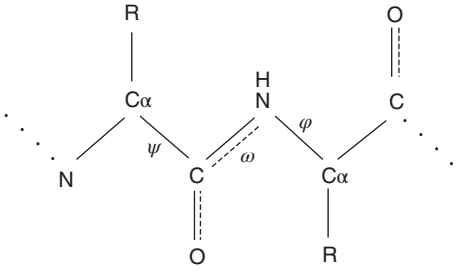

Figure 1.2 Amino acid and peptide bond structure. The figure shows two amino acids linked by a peptide bond. The R group varies between the different amino acids. The peptide chain would continue to the left and right (dotted lines). The bond between the N and Cα (φ) and the bond between Cα and C (ψ) are freely rotatable, but there are strong preferences for certain combinations of angles as a result of steric effects. The ω angle that describes the rotation about the C—N bond (the peptide bond) is constrained to be approximately 180° or 0° since the free electrons on the oxygen are delocalized, giving the bond partial double-bond characteristics. This angle is almost always approximately 180°, except when the following amino acid is proline.

Table 1.1 The genetic code.

		Second letter									
		U		C		A		G			
First letter	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U	**Third letter**
		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys		
		UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop		
		UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp		
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	C	
		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg		
		CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg		
		CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg		
	A	AUU	Ile	ACU	Thr	AAU	Asn	AUG	Ser	A	
		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser		
		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg		
		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg		
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	G	
		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly		
		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly		
		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly		

Triplets of RNA bases encode single amino acids.

since there are only 20 amino acids, the genetic code is redundant with most amino acids being encoded by multiple codons (Table 1.1).

The synthesis of proteins from the information encoded in DNA goes via an intermediate nucleic acid known as messenger RNA (mRNA). Proteins known as "transcription factors" bind to the DNA, opening the double helix to allow the initiation of transcription—the copying of the DNA code into mRNA. While DNA uses the four bases A, T, C, and G, RNA replaces T with U (uracyl). In eukaryotes, mRNA then undergoes extensive processing to remove regions known as introns that interrupt

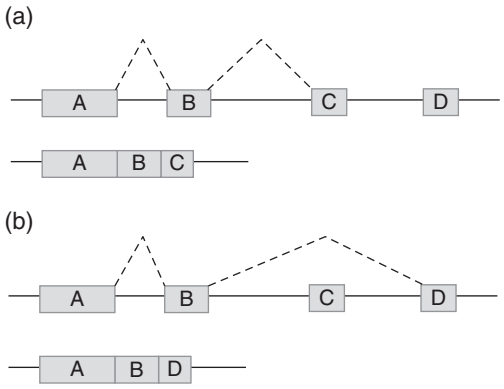

Figure 1.3 Splicing of messenger RNA. The RNA transcript contains exons (boxes A, B, C, D) an introns (black line). Splicing of the RNA (dotted lines) removes the introns leaving a mature transcript in which the exons are spliced together. The figure shows an example of alternative splicing: in (a) the exon D is discarded, while in (b) exon C is discarded leading to two alternatively spliced forms of the resulting protein.

the protein-coding DNA (see Figure 1.3); prokaryotes do not have introns. The transcribed mRNA is then translated into protein on a cellular machine known as the "ribosome." This structure—a complex of proteins and RNA—orchestrates the binding of transfer RNA (tRNA) molecules that have "anti-codons" to bind the mRNA and carry the correct amino acid. The amino acids are joined to one another in order to synthesize a protein chain—a linear sequence of amino acids. The synthesis of this chain always starts at an ATG codon, which encodes the amino acid methionine although this, together with signal sequences that target the protein to particular parts of the cell, is often later cleaved from the protein. Three of the 64 codons are reserved as stop signals that indicate the end of a protein chain.

Proteins have an enormous multiplicity of functions. While the importance of particular RNAs in cellular function is gradually being revealed (as will be described later), most cellular functions are mediated by proteins. One of their best known functions is as enzymes—highly specific biological catalysts. Enzymes are involved in every metabolic pathway ranging from core processes such as production of energy for the cell, to specific signaling pathways. Other functions of proteins include providing a purely structural role (e.g., supporting the cell or forming the eye lens), contractile functions (from changing the shape of a cell to moving a muscle), mediating interactions between cells (e.g., adhesins), and acting as messengers (e.g., hormones, growth factors, and cytokines) and as receptors for those messengers.

It is this huge diversity of protein function, and the ability of some proteins to perform multiple functions, that is the basis of protein moonlighting.

1.4 Protein Structure and Function

As described earlier, proteins are formed from 20 amino acid building blocks assembled in a linear sequence. The order in which the amino acids are connected to one another determines the way in which the protein folds in three dimensions and ultimately determines the function of the protein.

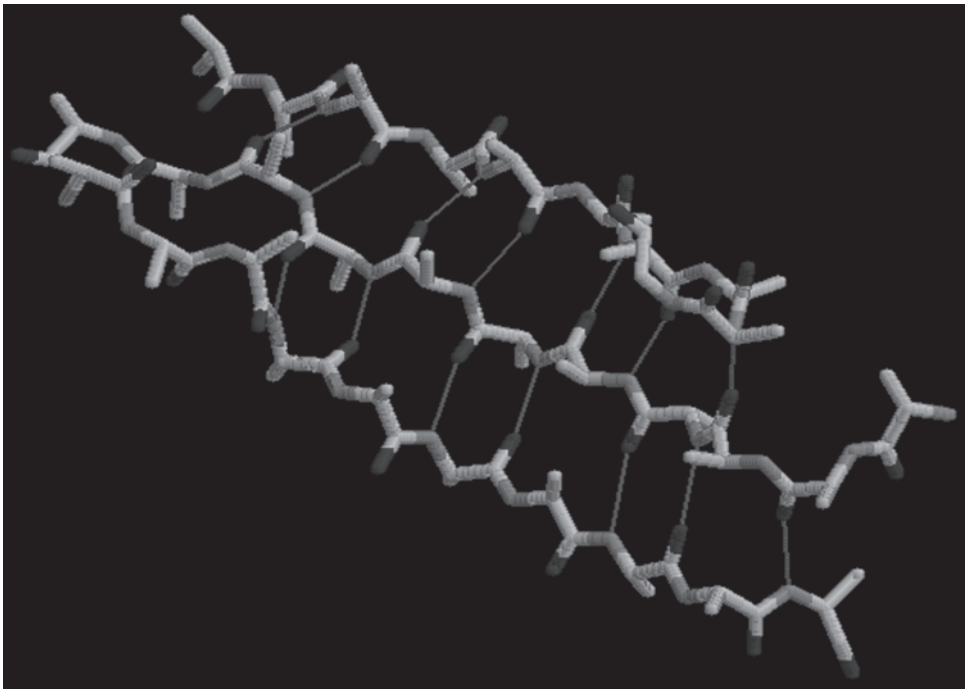

Figure 1.4 An example of an antiparallel β-sheet. The peptide in the strands is in a fully extended conformation and the strands are stabilized by hydrogen bonding of the backbone between the strands. In the antiparallel β-sheet (as shown) the adjacent strands run in opposite directions, while in a parallel sheet, the strands run in the same direction. (*See insert for color representation of the figure.*)

Amino acids are joined via peptide bonds. The carboxyl group of one amino acid links to the amide group of another in a condensation reaction to form the peptide bond (Figure 1.2). This bond has partial double-bond characteristics, as the lone pairs of electrons on the carboxyl oxygens are delocalized. This means that it has a planar conformation, only being able to adopt torsion angles of ~0° and ~180° referred to as *cis* and *trans,* respectively. The peptide bond is almost always in the *trans* conformation as this minimizes any steric hindrance. The exception is in peptide bonds preceding a proline, which strictly is an imino acid rather than an amino acid, because its sidechain links back onto the backbone nitrogen forming a five-membered ring. The location of the sidechain atoms means that the energy difference between *cis* and *trans* forms is much reduced and consequently the *cis* form is much more common.

Locally, the protein chain often folds into so-called "secondary structures." In a chain's most extended form, it creates a β-strand. These associate next to one another in either parallel or antiparallel arrangements to form β-sheets. β-sheets are stabilized by "backbone" hydrogen bonds that form between the nitrogens and oxygens of the peptide bonds on adjacent strands (see Figure 1.4). The second major type of secondary structure is the α-helix. This has 3.6 amino acids per turn and is an internally stabilized structure with backbone hydrogen bonds running parallel to the helix, the hydrogen on the backbone nitrogen of amino acid *n* bonding to the carboxyl oxygen of amino acid *n – 4* (see Figure 1.5). Owing to its cyclic sidechain, proline does not have a hydrogen

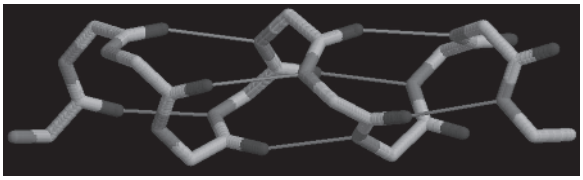

Figure 1.5 An α-helix showing the characteristic hydrogen bonding pattern in green. (*See insert for color representation of the figure.*)

on the backbone nitrogen and, therefore, cannot participate in this hydrogen bonding. It is therefore referred to as a "helix-breaker," although it is better thought of as a helix destabilizer, as it is able to adopt the conformation required for an α-helix. Proline can only be part of a β-sheet when it is in an edge strand of a sheet in a position where the nitrogen would not be involved in hydrogen bonding to the adjacent strand.

Secondary structure elements are linked by regions often known as "random coil." This does not mean that the structure of such regions is truly random—indeed the residues generally adopt a very distinct conformation. It simply means that they do not adopt one of the repeating structures described before. Within coil regions, there may be turns that adopt very specific conformations, which have been classified in detail (Sibanda and Thornton 1985). These generally form tight turns linking two strands of antiparallel β-sheet. In some cases, however, regions of proteins are truly random and disordered—in other words, these regions are highly dynamic and do not have a fixed structure. Such regions are often involved in low-affinity binding where they become structured once bound; functions such as signaling and reversible DNA binding often employ disordered regions.

The secondary structure elements pack together to form a tertiary structure often referred to as the "protein fold." Folded protein chains can then assemble into quaternary structures. These may be homo- or hetero-assemblies. For example, hemoglobin has two alpha chains and two beta chains forming a heterotetramer.

Generally, larger proteins fold into regions of 100–150 amino acids known as "domains." A domain is difficult to define unambiguously, but is generally regarded as a self-contained folding unit. Domains can be defined purely on the basis of structure, as is done in the CATH database (Cuff et al. 2011), or on the basis of evolution as is done in the SCOP2 database (Andreeva et al. 2014). CATH and SCOP2 are described later.

Often domains—at least in the sense used by SCOP2—are associated with a particular function. During evolution, such functional domains can then be assembled like parts of a kit to produce a novel protein. For example, one domain may be the power house of a protein, extracting energy stored in ATP, while another binds a particular substrate and catalyzes a reaction. Of course, function itself is also difficult to define. When referring to this example of a protein having two domains (one providing energy and the other carrying out a reaction), the function can be ascribed to the protein as a whole, or two functions can be ascribed—one to each of the domains. At the level of the protein, "function" can be thought of in a hierarchical sense—the function of a protein may be that it is an *enzyme*, or an enzyme that *catalyzes a particular type of reaction*, or an enzyme that *catalyzes this reaction for a particular substrate*. The hierarchical nature of enzyme nomenclature is captured in the Enzyme

Commission (EC) classification, which gives a four-digit identifier to each enzyme (or more strictly to each enzyme reaction). The gene ontology (GO) is a more holistic description of protein function, dividing function into three domains: cellular component, biological process, and molecular function (The Gene Ontology Consortium 2000)—see Chapter 4.

Thus, the notion of protein moonlighting is tied to the concept of protein domains. It is very common for different domains to have different functions, but as part of the same overall function of the protein—this is *not* protein moonlighting! However, if a protein, or indeed a domain, has two unrelated functions, then we have a case of protein moonlighting.

1.5 Protein Sequence Determination, Structures, and Bioinformatics

Protein sequence data used to be obtained solely by a process known as Edman degradation, in which the amino-terminal residue is labeled using phenylisothiocyanate to form a cyclical phenylthiocarbamoyl derivative and cleaved from the protein peptide as a thiazolinone derivative without disrupting the rest of the sequence. The thiazolinone amino acid is then extracted and stabilized and can be identified using chromatography or electrophoresis. The technique can accurately sequence up to 30 amino acids and modern machines are capable of over 99% efficiency per amino acid.

However, these days, proteins are normally sequenced using mass spectrometry which, in principle at least, can sequence a protein of any size. The protein is first fragmented using an endopeptidase—an enzyme that cleaves proteins at sites within the sequence (rather than at the ends). The resulting peptides are then separated using high-pressure liquid chromatography (HPLC) and sprayed into a mass spectrometer where they are fragmented and mass-to-charge ratios of the fragments are measured. The resulting spectrum is analyzed and compared against databases of known protein sequence data to determine the sequences of the fragments. The process is then repeated using different enzymes, which cleave at different sites, in order to obtain information about how the peptides overlap.

However, it is much more common to sequence DNA, which is now very rapid and cheap. The problem with genomic DNA is finding the regions that code for protein—particularly in the case of eukaryotes, where coding DNA typically makes up less than 5% of the genome and, as described before, the coding regions are interrupted by introns. Thus, computational methods have to be used to predict the location of genes, coding exons and introns—a process that is far from 100% accurate. Consequently, the best way of obtaining the coding information for a protein is to collect mRNA and reverse transcribe it into complementary DNA (cDNA)—that is, DNA that contains only the coding sequence derived from the RNA after the introns have been removed. The classical Sanger sequencing method has largely been replaced by a plethora of "next-generation" sequencing methods that allow very rapid sequencing of thousands of short stretches of DNA in parallel. Sequencing of the human genome—completed in 2003, with the first drafts published in 2001—took 13 years and cost approximately $1 bn. In January 2014, Illumina announced that the latest version of their HiSeq X Ten sequencing system would sequence a human genome in

its entirety for less than $1000 in less than a day (http://www.wired.co.uk/news/archive/2014-01/15/1000-dollar-genome). In 2015, Veritas Genetics broke this $1000 barrier (http://www.popsci.com/cost-full-genome-sequencing-drops-to-1000).

Sequence data for DNA are stored in three databanks: GenBank, EMBL-ENA, and DDBJ. These three databanks, from the United States, Europe, and Japan, respectively, act as deposition sites and exchange data on a regular basis such that they contain the same information in somewhat different formats. In addition to DNA data, where appropriate, they contain protein translations that are also available in separate resources, Genpept and UniProtKB. UniProtKB is split into two sections: UniProtKB/trEMBL contains protein translations from EMBL-ENA with some automatically generated annotations, while UniProtKB/SwissProt contains sequences for which additional manually verified and detailed annotations have been provided.

As described earlier, protein structures are largely determined by X-ray crystallography where crystals of protein are bombarded with X-rays that diffract. Once the phase problem has been addressed, the electron density of the atoms in the protein can be calculated from the diffraction pattern using Fourier transforms and the atomic structure can be fitted into this electron density. Around 10% of structures are determined by a different technique called "nuclear magnetic resonance" where distance constraints between atoms can be derived, allowing structures to be solved by building models that satisfy all the distance constraints. A few structures are solved by other techniques such as electron diffraction and neutron diffraction. These are low-resolution techniques and generally are used for very large proteins and complexes, often in concert with high-resolution techniques to obtain the detailed structure of the components.

Protein structure data are made available in the Protein Databank (PDB). The World Wide PDB (wwPDB, http://www.wwpdb.org/) is an umbrella organization for repositories including the RSCB PDB (http://www.pdb.org), the PDB in Europe (PDBe, http://www.ebi.ac.uk/pdbe/), and the Protein Databank Japan (PDBj, http://www.pdbj.org/). The different suborganizations all exchange data on a regular basis and provide data in the same format, but provide different analysis, query, and deposition tools via their web sites. An alternative view of data in the PDB is available through PDBSum (http://www.ebi.ac.uk/pdbsum/). This web site provides overview structural analysis of each structure, including quality assessment, secondary structure analysis, and simplified views of interactions with any substrates.

Two other resources that analyze, classify, and present structural data in the PDB are CATH and SCOP2, maintained at University College London (UCL) and Cambridge University, respectively. CATH (http://www.cathdb.info/) identifies structural domains in proteins and classifies those domains on the basis of Class (C, the secondary structure content—mostly α, mostly β, mixed αβ, or no secondary structure), Architecture (A, the layout of secondary structure elements in space), Topology (T, the connectivity between the secondary structure elements—in other words, the protein fold), and Homology (H, the homologous family—have domains descended from a common ancestor during evolution). The Homology level is further subdivided on the basis of sequence identity.

SCOP2 (http://scop2.mrc-lmb.cam.ac.uk/) takes a rather different approach. It uses four categorizations of proteins: (1) Protein types: soluble, membrane, fibrous, and intrinsically disordered; (2) Evolutionary events: allowing the annotation of various structural rearrangements and other oddities observed amongst related proteins; (3) Structural classes: similar to the CATH C-level, this organizes proteins according to

their secondary structure content, but unlike CATH subdivides domains having both α and β secondary structures into α/β (where the elements alternate) and $\alpha+\beta$ (where the elements are segregated); (4) Protein relationships: consisting of three subcategories: (4a) Evolutionary: Species (corresponding to the individual full-length sequence gene product), Protein (which groups orthologous proteins; in general, it is the same as the species grouping with the exception of fusion proteins found in some species), Family (corresponding to a conserved sequence region shared by closely related proteins and roughly equivalent to the Homology level in CATH), Superfamily (a common structural region shared by different protein families and roughly equivalent to the Topology level in CATH), Hyperfamily (a common region shared by different superfamilies, typically smaller than a structural domain); (4b) Structural Relationships: Fold (defined strictly on the basis of global structural features)—the composition of secondary structures and their architecture and topology—roughly equivalent to the C, A, and T levels of the CATH classification respectively; (4c) Other relationships: (i.e., internal structural repeats, common motifs, and subfolds). SCOP2's separation of evolutionary and structural classifications allows for the rare occasions on which homologous proteins (i.e., those that have descended from a common ancestor) have different structures, something that the monolithic classification used by CATH has difficulty in representing correctly.

While CATH uses a purely structural definition of protein domains, SCOP2 defines a domain as a "unit of relationship" whose boundaries are dependent on the relationship in question. Consequently, "fold" is related to a single structural domain, while the domains representing "Family" and "Superfamily" can span one or more structural domains. Thus, a "Family" domain generally represents a unit of inheritance—one or more structural domains that are inherited together and can be associated with other domains in a complete protein.

1.6 Regulation of Protein Synthesis

As mentioned earlier, in eukaryotes, RNA is processed before it is translated into protein. Introns that interrupt the coding region are removed and the coding exons are joined to one another. However, for many genes, the exons can be spliced together in different ways—they always appear in the same order, but exons may be skipped or left out of the complete spliced sequence (see Figure 1.3). In humans, it is estimated that at least one-third of genes undergo alternative splicing and that on average there are three splice variants per gene, meaning that the approximately 20 000 human genes can actually encode approximately 60 000 proteins.

Earlier, it was stated that proteins carry out most of the important roles in the cell. However, the importance of noncoding RNA (ncRNA) as a functional molecule, instead of just being involved in protein synthesis, has recently been realized. As well as tRNA, ribosomal RNA (rRNA), many other types of ncRNA have been identified including snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs, Xist, and HOTAIR. Recent transcriptomic and bioinformatic studies suggest that there are thousands of ncRNAs (Cheng et al. 2005; Morris 2012). However, it is possible that many of these are not functional (Hüttenhofer et al. 2005). ncRNAs fall into several groups involved in many cellular processes. In addition to those involved in protein synthesis, in

eukaryotes a molecular machine called the "spliceosome" that, like the ribosome, contains proteins and RNA performs RNA splicing, and in mammals, this process can be regulated by snoRNAs.

However, one of the most important roles of ncRNAs is in the regulation of many thousands of genes. This process can occur in two ways. In higher eukaryotes, *trans*-acting ncRNAs (which are encoded in parts of the genome not associated with the targets of their activity) such as micro-RNAs (miRNAs) regulate gene expression through partial complementary to mRNA molecules, generally in untranslated regions at the downstream end of the mRNA (3′ UTRs). In general, this has the effect of down-regulating gene expression.

Second, a number of *cis*-acting ncRNAs are encoded within the 5′ UTRs of protein-coding genes. For example, in prokaryotes, the regulation of amino acid-synthesizing operons (groups of genes involved in a single biosynthetic pathway) is mediated by RNA leader sequences upstream of the first gene. This regulatory mechanism has been seen in the synthesis of histidine, leucine, threonine, and tryptophan. Similarly, a riboswitch is a regulatory segment of an mRNA molecule that is able to bind a small molecule controlling the rate at which proteins, encoded by the mRNA, are synthesized (Tucker and Breaker 2005; Batey 2006). Regulation can also be indirect. For example, iron binds to iron-response proteins (IRPs) and the IRP–Fe complex can then bind to iron-response elements (IREs) found in UTRs of various mRNAs, which encode proteins involved in iron metabolism.

1.7 Conclusions

This chapter has set the scene for understanding protein moonlighting. It has precised the history of the discovery of proteins and has gone on to look at their composition and basic structure. The way in which the huge diversity of protein function can be categorized has been described in brief, and the problem of defining function and attributing function to a protein and to its constituent domains has been discussed. This is key to understanding the concepts of protein moonlighting. The methods by which protein sequence and structure information can be obtained have briefly been surveyed and bioinformatics resources for storing and cataloguing these data have been described. Finally the way in which proteins are synthesized has been described together with the importance of RNA processing and the role of ncRNAs in regulating gene expression.

In Chapter 2, the discussion switches to a consideration of the mechanisms involved in the evolution of the function of proteins. This will provide the reader with a background to the discussion of the potential mechanisms responsible for the evolution of moonlighting sites.

References

Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin A (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res* 42:D310–D314.

Batey RT (2006) Structures of regulatory elements in mRNAs. *Curr Opin Struct Biol* 16:299–306.

Bernal JD, Crowfoot D (1934) X-Ray photographs of crystalline pepsin. *Nature* 133:794–795.

Boradia VM, Malhotra H, Thakkar JS, Tillu VA, Vuppala B, Patil P, Sheokand N, Sharma P, Chauhan AS, Raje M, Raje CI (2014) *Mycobacterium tuberculosis* acquires iron by cell-surface sequestration and internalization of human holo-transferrin. *Nat Commun* 5:4730.

Bragg WH, Bragg WL (1913) The reflection of X-rays by crystals. *Proc R Soc* A88:428–438.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308:1149–1154.

Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones DT, Thornton JM, Orengo CA (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res* 39:D420–D426.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet* 23:5866–5878.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25:25–29.

Hall KT (2014) *The Man in the Monkey Nut Coat: William Astbury and the Forgotten Road to the Double Helix*. Oxford University Press: Oxford.

Henderson B, Fares MA, Lund PA (2013) Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions. *Biol Rev Camb Philos Soc* 88:955–987.

Hernández S, Ferragut G, Amela I, Perez-Pons J, Piñol J, Mozo-Villarias A, Cedano J, Querol E (2014) MultitaskProtDB: a database of multitasking proteins. *Nucleic Acids Res* 42(Database issue):D517–D520.

Hüttenhofer A, Schattner P, Polacek N (2005) Non-coding RNAs: hope or hype? *Trends Genet* 21:289–297.

Kendrew J, Bodo G, Dintzis HM, Parrish RG, Wyckoff H (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181:662–666.

Kendrew J, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 185:422–427.

Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, Zabad S, Patel B, Thakkar J, Jeffery CJ (2015) MoonProt: a database for proteins that are known to moonlight. *Nucleic Acids Res* 43(Database issue):D277–D282.

Morris, KV (editor) (2012) *Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection*. Caister Academic Press: Poole.

Mulder GJ (1838) Zusammensetzung von Fibrin, Albumin, Leimzucker, Leucin u.s.w. *Annalen Pharmacie* 28:73–82.

Sheokand N, Kumar S, Malhotra H, Tillu V, Raje CI, Raje M (2013) Secreted glyceraldehye-3-phosphate dehydrogenase is a multifunctional autocrine transferrin receptor for cellular iron acquisition. *Biochim Biophys Acta* 1830:3816–3827.

Sibanda BL, Thornton JM (1985) β-Hairpin families in globular proteins. *Nature* 316:170–174.

Simoni RD, Hill RL, Vaughan M (2002) The discovery of the amino acid threonine: the work of William C. Rose [classical article]. *J Biol Chem* 277:E25.

Sirover MA (2014) Structural analysis of glyceraldehyde-3-phosphate dehydrogenase functional diversity. *Int J Biochem Cell Biol* 57:20–26.

Tanford C, Reynolds J (2003) *Nature's Robots: A History of Proteins*. Oxford University Press: Oxford/New York.

Tucker BJ, Breaker RR (2005) Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* 15:342–348.

Vickery HB, Schmidt CLA (1931) The history of the discovery of the amino acids. *Chem Rev* 9:169–318.