# The important properties of proteins and how to explore them

# 5

## 5.1 Introduction: the context for studies and data analysis

**KEY CONCEPT**

- Appreciating the complexity of cellular systems in terms of the numbers of distinct protein species present

*'There's no big mystery to being an enzymologist. All you have to have is a razor blade and a liver'. Gordon Tomkins to Julius Axelrod, circa 1950.*

The characterization of proteins, including enzymatic proteins, requires a good understanding of their important properties and the approaches employed to explore them. This chapter, underpinned by the conceptual toolkit in Chapter 1, aims to enhance your understanding of the goals sought and methods used in characterizing proteins. The simplicity of the approach suggested by Gordon Tomkins fails to convey the challenges associated with applying the key concepts and tools, outlined in Chapters 2–4, to the characterization of proteins within complex biological systems. A typical eukaryotic genome may well have in excess of 10 000 genes which, in turn, can encode probably over 10 times as many distinct proteins, resulting from differential RNA processing and post-translational modification. To establish the role of the several hundred thousand proteins would require the structural and functional characterization of each one. The enormity of this task is compounded by the diverse nature of protein function and structure. In section 5.2, we shall consider how to establish the function and structure of a protein and how its activity may be regulated. In section 5.3, the range of assays used to monitor the biological activity of proteins is outlined; such assays underpin protein characterization and purification. The classical approach to studying proteins requires their purification from their native source (section 5.4.1) using bespoke purification procedures for individual proteins. This task has been simplified greatly with the advent of protein expression in recombinant systems (section 5.4.2). In addition, the overexpression of recombinant proteins has enhanced the structural characterization of proteins, as reflected in the large

As examples, the genomes of the eukaryotic species budding yeast *Saccharomyces cerevisiae*, nematode worm *Caenorhabditis elegans* and human contain about 6000, 12 000 and 23 000 genes, respectively.

increase in the number of structures deposited in the Protein Data Bank (PDB) in recent years. Section 5.5 will present a brief outline of the methods employed to determine the structures of proteins.

Having established the structure and function of a protein, it is important to understand how it is regulated within the cellular environment and the types of interactions in which it is involved. Typically, this is achieved by monitoring the effects of a number of physical and chemical variables on protein activity, as described in sections 5.6 and 5.7. In section 5.8, we shall consider the use of bioinformatics in exploring the properties of proteins and, finally, experimental design will be outlined in section 5.9.

This chapter should lead to a sound understanding of the goals and methods employed to separate, identify, and characterize proteins that will enable data acquisition and handling in the specific examples outlined in subsequent chapters.

As described in section 5.8, the PDB is the database for all three-dimensional structures of proteins solved to atomic resolution by X-ray crystallography or nuclear magnetic resonance (NMR). As of July 2008, there were approximately 47 000 structures in the PDB, of which about 85% had been solved by X-ray crystallography. It should be noted that many structures in the PDB refer to the same protein in different forms or complexes and that relatively few structures of membrane proteins have been determined.

## 5.2 The key questions about a protein

**KEY CONCEPTS**

- Being aware of the range of functions of proteins
- Understanding the levels of protein structure
- Identifying appropriate methods to explore protein structure and protein interactions

All proteins share the common structural feature of being composed of amino acids which are linked by peptide bonds (Chapter 1, sections 1.1–1.3); however, it is the *sequence* of amino acids within a given protein that dictates its unique function and structure. The characterization of a novel protein requires an appreciation of the diversity of protein function and structure. It is also important to appreciate that biological systems are not static entities; they respond to environmental, developmental and metabolic signals with concomitant changes in the structure and function of proteins associated with these processes. Finally, while it is convenient to study proteins in isolation (section 5.4), it must be remembered that they almost always occur within complex cellular environments, interacting with other proteins, metabolites and cellular structures. Thus, to achieve complete characterization, we need to establish how proteins interact with other molecules under physiologically relevant conditions.

### 5.2.1 What is the function of the protein?

Proteins fulfil a diverse range of roles within the cell which can be categorized into the following general groups:

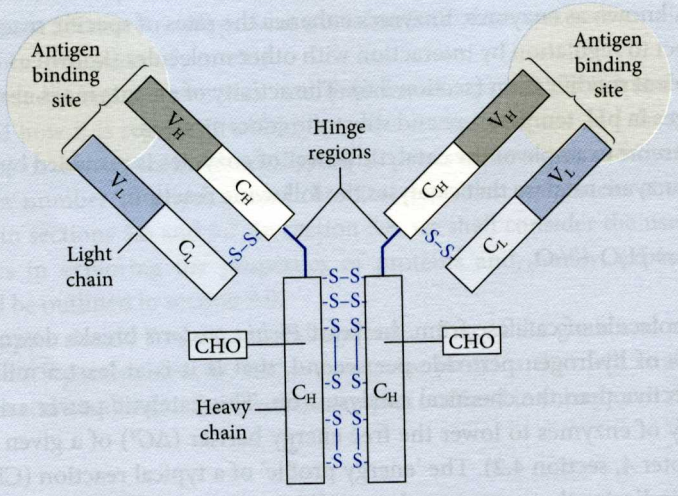



**Fig. 5.1** Representation of general structure of an antibody molecule with the antigen binding sites located at the N-termini of the heavy ($V_H$) and light ($V_L$) chains. The constant regions ($C_H$ and $C_L$) are linked by disulphide bonds (S–) and contain a carbohydrate modification (CHO).

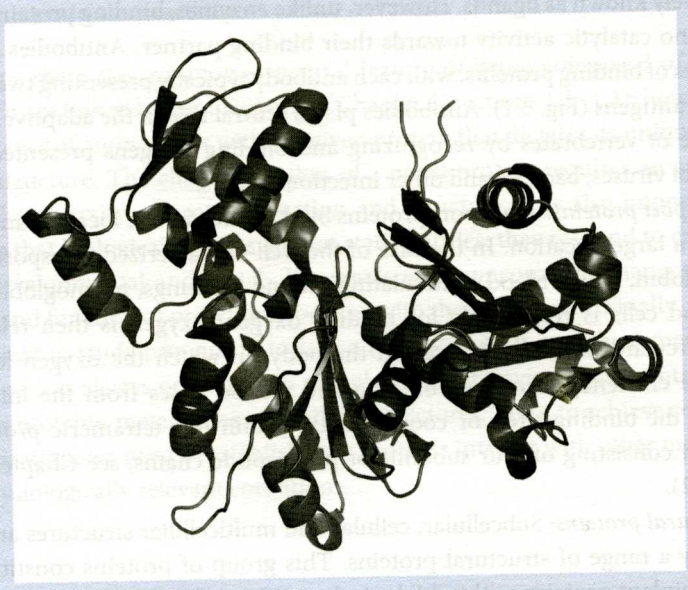


**Fig. 5.2** Structure of a monomeric form of actin (Protein Data Bank code 2A5X). This structural protein usually occurs as part of multimeric structures, such as actin filaments.

*Signalling proteins*: Communication between cells relies on the production of signal molecules in one cell type that are detected by receptors (proteins) located on the surface of a second cell type or target cell; the interaction between any given signal and its receptor leads to a cellular response, a process termed 'signal transduction'. Cellular signal molecules can take the form of small molecules (such as adrenalin, also known as epinephrine) or macromolecules. Insulin, a protein produced in the β cells of the pancreas is one such signal molecule which is detected by insulin receptors located in the cell membrane of many cell types. Insulin is produced in response to high blood glucose levels and the subsequent binding of the hormone to insulin receptors leads to a reduction in blood glucose.

The terms adrenalin(e), which is used in the UK and Europe, and epinephrine (used in the USA) are both derived from the location of the secretory adrenal gland that is adjacent to the kidneys. The roots are: Latin, *ad* (against), *ren* (kidney) or Greek, *epi* (close to), *nephros* (kidney).

*Motor proteins*: Movement within biological systems is a process that is accompanied by the utilization of ATP. A number of motor proteins, including dynein and myosin, can harvest the energy released by ATP hydrolysis to generate movement. Within muscle structures, ATP hydrolysis drives the movement of myosin relative to actin filaments resulting in muscle contraction.

It has been estimated that the actin–myosin system in skeletal muscle can be up to about 60% efficient in converting the chemical energy of the fuel (ATP hydrolysis) into mechanical work. This is more efficient than power stations powered by coal or gas, which can achieve efficiencies in the range 30–40%.

*Storage proteins*: Storage proteins fulfil an essential role within the cell by storing minerals or essentially acting as a source of amino acid nutrients, poised for release in response to an appropriate metabolic or developmental signal. Seed storage proteins are released and degraded on seed germination to provide essential nutrients for the developing seedling. A number of storage proteins, such as the iron storage protein, ferritin, sequester ligands, which may prove toxic to the cell. Iron would tend towards its toxic ferric state within biological systems; however, this essential mineral is stored safely within proteins such as ferritin until it is required for processes such as haem synthesis.

Under physiological conditions, iron can exist in the ferrous ($Fe^{2+}$) or ferric ($Fe^{3+}$) state. The ferric state forms large complexes with anions and hydroxide ions, which are highly insoluble and toxic. The role of ferritin is to sequester iron in the ferric state, complexed with phosphate and hydroxide ions, until it is required for processes such as haem biosynthesis. Ferritin is composed of 24 identical subunits which associate to form a hollow spherical structure. Each 24-mer can accommodate up to 4500 iron ions within this hollow structure (see Fig. 5.3).

In all cases, the function of each of these protein groups is dependent on the structure of the protein, i.e. form fits function: catalytic proteins have residues in and around the active site, which present an environment to promote specific chemical reactions and binding proteins present specific binding sites to allow recognition and binding of target molecules.

Increasingly, the function of a novel protein is determined by establishing its amino acid sequence (either directly using amino acid sequencing or indirectly by translating the sequence of the gene encoding the novel protein) and then conducting a homology search of the ever-expanding sequence databases (section 5.8.5). Confirmation of the probable function requires the purification of the protein and an assay to test its function, e.g. an enzyme assay to measure the activity of the putative catalytic protein. Assays can also be used to test the possible influence of environmental and metabolic effectors on the activity of the protein. A complete understanding of protein function requires solving (or prediction) of its structure.

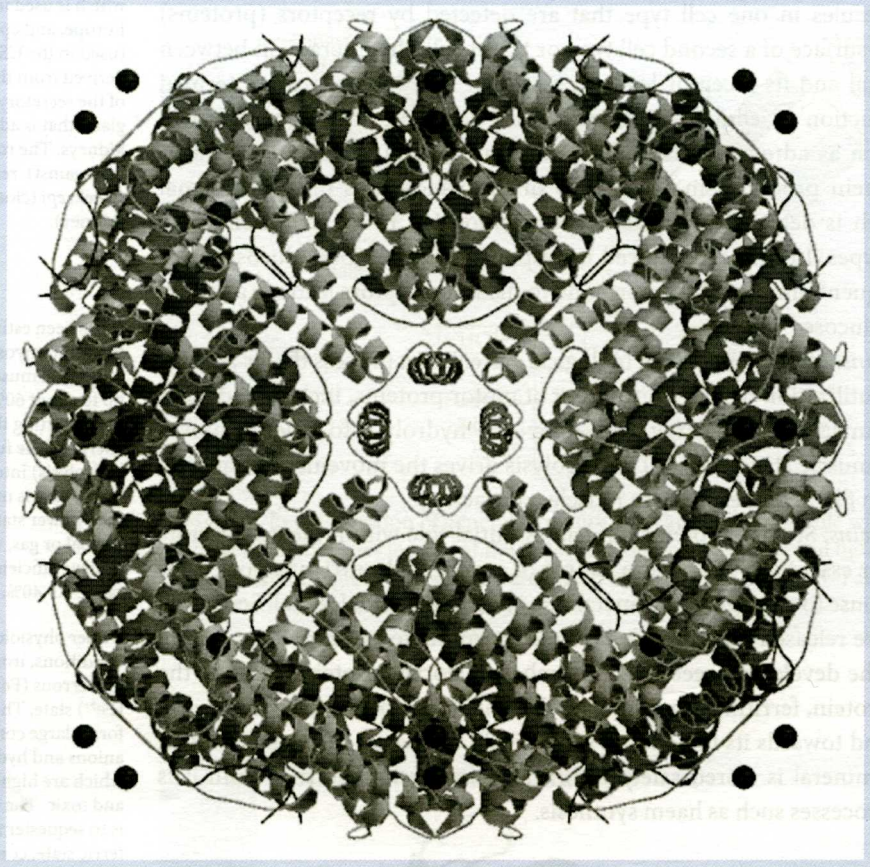


**Fig. 5.3** 24-mer structure of the iron storage protein ferritin (Protein Data Bank code 1FHA). 24 identical subunits pack together to form a hollow shell that can accommodate up to 4500 iron ions in the ferric ($Fe^{3+}$) state.

### 5.2.2 What is the structure of the protein?

Over the past 20 years, our understanding of protein structure has been greatly enhanced by the near exponential growth in the number of structures that have been solved (see Fig. 5.4). A number of factors have contributed to this growth, including: the ability to overexpress many proteins (which has provided the quantity and quality of material required for structural studies), enhanced computing capabilities and developments in the biophysical techniques employed to determine protein structure. Complete structural characterization of a protein requires:

- *Determination of the amino acid sequence*: This may be deduced from the nucleotide sequence of a gene encoding a particular protein or from direct amino acid sequencing (Chapter 8, section 8.3). The amino acid sequence of a protein can be used to generate a wealth of structural information, including the

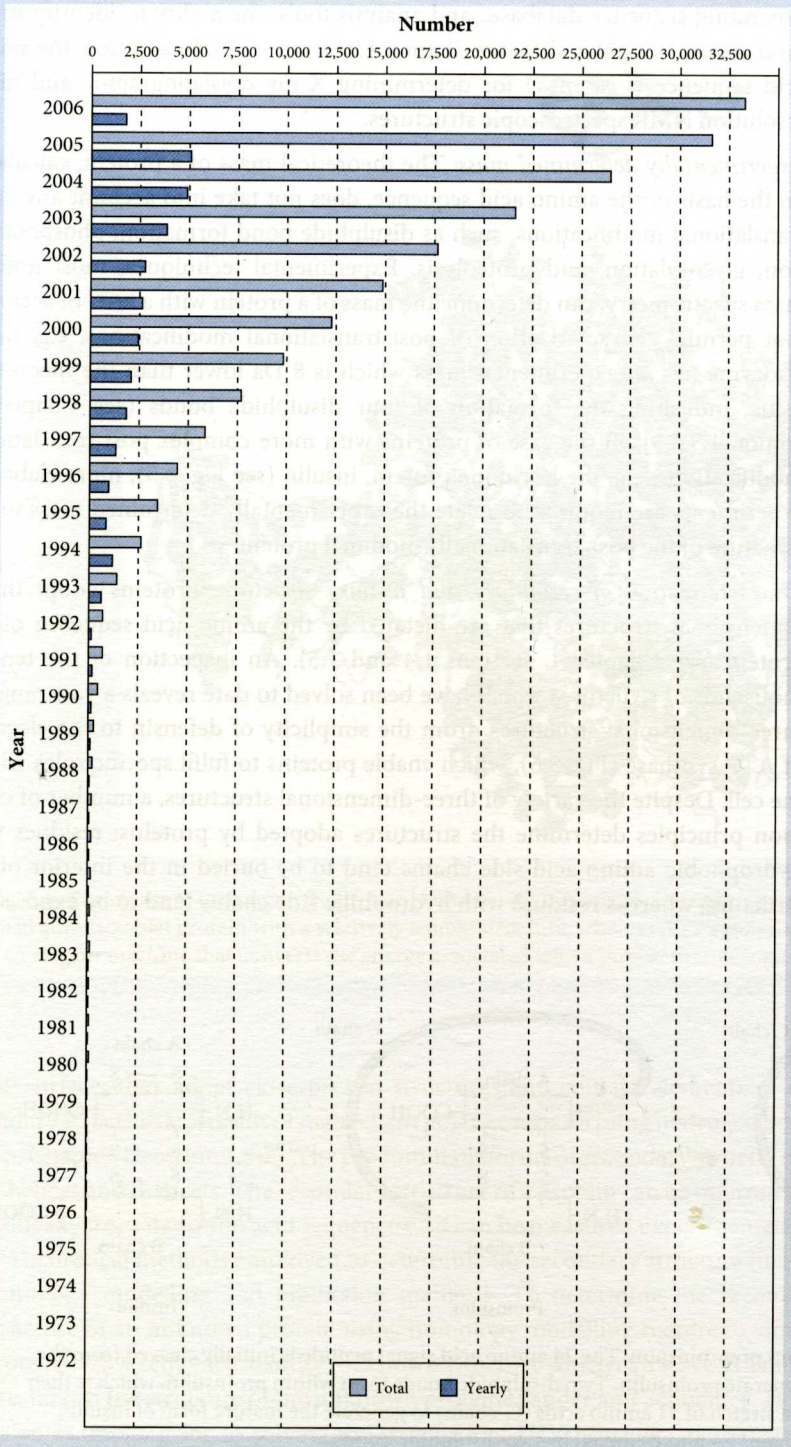



**Fig. 5.4** Growth of protein structures solved over the past 35 years.

Many of the DNA-derived protein sequences in the databases correspond to proteins of unknown function or proteins which may not be found in cells, certainly under normal growth conditions.

theoretical mass, potential post-translational modification sites and, with ever-expanding sequence databases and analysis tools, the ability to identify structural and functional motifs (see Chapter 1, section 1.3.2). In addition, the amino acid sequence is essential for determining X-ray crystallographic and high-resolution NMR spectroscopic structures.

- *Experimentally determined mass*: The theoretical mass of a protein, calculated on the basis of the amino acid sequence, does not take into account any post-translational modifications, such as disulphide bond formation, phosphorylation, glycosylation, and proteolysis. Experimental techniques, most notably mass spectrometry, can determine the mass of a protein with a level of accuracy that permits characterization of post-translational modifications, e.g. horse lysozyme has an experimental mass which is 8 Da lower than the theoretical value, indicating the formation of four disulphide bonds (see Chapter 1, section 1.3.2.5). In the case of proteins with more complex post-translational modifications, e.g. the hormone protein, insulin (see Fig. 5.5), more elaborate experiments are required to relate the experimentally determined mass to the structure of the post-translationally modified protein.
- *Characterization of secondary and tertiary structure*: Proteins adopt three-dimensional structures that are dictated by the amino acid sequence of the protein (see Chapter 1, sections 1.4 and 1.5). An inspection of the tens of thousands of structures which have been solved to date reveals a vast range of three-dimensional structures, from the simplicity of defensin to the elegance of ATP synthase (Fig. 5.6), which enable proteins to fulfil specific roles within the cell. Despite the variety of three-dimensional structures, a number of common principles determine the structures adopted by proteins: residues with hydrophobic amino acid side chains tend to be buried in the interior of the structure, whereas residues with hydrophilic side chains tend to be exposed on

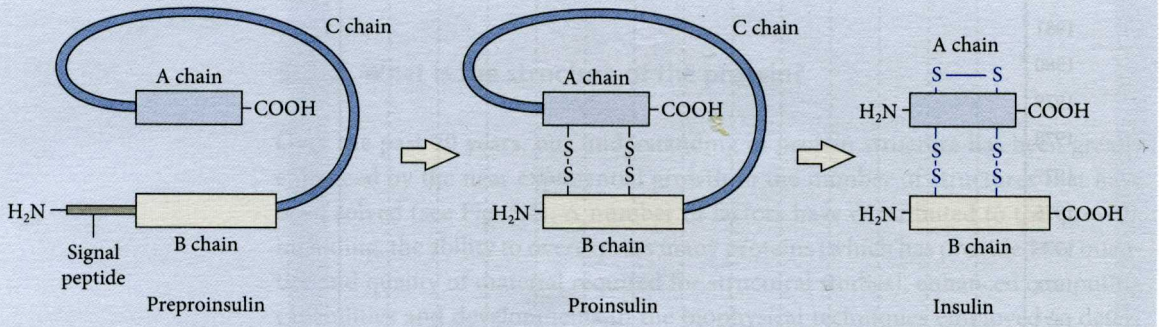



**Fig. 5.5** Maturation of insulin from preproinsulin. The 24 amino acid signal peptide is initially cleaved from the N-terminus of preproinsulin to generate proinsulin. Two disulphide bonds form within proinsulin, which is then proteolytically cleaved to remove a stretch of 51 amino acids (C-chain) to generate the mature form of insulin, which contains two polypeptide chains (A and B) linked by two disulphide bonds.

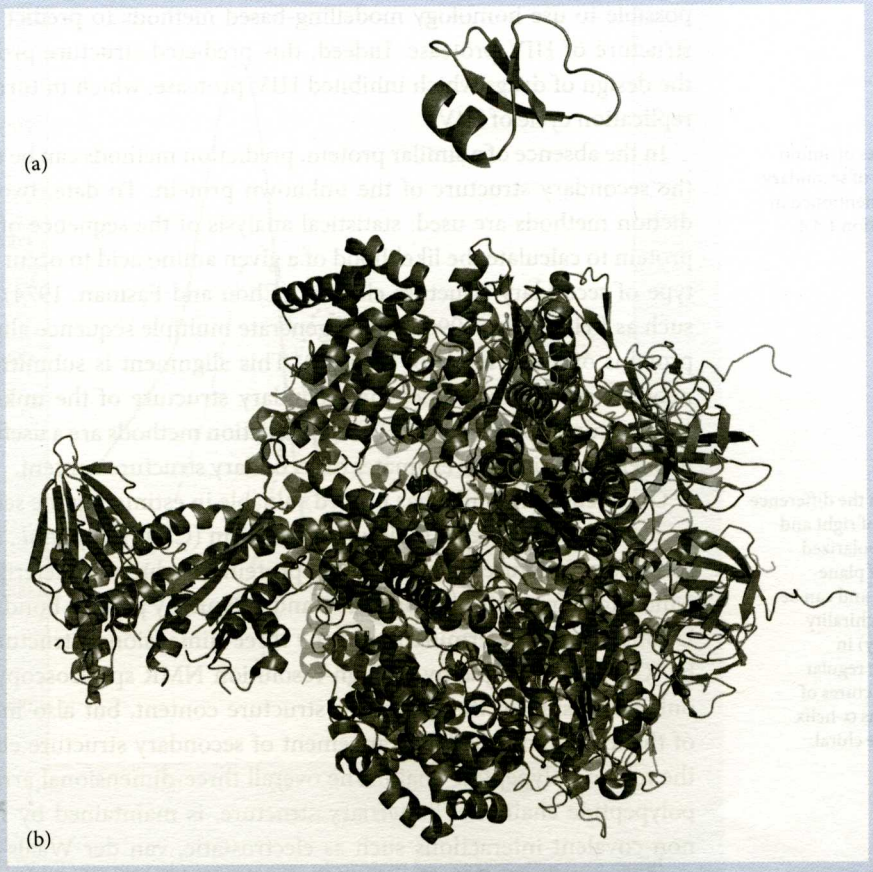



**Fig. 5.6** Structures of (a) human β-defensin-1 and (b) ATP synthase. Defensin (Protein Data Bank code 1IJU) is a small antimicrobial protein with a relatively simple structure, whereas ATP synthase (Protein Data Bank code 1QO1) is an elegant machine that converts the energy associated with a proton motive force to generate ATP.

the surface; they adopt close-packed structures and contain elements of secondary structure as a result of main chain polar groups forming hydrogen bonds (see Chapter 1, section 1.5.1). The predominant forms of secondary structure are α-helices and β-sheets. The secondary structure of a protein can be inferred theoretically from its amino acid sequence or it can be measured experimentally.

Theoretical methods employed to determine the secondary structure include homology modelling and prediction methods. To determine the secondary structure of an unknown protein using homology modelling requires a similar protein (>25% sequence identity) of known structure which can serve as a model to establish the structure of the unknown protein (Sander and Schneider, 1994). As an example, prior to solving the structure of HIV protease, the structures of aspartic proteases from a number of sources had been determined. As HIV

protease shared >25% sequence identity with these aspartic proteases, it was possible to use homology modelling-based methods to predict the secondary structure of HIV protease. Indeed, this predicted structure proved pivotal in the design of drugs which inhibited HIV protease, which in turn inhibited the replication cycle of HIV.

The preferences of amino acids for types of secondary structure are mentioned in Chapter 1, section 1.4.4.

In the absence of a similar protein, prediction methods can be used to predict the secondary structure of the unknown protein. To date, two types of prediction methods are used: statistical analysis of the sequence of the unknown protein to calculate the likelihood of a given amino acid to occur in a particular type of secondary structure element (Chou and Fasman, 1974) or techniques such as PHD (Rost, 1996), which generate multiple sequence alignments (with proteins of lower levels of identity). This alignment is submitted to a neural network system to predict the secondary structure of the unknown protein. With accuracies of up to 72%, such prediction methods are a useful tool to complement experimental estimates of secondary structure content.

CD is based on the difference in absorption of right and left circularly polarized components of plane-polarized light and can be used to detect chirality (optical activity) in molecules. The regular secondary structures of proteins such as α-helix and β-sheet are chiral.

One technique which has proved valuable in estimating the secondary structure content of proteins is circular dichroism (CD) (Kelly *et al.*, 2005). Regular secondary structure elements within proteins produce characteristic CD spectra which arise from absorption at 190 and 220 nm by peptide bonds (see Fig. 5.7). Ultimately, the determination of the three-dimensional structure of a protein by X-ray crystallography or high-resolution NMR spectroscopy provides not only a measure of the secondary structure content, but also molecular detail of the length and spatial arrangement of secondary structure elements within the folded polypeptide chain. The overall three-dimensional arrangement of a polypeptide chain, i.e. its tertiary structure, is maintained by multiple weak, non-covalent interactions such as electrostatic, van der Waals', hydrophobic interactions, and hydrogen bonds (see Chapter 1, section 1.5). These interactions are also involved in maintaining subunit–subunit contacts, i.e. the quaternary structure of proteins.

- *Quaternary structure*: In general, larger proteins (typically >50 kDa) tend to exist as multiple subunits giving rise to the level of structure known as the quaternary structure. The quaternary structure can range in complexity from two identical subunits, e.g. ribulose-bisphosphate carboxylase (Rubisco: EC 4.1.1.39) from photosynthetic bacteria (2 × 55 kDa) to multiple non-identical subunits, e.g. Rubisco from plants and algae which exists as hexadecamer with eight large subunits and eight small subunits (8 × 55 kDa and 8 × 15 kDa). Characterization of multi-subunit proteins requires determination of the overall molecular mass of the protein, identification of the types of subunits and the molecular mass of each type, calculation of the number of each subunit type within the protein and the structural arrangement of the subunits. The molecular mass of multi-subunit proteins must be determined under non-denaturing conditions to maintain subunit–subunit interactions, using techniques such as gel filtration chromatography and ultracentrifugation (Chapter 8, section 8.2). Analysis by SDS-PAGE

Rubisco catalyses the first dark reaction of photosynthesis (i.e. the addition of $CO_2$ to the five-carbon sugar ribulose bisphosphate). It is a relatively inefficient enzyme in catalytic terms and is thought to be the most abundant protein on earth.

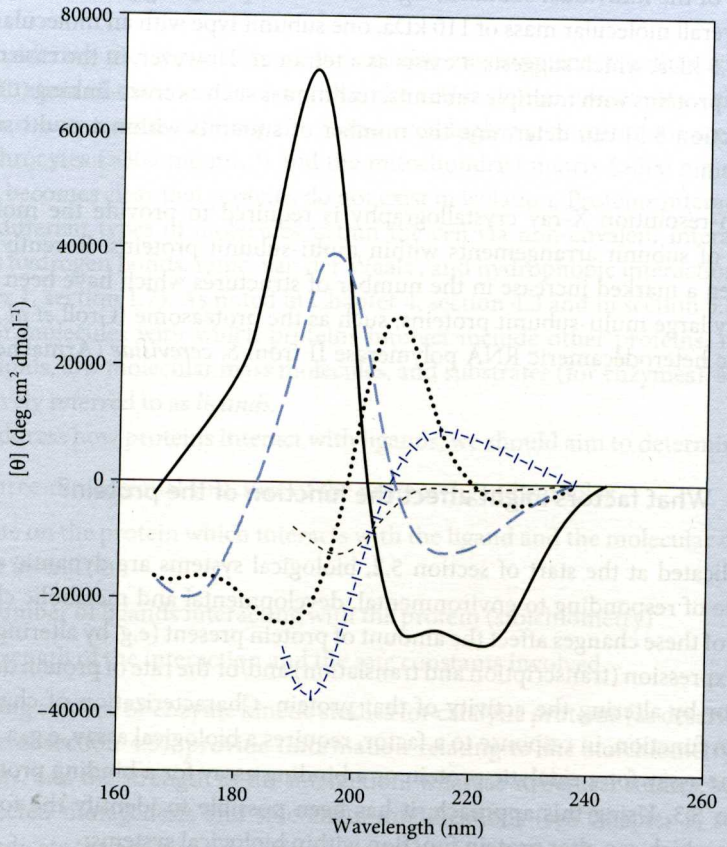



**Fig. 5.7** Far-UV CD spectra of various types of protein secondary structure. Solid line, α-helix; long dashed line, anti-parallel β-sheet; dotted line, type I β-turn; cross dashed line, extended $3_1$-helix or poly (Pro) II helix; short dashed line, irregular structure.

(see Chapter 8, section 8.2.1) can reveal the molecular mass of individual subunits and the number of subunit types within a multi-subunit protein: if only one species appears on an SDS-PAGE gel, it usually indicates that only one type of subunit exists within the protein; two species can suggest two subunit types, and so on. It is worth noting that there may be instances when non-identical subunits may have similar mobilities on SDS-PAGE (i.e. subunits of similar size but different amino acid sequence appear as one band) and further analysis, such as mass spectroscopy or amino acid sequencing, will be required to confirm whether all subunits are identical or not. It is often possible to determine the number of subunits within a multi-subunit protein knowing the molecular mass of the multi-subunit protein, the number of subunit types and the molecular

It should be noted that SDS-PAGE is not a high-resolution technique and it is doubtful if two proteins which had subunits of similar molecular masses (within 2%) would be efficiently separated by this technique. Mass spectrometry (see Chapter 8, section 8.2.5) would be useful for resolving proteins of similar masses.

mass of the individual subunits, e.g. *S. cerevisiae* phosphoglycerate mutase has an overall molecular mass of 110 kDa, one subunit type with an molecular mass of 27.5 kDa, which suggests it exists as a tetramer. However, in the case of very large proteins with multiple subunits, techniques such as cross-linking (Chapter 8, section 8.5) can determine the number of subunits within a multi-subunit protein.

High-resolution X-ray crystallography is required to provide the molecular details of subunit arrangements within multi-subunit proteins. Recently, there has been a marked increase in the number of structures which have been solved for very large multi-subunit proteins, such as the proteasome (Groll *et al.*, 2001) and the heterodecameric RNA polymerase II from *S. cerevisiae* (Armache *et al.*, 2005).

### 5.2.3 What factors might affect the function of the protein?

The turnover rates of different proteins in eukaryotic cells vary enormously, with some proteins having a very short half-life of only a few minutes, whereas others exist within the cell for weeks. The half-life of a protein depends on its function, e.g. proteins involved in the regulation of cell division and transcription have a short half-life (minutes), whereas central metabolic enzymes have a long half-life (several days).

As indicated at the start of section 5.2, biological systems are dynamic and are capable of responding to environmental, developmental and metabolic changes. Many of these changes affect the amount of protein present (e.g. by altering rate of gene expression (transcription and translation) and/or the rate of protein degradation) or by altering the activity of that protein. Characterization of changes in protein function, in response to a factor, requires a biological assay, e.g. a specific enzyme assay for a catalytic protein or a binding assay for a binding protein, see section 5.3. Using this approach, it has been possible to identify the following factors which can alter protein function within biological systems:

*Non-covalent binding of other molecules* ranging from small molecules to regulatory protein subunits and specific macromolecules, collectively known as *effectors*. It is possible to characterize these interactions *in vitro* using binding assays (see Chapter 6, section 6.4) coupled with appropriate analysis of the saturation curves (see Chapter 4, section 4.3).

*Availability of ligands*, including substrates for enzymes or cofactors, which can vary greatly within the cell. This effect is most obvious when varying the substrate concentration at values close to the Michaelis constant ($K_m$), which can lead to sizeable changes in activity (see Chapter 4, section 4.3.3 and Fig. 4.8).

For example, antibodies are available commercially that can distinguish between phosphorylated side chains of serine and tyrosine in proteins.

*Reversible covalent modification*, involving the addition and removal of specific chemical groups, can have a dramatic effect on the property of a protein. Many of these modifications are listed in Chapter 1, section 1.3.2.5 and can be readily identified by mass spectrometry or binding of antibodies which recognize specific post-translational modification groups such as the phosphoryl group.

*Irreversible covalent changes*, including the targeted proteolytic cleavage of inactive precursors to generate functionally active proteins, can be characterized using SDS-PAGE, size exclusion chromatography and mass spectrometry.

### 5.2.4 How does the protein interact with other molecules?

By considering the complexity of the cellular environment, together with the concentration of macromolecules within the cell (a typical prokaryotic cell has a protein concentration >200 mg $mL^{-1}$, and even higher concentrations are found in erythrocytes (>300 mg $mL^{-1}$) and the mitochondrial matrix (>500 mg $mL^{-1}$)), it soon becomes clear that proteins do not exist in isolation. Proteins interact with many different types of molecules within the cell via non-covalent interactions such as hydrogen bonds, ionic, van der Waals', and hydrophobic interactions (see Chapter 1, section 1.7). As noted in Chapter 4, section 4.3 and in section 5.2.1 the types of molecules with which proteins interact include other proteins, nucleic acids, lipids, low molecular mass molecules, and substrates (for enzymes), and are collectively referred to as *ligands*.

To address how proteins interact with ligands, we should aim to determine:

- the three-dimensional structure of the protein–ligand complex
- the site on the protein which interacts with the ligand and the molecular details of the interaction
- the number of ligands interacting with the protein (stoichiometry)
- the strength of the interaction and the rate constants involved.

Binding studies, or enzyme kinetic studies for catalytic proteins (as described in Chapter 4, section 4.3), provide information relating to the stoichiometry, rate constants, and the strength of the interaction, whereas structural studies, such as site-directed mutagenesis and side chain modifications (see Chapter 9, section 9.10), indicate the amino acids that are important for protein–ligand interactions.

## 5.3 Assays for biological activity

**KEY CONCEPT**

- Appreciating the range of assays for biological activity measurements and their limitations

A specific assay is required for purification of a protein and to address key questions, relating its structure and function. In general, an assay is used during protein purification to gauge the success of the purification protocol, i.e. enhancement of the specific activity and maintenance of a high yield of biologically active protein (Chapter 3, section 3.9). Assays are also used to assess the effect of factors which may influence the biological activity of a protein, providing information relating to the function and structure of the protein. A good assay is one which is quick, simple to conduct, highly specific for the protein of interest, and relatively inexpensive. Before looking at some specific examples of different assay types,

we must consider the care which must be exercised when interpreting assay data in general:

*Is the activity measurement due solely to the presence of the protein of interest?* There may be other factors present which may contribute to the activity measurement. A series of control assay measurements in the absence of biologically active protein will indicate whether the assay components make a contribution to the measured response, e.g. an increase in absorbance arising from non-enzyme-catalysed substrate degradation in an enzyme assay, or non-specific binding partners in immunoblots and ELISAs. The quality of assay components is critical to the success of the assay. A good example of this is provided by the enzyme-coupled assay for the glycolytic enzyme, phosphoglycerate mutase (PGAM): 3-phosphoglycerate is converted to 2-phosphoglycerate by PGAM and then 2-phosphoglycerate is then converted to phosphoenolpyruvate (PEP) by the coupling enzyme enolase (see Fig. 5.8). The assay is started by the addition of PGAM, with the resultant formation of phosphoenolpyruvate (PEP) monitored by measuring the increase in absorbance at 240 nm. During a study of yeast PGAM, it was noted that prior to the addition of this enzyme to the assay, some unexpectedly high-activity measurements were obtained. Further analysis revealed that the commercial preparation of enolase (purified from horse liver) was contaminated with PGAM.

In a coupled assay, the coupling enzyme(s) is (are) added in a large excess, perhaps 50-fold more than the enzyme being assayed. Any coupling enzymes should therefore be extremely pure.

*Is the activity measurement proportional to the amount of protein present?* If twice as much protein is added to the assay, is a doubling of the activity observed? If the answer is no, this may reflect the fact that some component of the assay is limiting the activity measurements.

Such limiting factors could include an insufficiency of a coupling enzyme, or that the reaction occurs too quickly for the detection system to give an accurate measurement of the rate.

The most convenient type of assay is the continuous assay in which a response is measured directly, e.g. change in absorbance, fluorescence or pH, following the addition of protein and the response can be recorded continuously throughout the course of the reaction. Less convenient, but no less useful, are discontinuous assays that involve reactions which are initiated by the addition of a protein and then samples are removed at specific time intervals. These samples are subsequently quenched (i.e. biological activity is stopped) and analysed.

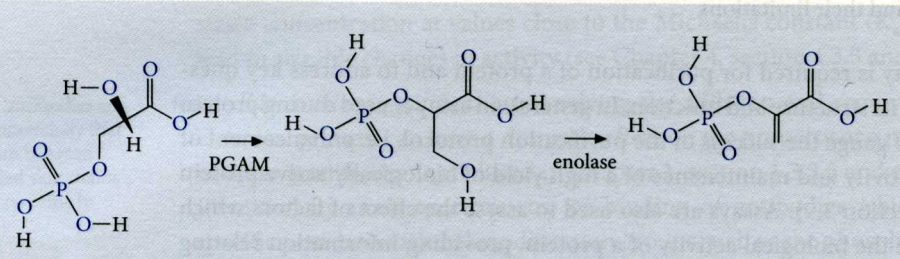



**Fig. 5.8** Phosphoglycerate mutase (PGAM) assay contains 3-phosphoglycerate and the coupling enzyme enolase. The reaction is initiated by the addition of PGAM, which converts 3-phosphoglycerate to 2-phosphoglycerate and then enolase converts this to phosphoenol pyruvate (PEP), which is detected at 240 nm.

### 5.3.1 Catalytic proteins (enzymes)

The aim of assaying an enzyme is to measure the rate of product formation or substrate utilization. This typically relies on a difference in the spectroscopic properties between the substrate and the product. For example, in the following reaction catalysed by alcohol dehydrogenase (EC 1.1.1.1):

ethanol + $NAD^+$ → acetaldehyde (ethanal) + NADH + $H^+$

the reduced form of nicotinamide dinucleotide coenzyme (NADH) absorbs radiation at 340 nm, however, the oxidized form ($NAD^+$) does not, allowing the direct continuous monitoring of product formation. Not all enzyme-catalysed reactions have natural reactants and products with suitable spectroscopic properties. In such cases, synthetic chromogenic substrates may prove useful, e.g. the enzyme 4-nitrobenzyl esterase, which is used in the synthesis of the antibiotic, Loracarbef, catalyses a reaction which produces little spectroscopic change. However, a nitrophenyl derivative of the substrate generates the product 4-nitrophenol, which is readily detected at 400 nm.

Loracarbef is s cephalosporin-derived antibiotic. During the synthesis of loracarbef, carboxyl groups are protected with 4-nitrobenzyl alcohol, which is subsequently removed by the enzyme 4-nitrobenzyl esterase.

Alternatively, the reaction of interest can be coupled to a second reaction, which will produce a spectroscopic change, e.g. the assay for hexokinase (EC 2.7.1.1) involves the addition of the coupling enzyme glucose-6-phosphate dehydrogenase (EC 1.1.1.49).

Hexokinase:

ATP + D-glucose → ADP + D-glucose 6-phosphate

Glucose-6-phosphate dehydrogenase:

D-glucose 6-phosphate + $NADP^+$ → D-glucono-1,5-lactone 6-phosphate + NADPH + $H^+$

Whilst the reaction catalysed by hexokinase produces no spectroscopic change, glucose-6-phosphate dehydrogenase generates NADPH which absorbs at 340 nm, providing a indirect continuous assay for hexokinase activity.

Alternative detection methods, such as change in pH, may provide a means of monitoring enzyme activity.

### 5.3.2 Binding proteins

Protein binding assays can be used to both identify ligands and to characterize the nature of the protein–ligand interaction. Binding assays can be subdivided into two categories: those which are based on biophysical changes in the protein or ligand upon protein–ligand complex formation and those which employ direct quantitation of free and bound ligand. Biophysical changes can be monitored using a range of spectroscopic methods, such as absorbance, fluorescence, CD and NMR. These techniques rely on significant changes in the spectra of the unbound

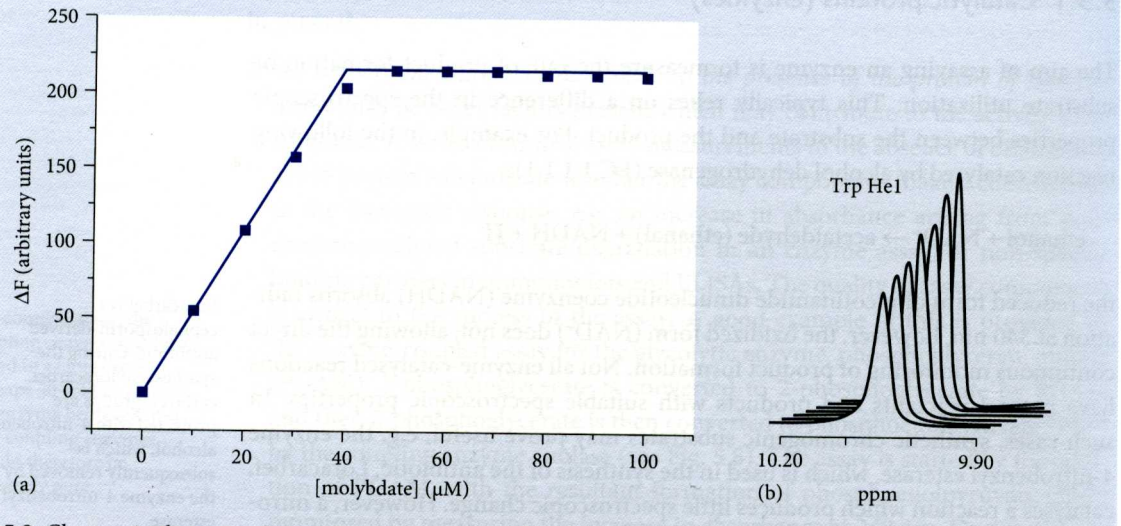



**Fig. 5.9** Changes in the spectral properties of proteins on the addition of a ligand. (a) Change in fluorescence. Binding of molybdate ions to the molybdate-sensing protein ModE from *E. coli*. Aliquots of ligand were added to a solution of protein (40 μM) and changes in the fluorescence at 350 nm were monitored. Saturation occurs at a concentration of 40 μM ligand, showing that there is one binding site per polypeptide chain. Binding of ligand leads to an approximately 50% quenching in fluorescence (Boxer *et al.*, 2004). (b) Chemical shift in NMR spectrum. This stack plot shows the spectral effects on addition of increasing amounts of Cu(II) on the tryptophan residue within a prion protein peptide (PHGGGWGQ). The $CuSO_4$ was added in aliquots of 0.0033 mole-equivalents up to 0.02 mole-equivalents (Viles *et al.*, 1999).

The chromatin-associated protein Hbsu from *Bacillus subtilis* binds DNA in a sequence-independent manner and is important in bacterial nucleoid formation. Wild-type Hbsu does not contain any Trp residues and by introducing a Trp at position 47 it was possible to generate a mutant form of the protein, which was indistinguishable from the wild type while providing a spectroscopic means for determining dissociation constants (Groch *et al.*, 1992).

protein or ligand and the complexed state. In titration studies, where the degree of spectral change is assumed to be directly proportional to the ligand concentration, it is possible to determine the degree of binding (see Fig. 5.9). In the absence of a suitable intrinsic spectroscopic change, it is possible to design chemically modified versions of the protein or ligand to facilitate a simple binding assay. A variety of means can be used to introduce spectroscopic labels into proteins. For example, site-directed mutagenesis can be employed to substitute a non-fluorescent amino acid by the fluorescent tryptophan (Trp) at a selected position in the protein.

An alternative approach to introducing fluorophores involves *in vitro* chemical modification of amino acids; a fluorescent group can be introduced by reaction of a suitable reagent with Cys side chains, for example the introduction of the fluorescent probe IAEDANS to Trp repressor protein mutants to characterize tryptophan and DNA binding (Chou and Matthews, 1989). Recent advances in bacterial and yeast expression systems allow efficient site-specific introduction of unnatural amino acids *in vivo*; use of engineered tRNA and aminoacyl tRNA synthase within these systems permits the introduction of a range of unnatural amino acids, including fluorophores (Magliery, 2005). Site-specific, *in vivo*, incorporation of unnatural fluorescent amino acids has been used to modify green fluorescent protein (GFP) at residue 66. The mutant GFPs were successfully overexpressed and purified and were found to have unique spectral properties (Wang *et al.*, 2003).

Another possibility is to introduce the Trp analogue 7-azaTrp in place of Trp in an expressed protein by growth of the host organism on a medium containing this amino acid, with the biosynthetic pathway for Trp inhibited. 7-azaTrp has spectroscopic properties which can be readily distinguished from those of Trp.

Replacement of all tryptophan residues in λ bacteriophage lysozyme with 7-aza Trp has been used to probe the structure and function of this enzyme. In addition, the 7-aza Trp-modified version of lysozyme facilitated its successful crystallization under the microgravity conditions on space shuttle flights (Evrard *et al.*, 1998).

In the case of proteins that undergo significant conformational changes as a result of ligand binding, it may be possible to monitor these changes by determining the sedimentation coefficients using ultracentrifugation.

Traditional direct quantitation methods rely on partitioning techniques, such as equilibrium dialysis and membrane filtration, in which the protein and bound ligand are separated from free ligand. A typical dialysis binding assay would involve placing the protein within a dialysis membrane, which is then placed in a solution of ligand. At equilibrium, the concentration of free ligand will be the same inside and outside the dialysis membrane. The concentration of bound ligand can either be calculated from the difference between the free ligand concentration at the start of dialysis and the free ligand concentration at equilibrium, or from the measured concentrations of ligand on the protein side of the membrane (free ligand plus bound ligand) and on the other side of the membrane (free ligand). Similarly, the ability of membrane filtration to retain protein and protein complexed with ligand, but not free ligand, can be exploited to detect ligand binding. This simple technique requires a means of measuring the amount of ligand retained by the filter, i.e. complexed with protein.

Binding studies using equilibrium dialysis or membrane filtration are greatly facilitated if the ligand has some convenient spectroscopic property or is radioactively labelled.

More recently, the use of solid phase techniques, such as surface plasmon resonance (SPR), have been employed to detect and characterize protein–ligand interactions. This technique is based on the detection of an increase in mass resulting from protein–ligand complex formation. A typical assay would involve immobilization of the protein on the surface of a sensor, followed by introduction of a ligand. Protein–ligand interactions, resulting in an increase in mass, give rise to an increase in signal. Likewise, dissociation of a ligand from immobilized protein results in a decrease in mass, producing a decrease in signal (see Fig. 5.10). Thus, solid phase techniques are proving useful in identifying potential ligands and in determining the kinetic and affinity properties in binding assays (see Chapter 10, section 10.7).

Essentially, the SPR technique measures the rate constant for the association (complex formation) and dissociation (complex breakdown) steps. As described in Chapter 4, section 4.3, the equilibrium constant can be derived from the ratio of these rate constants.

### 5.3.3 Transport proteins

There is a whole family of glucose transporter (GLUT) proteins which share some common structural features, but have different tissue distributions and affinities for glucose. For example, GLUT2 has a low affinity for glucose and is found in the liver and pancreas. GLUT4 has a higher affinity for glucose and is stimulated by insulin; it plays a particularly important role in glucose uptake by muscle and adipose tissue.

Transport protein assays are designed to measure the rate of transport of a ligand from one location to another, e.g. transport of glucose into erythrocytes by an integral membrane glucose transporter. While binding assays (see section 5.3.2) on purified transport proteins provide information relating to the stoichiometry and affinity of the protein–ligand interaction, they do not necessarily provide a measure of transport. *In vivo* and *in vitro* transport assays can involve relatively complex systems, such as whole cells or proteoliposomes, and require some mechanism to monitor the initial ligand concentrations in one location and ligand concentration in its final destination. One of the most direct methods employed to

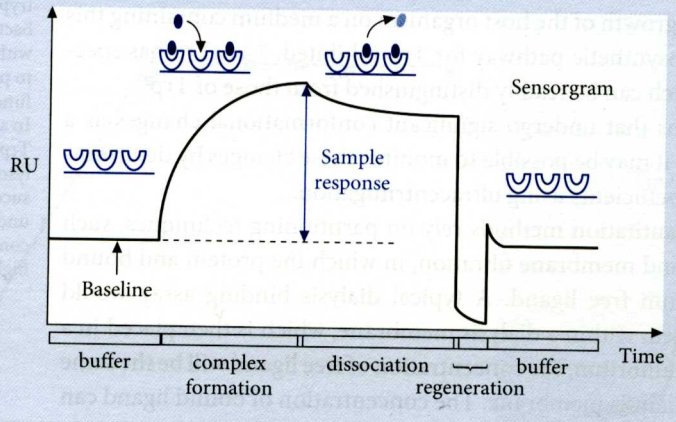



**Fig. 5.10** A typical surface plasmon resonance (SPR) sensogram. A baseline signal is produced by the continuous flow of buffer over the protein immobilized on the surface of a sensor chip. Introduction of ligand into the flow of buffer may result in an association of the ligand with the immobilized protein and this is detected by an increase in signal. Subsequent removal of the ligand and a return to the continuous flow of buffer alone will generate a decrease in signal which reflects the dissociation of the ligand from the immobilized protein. The rate constants for the association and dissociation steps are obtained by curve-fitting procedures. SPR occurs when light is reflected off thin metal films. SPR has been exploited to detect protein–ligand interactions by immobilizing a protein on the surface of a thin metal film (sensor chip) and subsequently adding ligands to produce a change in mass (protein plus ligand), which in turn alters the surface of the metal film. Surface changes produce a change in the angle of the reflected light, i.e. a change in the SPR signal. RU, response units.

measure the activity of transport proteins involves the use of isotopically labelled ligands. In a typical assay, a known amount of isotopically labelled ligand is added to the system and incubated for a fixed period of time, after which the reaction is stopped by introducing an inhibitor or by rapid isolation of the final destination of the ligand, e.g. isolation of cells or proteoliposomes using centrifugation or size exclusion chromatography. Functional studies of transport proteins require assays to be conducted in the presence of varying amounts of ligand and in the presence or absence of effectors (see section 5.2.3).

### 5.3.4 Other types of proteins

This section will describe how to test the biological function of proteins which cannot be measured using the more conventional assays outlined in the previous sections. GFP, which occurs naturally in the jellyfish *Aequorea victoria* and has been used as a reporter molecule in many prokaryotic and eukaryotic systems, is assayed based on its ability to fluoresce at 510 nm, following excitation at 395 nm. A more unusual assay has been developed for the taste-modifying glycoprotein, miraculin. Miraculin, which is isolated from the red berries of the West African

shrub, *Richadella dulcifica*, has the unusual property of making sour tastes seem sweet. The sweet-inducing activity of miraculin is measured by administering a small amount of miraculin to subjects, followed by sour citric acid solutions; the subjects then assign an apparent sweetness value to the citric acid solutions (Theerasilp and Kurihara, 1988).

One final example of less conventional assays is that used to monitor the effects of cytokines. Cytokines are a family of proteins, secreted primarily by leukocytes, which allow cell–cell communication. Cytokines are often assayed by monitoring the effects they have on cell cultures, such as cell proliferation, differentiation, and stimulation of immune functions.

The cytokines are a diverse group of signalling proteins that include interferons, several interleukins, and a range of growth factors. Cytokines are secreted by many different cell types which then bind to specific receptor proteins located on the surface target cells, resulting in a biological response. Each cytokine acting on specific target cells produces a specific response that may be cell differentiation, growth, or tissue development.

In the case of proteins with biological functions which cannot be measured easily using direct assays, the advent of heterologous expression systems has presented a convenient means of monitoring the purification of such proteins, circumventing the need for an assay. Overexpressed proteins can account for a substantial proportion of the total cell protein (see section 5.4.2) and as a result can be readily identified by measurements of molecular mass using SDS-PAGE; the most abundant species with the correct molecular mass (theoretical or known) can be identified in crude extracts and in samples throughout the purification stages (see Fig. 5.11). When the protein with the correct mass is purified, its identity should be confirmed by mass spectrometry and by partial amino acid sequencing or peptide mass fingerprinting (Chapter 8, section 8.3).

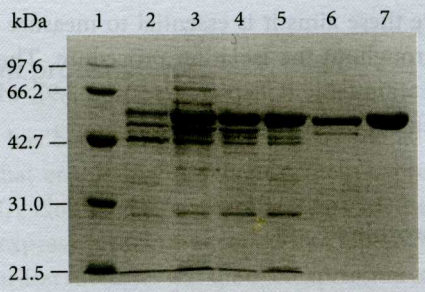



**Fig. 5.11** SDS-PAGE analysis of tomato leucine aminopeptidase (LAP-A) overexpression and purification. LAP-A was overexpressed in *E. coli* and purified using a four-step procedure, during which LAP-A could be followed as the most intense band, measuring 55 kDa on SDS-PAGE analysis (from Gu *et al.*, 1999) Lane 1, protein molecular weight markers; lane 2, *E. coli* lysate prior to IPTG induction; lane 3, *E. coli* lysate following IPTG induction; lane 4, heat denaturation; lane 5, ammonium sulphate precipitation; lane 6, MonoQ chromatography; lane 7, hydrophobic interaction chromatography.

## 5.4 Purification of proteins

**KEY CONCEPTS**

- Knowing the objectives of protein purification
- Being aware of the experimental considerations required to meet these objectives

As outlined in the previous sections, structural and functional studies can yield information about the important properties of proteins. This information and its interpretation are simplified greatly by studying purified proteins, i.e. structural and functional properties derived from experiments involving 'protein X' can be attributed to 'protein X' alone and not to some other molecular species. The aims of any protein purification scheme are:

*To retain maximum biological activity*: The success of structural and functional studies hinges on the purified protein behaving in a manner similar to its behaviour within the cell, i.e adopting its native structure and being fully functional.

*To ensure the protein is indeed pure*: When characterizing a protein it is essential that the properties measured can be attributed to the protein of interest and are not due to the presence of contaminating macromolecules.

*To maximize the amount of protein recovered*: An efficient purification scheme will recover as much biologically active, pure protein as possible from the source material.

In order to achieve these aims it is essential to measure protein content and biological activity throughout the isolation procedure. These data can be used to construct a purification table (Chapter 3, section 3.9), which will give a clear indication of the efficiency of the procedure.

### 5.4.1 Wild-type proteins

The isolation of proteins from their natural source exploits their heterogeneous biochemical properties. A range of purification techniques have been developed to isolate individual proteins according to their unique set of biochemical characteristics, encompassing molecular mass, charge, stability, hydrophobicity, solubility, and specific ligand binding sites. Most isolation procedures consist of a number of steps, each employing a different purification technique, although it is often preferable to minimize the number of steps to ensure retention of the maximum quantity of biologically active protein.

The task of devising a new purification procedure can be made easier by gathering as much biochemical information relating to the protein of interest as possible. Predicted properties such as molecular mass, pI, hydrophobicity,

post-translational modifications, and putative ligands can be deduced from the amino acid sequence of the protein or a homologue (section 5.8). Alongside experimental information relating to the purification of similar proteins, an isolation strategy can be readily developed. Careful selection of the starting material will improve the chance of devising a successful strategy; ideally, the starting material should be readily available and rich in the protein of interest (Chapter 7, section 7.1). The starting material, together with subsequent purification steps, influences cell lysate preparation. This requires a balance between optimal release of active protein from the staring material with providing conditions conducive to subsequent purification steps. As a result it is important to optimize the following factors in the cell lysate preparation: pH, ionic strength, temperature, solubility, protease inhibitors, reducing agents, and ligands to confer stability/activity.

Many of the preliminary steps in protein isolation procedures involve precipitation methods, which employ precipitants such as ammonium sulphate, polyethylene glycol, or ethanol to separate proteins by solubility. Precipitants weaken the forces between the protein and the aqueous solvent that keep the protein in solution. Precipitation induced by changes in pH and temperature can also be employed to separate proteins. Most subsequent separation techniques in an isolation procedure employ chromatographic methods, which separate proteins according to size, charge, hydrophobicity, and ligand binding (Chapter 7, section 7.2). Most procedures require a few chromatographic steps and in some favourable cases it may be possible to isolate the protein in one simple chromatographic step. A successful isolation procedure will produce material with a high specific activity and purity, suggested by SDS-PAGE analysis or mass spectrometry.

An example of a single-step purification is provided by the purification of fructose-bisphosphate aldolase from rabbit muscle. The protein precipitating between 50% and 52% saturation ammonium sulphate is essentially pure enzyme. (Note that at 25°C, a saturated solution of ammonium sulphate is about 4 M). Use of ammonium sulphate precipitation in protein purification is detailed in Chapter 7, section 7.2.1.

### 5.4.2 Recombinant proteins

Protein purification has been simplified greatly by the advent of recombinant DNA technologies that enable high levels of protein expression. Following the development of DNA cloning in bacterial systems in the mid-1970s, it soon became possible to express proteins, encoded by cloned DNA, in heterologous systems. One of the earliest successes was the overexpression of mouse dihydrofolate reductase in *Escherichia coli* (Chang *et al.*, 1978). A large number of prokaryotic and eukaryotic overexpression systems are now available, including *E. coli*, *Aspergillus nidulans*, *S. cerevisiae*, *P. pastoris*, insect cells, mammalian cells, and transgenic plants and animals. In addition to providing large quantities of material, these systems have enabled the expression of site-directed mutants and the introduction of purification tags.

In most cases the overexpressed protein accounts for between 1% and 10% of the total cell protein in the expression system. In extreme cases, this proportion can be in excess of 25%, e.g. the therapeutically important proteins human insulin and interleukin-2 are produced in *E. coli* at levels in excess of 25% total cell protein.

To achieve maximum expression of biologically active protein, it is important to consider a number of factors, including the natural source of the protein, whether it is post-translationally modified, the level of expression required, whether it is secreted, its subcellular location, its toxicity to the expression host, and whether a purification tag is required. As a general rule, optimal expression systems closely

mimic the natural production of the protein, e.g. if producing a eukaryotic, post-translationally modified, secreted protein, overexpression may be achieved in an *S. cerevisiae* or *P. pastoris* system in which the gene encoding the protein is linked to an appropriate signal sequence that directs the protein out of the cell. In many cases, the use of these signal sequences can simplify the purification of the expressed proteins from the growth medium, without the need to make cell extracts.

In some cases, attempts to overexpress proteins in *E. coli* can result in the formation of inclusion bodies which are composed of partially folded, overexpressed protein. Although inclusion bodies present a rich source of pure and stable protein, which can be readily isolated by centrifugation, recovery of folded, fully functional protein from this partially folded state can prove challenging. Typically recovery involves solublization of the inclusion body protein using a denaturant and then removing the denaturatant slowly (by dialysis, dilution or filtration) to promote refolding of the protein.

*E. coli* expression systems are by far the most commonly used, by virtue of their well-characterized molecular biology and the high levels of overexpression which can be achieved. Although mostly limited to producing intracellular proteins which are not post-translationally modified, the vast range of *E. coli* expression vectors, which can produce proteins fused to purification tags, makes *E. coli* the system of choice. Engineering a tagged protein to facilitate purification requires adding DNA encoding the tag to either the 5′ or 3′ end of the gene encoding the protein of interest to generate recombinant protein with a tag at the N- or C-terminus (see Fig. 5.12). Purification tags can take the form of enzymatically active

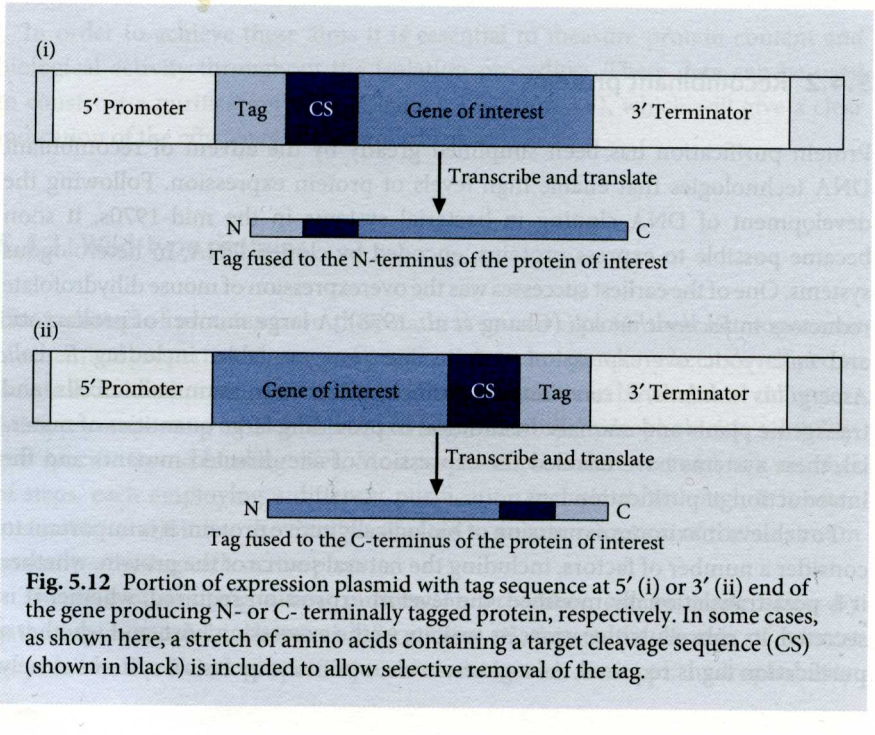
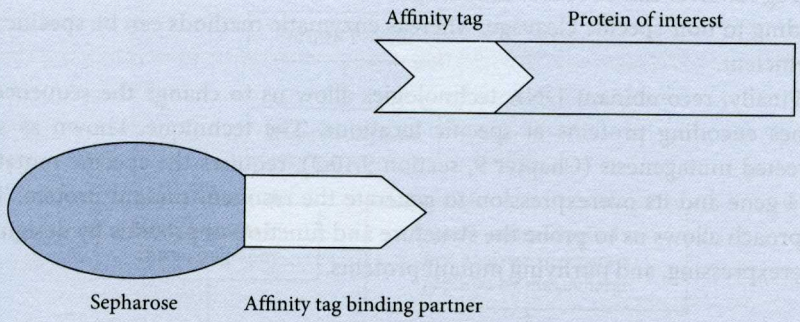



**Fig. 5.12** Portion of expression plasmid with tag sequence at 5′ (i) or 3′ (ii) end of the gene producing N- or C- terminally tagged protein, respectively. In some cases, as shown here, a stretch of amino acids containing a target cleavage sequence (CS) (shown in black) is included to allow selective removal of the tag.




| Immoblized binding partner of affinity tag | Affinity tag fused to N- or C-terminus of protein |
|---|---|
| TPEG (substrate analogue of β-galactosidase) | β-galactosidase |
| Glutathione | Glutathione-S-Transferase |
| Immunoglobulin G | Protein A |
| Cu II, Co II or Ni II | poly His or poly Cys |

**Fig. 5.13** Affinity purification of tagged proteins. A tag is fused to the N- or C-terminus of the protein of interest to facilitate purification, which relies on a specific interaction between the affinity tag and the immobilized binding partner of the affinity tag.

fusion proteins such as β-galactosidase and glutathione-*S*-transferase (Chapter 7, section 7.3.4), affinity proteins (e.g. protein A, exploiting its affinity for IgG) or metal-chelating affinity tags (see Fig. 5.13). Metal affinity tags, which include poly-histidine and poly-cysteine tags, have a high affinity for divalent metal ions such as copper, nickel, and cobalt. This property forms the basis of immobilized metal affinity chromatography: divalent metal ions are immobilized on a chelating chromatography medium (e.g. NTA (nitrilotriacetic acid)-agarose) and selectively retain proteins fused to poly-histidine or poly-cysteine affinity tags. Subsequent elution, for example with imidazole solutions in the case of His-tags, can produce large quantities of pure tagged protein in one chromatographic step (Chapter 7, section 7.3.8). Removal of purification tags may be necessary to restore function to the protein of interest or to enhance its solubility. Removal of a tag can be achieved by chemical or, more commonly, enzymatic methods. A number of examples of cleavage target sites and associated cleavage agents are given below.

| Cleavage site | Cleavage agent |
|---|---|
| Asp–Asp–Asp–Asp–Lys↓–X | Enteropeptidase |
| Leu–Val–Pro–Arg↓–X | Thrombin |
| Ile–(Glu/Asp)–Gly–Arg↓X | Factor Xa |
| Met↓–X | Cyanogen bromide |
| Asn–Gly | Hydroxylamine |

Tag removal can be somewhat complex: chemical conditions tend to be harsh, leading to non-specific cleavage, whereas enzymatic methods can be specific but inefficient.

Finally, recombinant DNA technologies allow us to change the sequence of genes encoding proteins at specific locations. The technique, known as site-directed mutagenesis (Chapter 9, section 9.10.3), requires the specific mutation of a gene and its overexpression to generate the resultant mutant protein. This approach allows us to probe the structure and function of proteins by designing, overexpressing, and purifying mutant proteins.

## 5.5 Structure determination

### KEY CONCEPTS

- Understanding the tools required to explore the different levels of protein structure
- Appreciating the importance of an integrated experimental approach to provide a more complete picture of protein structure

Prior to the early 1980s, many protein crystal structures were studied without the availability of their amino acid sequences. Nevertheless, in each case it was usually possible to produce an atomic interpretation of the X-ray diffraction pattern and to trace the path of the polypeptide chain, although the later availability of the sequence allowed detailed interpretation of the diffraction data to be undertaken. Many of these proteins, e.g. α-chymotrypsin, lactate dehydrogenase, subtilisin, and myoglobin, were studied because they could be readily purified from their natural source and formed crystals which gave rise to clear, interpretable, X-ray diffraction patterns.

The ultimate aim of protein structure determination is to gather and interpret data to reveal the three-dimensional structure of the protein at an atomic level. Whilst X-ray crystallography and high-resolution NMR spectroscopy (for proteins of molecular mass <30 kDa) generate structural data, additional experimental information relating to the primary, secondary, tertiary, and quaternary structure plus post-translational modifications, is required to interpret these data and solve the structure.

The apparent molecular mass, which is indicative of the number of amino acids within individual subunits, can be estimated by SDS-PAGE, gel filtration, ultracentrifugation, and mass spectrometry. The order of the amino acids within a polypepticle chain can be determined directly or indirectly. The direct approach would use Edman degradation or tandem mass spectrometry (Chapter 8, Section 8.3) to degrade sequentially peptides derived from the protein (see Fig. 5.14). The indirect approach would require translation of the gene encoding the protein. A comparison of the theoretical molecular mass calculated from the primary structure with the experimentally determined mass can provide evidence of post-translational modification. The nature of the post-translational modification can be determined by SDS-PAGE combined with specific removal of modifications, immunoblotting using antibodies raised to specific post-translational modifications, and mass spectrometry to analyse the mass of peptides with modifications (Chapter 8, section 8.2).

A number of techniques, including chemical modification of surface exposed residues, fluorescence, CD, and NMR spectroscopy, can indicate the nature of the secondary and tertiary structure of a protein (Chapter 8, section 8.4). The availability of secondary and tertiary structure prediction tools (section 5.2.2) together

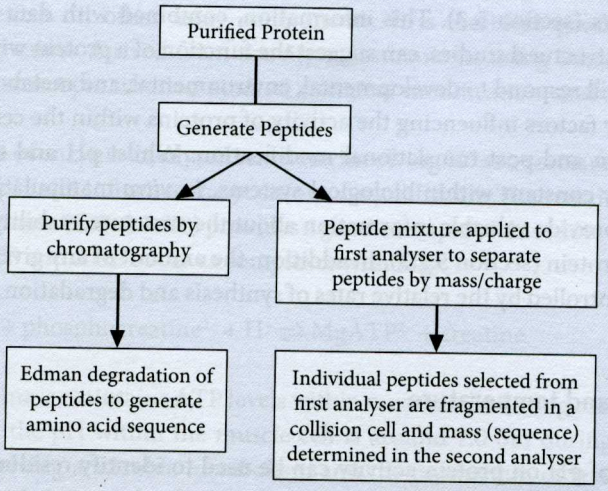



**Fig. 5.14** Ouline of scheme for the 'direct' determination of the primary structure of a protein using Edman degradation or tandem mass spectrometry.

with the primary structure can be employed to complement experimentally derived data.

Determination of the quaternary structure of a protein requires combining information relating the molecular mass and number of types of individual subunits to the apparent mass of the native multi-subunit protein, estimated by gel filtration or ultracentrifugation. Complex multi-subunit proteins may require further characterization using techniques, such as cross-linking (Chapter 8, section 8.5), to establish the number of subunits present.

Combining all of this experimental evidence permits the accurate interpretation of data generated by X-ray crystallography or NMR spectroscopy, and eventual structure determination. The advent of bioinformatics has enabled the determination of theoretical protein structures, using homology searches and prediction methods (sections 5.2.2 and 5.8). Again, this approach can be used to complement experimentally derived data.

## 5.6 Factors affecting the activity of proteins

**KEY CONCEPTS**

- Defining the major factors which influence the activity of proteins
- Understanding how to monitor their effects on protein structure and function

Studying the effects of factors such as post-translational modification, ligands, pH, and temperature on protein activity can provide insight into the structure and

Note that while we can draw some conclusions about how the protein may function within the cell, the *in vitro* assay conditions will be very different from *in vivo* conditions (e.g. high protein concentration, low concentration of ligands, protein complex formation, compartmentalization).

function of proteins. The effects are measured by monitoring their impact on activity assays (section 5.3). This information, combined with data from other activity and structural studies, can suggest the function of a protein within the cell and how it will respond to developmental, environmental, and metabolic signals.

The major factors influencing the activity of proteins within the cell are ligand concentration and post-translational modification. Whilst pH and temperature are normally constant within biological systems, *in vitro* manipulation of these factors can provide valuable information about the structure, stability, and function of the protein (section 5.2.3). In addition, the amount of any given protein in the cell is controlled by the relative rates of synthesis and degradation.

### 5.6.1 pH and temperature

The effects of pH on protein activity can be used to identify residues which are functionally or structurally important by exploiting the characteristic p$K_a$ values of amino acid side chains (see Chapter 1, section 1.2.3.2). A typical activity response to changes in pH is shown in Fig. 5.15. Under extreme pH conditions, the lack of activity is due to protein denaturation. pH conditions which elicit the highest level of activity (the pH optimum) promote side chain side ionization states which are essential for optimal activity. Identification of important residues requires activity measurements at pH values close to the pH optimum. Resultant changes in the ionization state of key residues produce changes in activity which can be used to estimate p$K_a$ values. In some cases, p$K_a$ values can be assigned to particular types of amino acids. In the example shown in Fig. 5.15, the measured p$K_a$ values of 4.0 and 10.5 are indicative of glutamic acid and lysine, respectively.

The effects of pH are not only important in studies of activity, also they are an essential consideration in developing successful purification strategies in which isolectric focusing (Chapter 6, section 6.2.3) or ion-exchange chromatography (Chapter 7, section 7.2.2) are employed.

Care must be exercised when interpreting apparent p$K_a$ values as the p$K_a$ of side chains of amino acids within proteins can be very different from the p$K_a$ of free

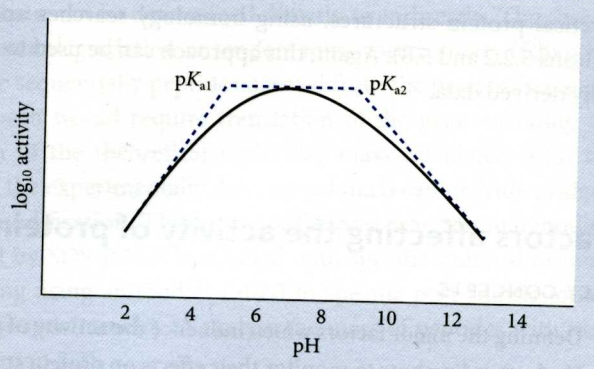



**Fig. 5.15** Change in protein activity at varying pH conditions with two ionizable groups involved.

amino acid side chains. It is also important to ensure that pH changes are not having an effect on other assay components. It is therefore good practice to confirm $pK_a$ values derived from activity measurements using other techniques such as spectroscopic titration studies, chemical modification, and site-directed mutagenesis.

In the case of enzyme-catalysed reactions involving $H^+$ as a reactant or product, a change in pH will influence the position of equilibrium. The reaction catalysed by creatine kinase (E.C. 2.7.3.2) provides a good example of this (note that in most reactions ATP and ADP occur as complexes with Mg):

$$MgADP^- + \text{phosphocreatine}^{2-} + H^+ \rightleftharpoons MgATP^{2-} + \text{creatine}$$

Creatine kinase maintains ATP levels within muscle cells during exercise. In the resting state, the pH within the muscle cell is around 7.0 but during prolonged exercise, involving anaerobic metabolism and the production of lactate, the pH can drop to about 6.3. As a result, a shift in equilibrium towards the right of this reaction occurs, to try to maintain the levels of ATP. Beyond a certain point, however, the build up of acidic substances will alter the activity of key proteins in muscle and lead to muscle fatigue.

By monitoring protein activity over a range of temperature conditions, it is possible to characterize structural changes which can be used to infer the functional properties of proteins. A typical response in protein activity is shown in Fig. 5.16. Activity measurements at lower temperatures indicate that as temperature increases, so the activity increases (see the Arrhenius equation, Chapter 4, section 4.2.3). Most mammalian enzymes tend to show maximal activity in the range 40–45°C, known as the optimum temperature; at temperatures above this range there is a rapid loss of activity resulting from protein denaturation.

The optimum temperature for an enzyme can depend on various factors such as the length of time the enzyme is incubated at the temperature used for assay. Biochemists now prefer to use the term 'apparent optimum temperature'.

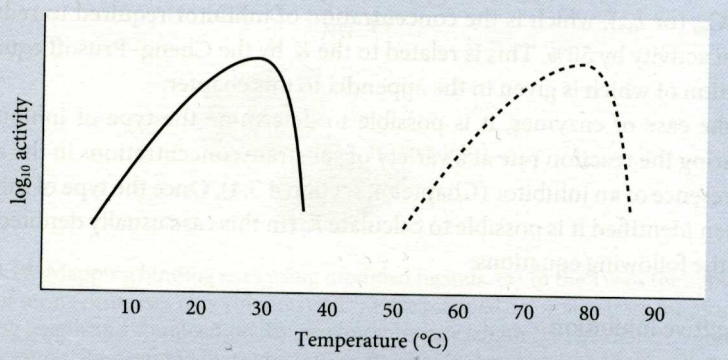



**Fig. 5.16** Effect of temperature on protein activity. Most proteins have a temperature optimum of 30–40°C (solid line) whereas thermophilic proteins are much more stable, exhibiting optimum temperature values of >70°C (dotted line).

Thermophilic organisms are categorized according to their growth temperatures: the optimal growth temperature for thermophiles is >50°C, for extreme thermophiles >65°C (e.g. *Thermus aquaticus*, the source of *Taq* polymerase), and for hyperthermophiles >90°C (e.g. *Pyrococcus furiosus*). Mesophilic organisms have growth temperature optima in the range 20–37°C.

Further information can be gleaned by monitoring activity changes in response to temperature changes under varying conditions, such as the presence of ligands and post-translational modifications. These types of experiments are important in characterizing the influence of ligands and post-translational modifications on protein activity. Generally, we can correlate temperature-induced activity changes with temperature-induced structural changes, using a number of methods such as spectroscopic techniques, chemical modification, and site-directed mutagenesis.

Proteins from thermophilic organisms respond to temperature changes in a similar way; however, in these cases, the optimum temperature is much higher. Although thermophilic proteins are generally structurally similar to their mesophilic counterparts, they have adopted a range of strategies to enhance their thermostability including additional electrostatic interactions, helix dipoles, helix capping, and shorter surface loops (Cowan, 1995).

### 5.6.2 Inhibitor and activator molecules

Inhibitors and activators are important effectors of protein activity within the cell. By measuring the influence of these effectors on protein activity *in vitro*, it is possible to establish how proteins are regulated in a cellular environment. The effect of inhibitor and activator molecules on protein activity assays (section 5.3) combined with saturation curve analyses (Chapter 4, section 4.3) can be used to calculate protein:effector stoichiometry, the strength of the interaction, and their influence on the rate constants of individual steps in the process.

Typical inhibition studies involve measuring protein activity in the presence and absence of inhibitor. The effect of the inhibitor can be quantified by calculating $K_i$, the binding constant of the inhibitor to the protein: a small $K_i$, value indicates a high-affinity inhibitor, whereas a large $K_i$, value suggests a low-affinity inhibitor.

Biochemists usually quantify the effects of inhibitors in terms of an inhibitor constant, $K_i$, whereas pharmacologists quantify the effects of an inhibitor with the term $IC_{50}$ (or $I_{0.5}$), which is the concentration of inhibitor required to reduce the protein activity by 50%. This is related to the $K_i$ by the Cheng–Prusoff equation, a derivation of which is given in the appendix to this chapter.

In the case of enzymes, it is possible to determine the type of inhibition by measuring the reaction rate at a variety of substrate concentrations in the absence and presence of an inhibitor (Chapter 4, section 4.3.4). Once the type of inhibition has been identified it is possible to calculate $K_i$ (in this case usually denoted as $K_{EI}$) using the following equations:

competitive inhibition

$$K_i = \frac{[I]}{\left(\frac{K_m^I}{K_m} - 1\right)} \qquad 5.1$$

non-competitive inhibition

$$K_i = \frac{[I]}{\left(\frac{V_{max}}{V^I_{max}} - 1\right)} \qquad 5.2$$

where $K_m$ and $V_{max}$ are the constants in the absence of inhibitor, whereas $K^I_m$ and $V^I_{max}$ are the constants in the presence of the inhibitor and these are determined as outlined in Chapter 4, section 4.4. A more accurate method of determining $K_i$ involves extending this approach to look at the effect of several concentrations of inhibitor to generate a series of Lineweaver–Burk plots (Engel, 1981).

The activation of proteins by effector molecules can be characterized by conducting protein activity assays in the presence and absence of activator. Changes in saturation curves are reflected in the rate constants that can be used to measure the affinity of the activator for the protein ($K_A$) and the type of activation, e.g. allosteric activation, in which the binding of the activator promotes cooperative substrate/ligand binding, analogous to oxygen binding to haemoglobin outlined in Chapter 4, section 4.3.2.

Repeating these experiments with modified versions of effectors can improve our understanding of protein/effector specificity and provide structural information concerning the nature of the effector binding site (see Fig. 5.17).

Over the last decade, generating libraries of modified protein ligands by a technique known as combinatorial chemistry has emerged as a powerful tool in protein characterization and drug discovery.

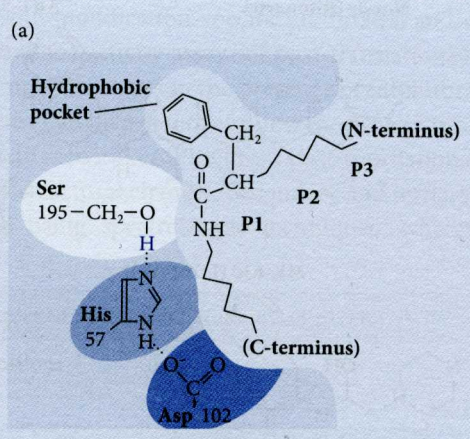



**Fig. 5.17** Mapping binding sites using modified ligands. (a) In the 1960s the active sites of serine proteases (e.g. chymotrypsin) were mapped using substrate mimics of varying length and composition. By varying substrate length (P1, P2, P3, etc.) and composition, it was noted that substrates with large hydrophobic residues at P1 were good substrates for chymotrypsin. Structure studies subsequently revealed that the substrate specificity observed relates to the shape and hydrophobic nature of the active site.

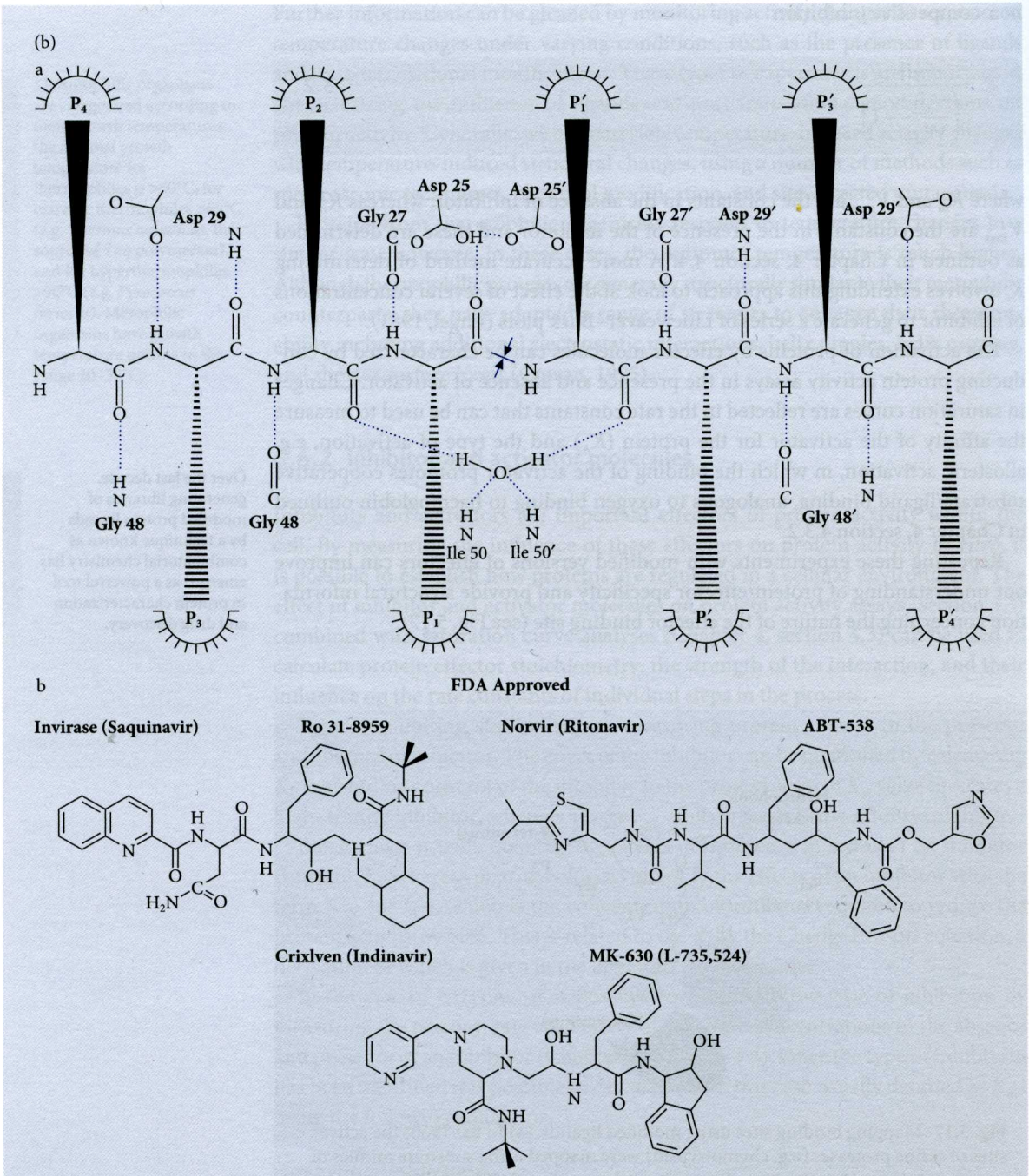



**Fig. 5.17** (b) A similar approach has been used more recently to map the active sites of therapeutically important proteases and subsequently to inhibit them. A good example of this is provided by the protease inhibitors designed for the effective treatment for HIV infection. a representation of HIV protease complexed with putative substrate, b number of effective HIV protease inhibitors which mimic the substrate (from Wlodawer and Vondrasek, (1998)).

In many cases, activity changes induced by the presence of inhibitors and activators are associated with measurable changes in protein structure. Effector-induced structural changes can be monitored using spectroscopic techniques, such as absorbance, fluorescence, CD, and NMR, and can provide valuable information which complements measurements of effector-induced activity changes.

### 5.6.3 Post-translational modifications

Post-translational modification is one of major effectors of protein activity within the cell, indeed, its importance has been re-examined recently following the completion of numerous genome sequencing projects, in particular those of higher organisms. These projects have revealed that the complexity of higher organisms is due to the differential RNA processing and post-translational modification of gene products. Thus, the human genome is thought to encode about 23 000 gene products, but these are thought to give rise to several hundred thousand distinct proteins. There are many forms of post-translational modification (see Table 5.1) which can influence protein activity. All of these covalent modifications are the result of enzyme-catalysed reactions which confer rapid and amplified changes in protein activity in response to small cellular signals.

Protein activity assays alongside saturation curve analyses (Chapter 4, section 4.3) can be used to characterize the effects of post-translational modifications. The exploration of these effects requires uniform protein preparations, both with and without post-translational modification. This can be checked using SDS-PAGE, Western blotting with modification-specific antibodies, and/or mass spectrometry. Developing standard activity assays of post-translationally modified protein to include activity measurements in the presence of inhibitors/activators and the effects of temperature can help determine how the protein is regulated *in vivo*.

Changes in protein activity resulting from post-translational modification may be accompanied by structural changes. Secondary and tertiary structural changes can be monitored using spectroscopic techniques, whereas changes in the

**Table 5.1 Examples of post-translational modifications**

| Amino acid modifications | |
|---|---|
| **Modifications** | **Example** |
| Cysteine: disulphide bond formation | Lysozyme |
| Lysine biotinylation | Acetyl CoA carboxylase |
| Serine phosphorylation | Glycogen phosphorylase |
| Threonine phosphorylation | Cyclin-dependent kinase |
| Tyrosine phosphorylation | Cortactin |
| Addition of prosthetic groups—thiamine diphosphate | Pyruvate dehydrogenase |
| Proteolytic processing | Chymotrypsin |
| Protein targetting (signal sequences) | Penicillin acylase |

quaternary structure can be determined using gel filtration, ultracentrifugation, and cross-linking studies (Chapter 8, section 8.2).

A good example of a protein which undergoes post-translational modification resulting in activity and structural changes is the enzyme glycogen phosphorylase (E.C. 2.4.1.1). Phosphorylase exists in two interconvertible forms: phosphorylase *a*, which is phosphorylated at Ser 14 by a specific kinase, and phosphorylase *b*, which is not phosphorylated. In the absence of AMP, phosphorylase *b* is inactive, whereas phosphorylase *a* actively catalyses the cleavage of glycogen to produce glucose-1-phosphate:

$$(1,4\text{-}\alpha\text{-D-glucosyl})_n + P_i \rightleftharpoons (1,4\text{-}\alpha\text{-D-glucosyl})_{n-1} + \alpha\text{-D-glucose 1-phosphate}$$

Conversion of phosphorylase *b* to phosphorylase *a* results in tertiary and quaternary structural changes which are associated with the regulatory properties of both forms of this key enzyme.

## 5.7 Interactions with other macromolecules

**KEY CONCEPTS**

- Appreciating the importance of protein–protein interactions *in vivo*
- Understanding the need to use appropriate experimental conditions to study these interactions

Whilst it is convenient to study isolated proteins *in vitro*, this does not reflect how they function *in vivo*. Within the cell, proteins are part of a highly organized system involving subcellular structures and high concentrations of macromolecules, including other proteins, nucleic acids, and lipids. Recent developments in techniques used to study protein interactions (section 5.3.2) are beginning to reveal the complexity of these interactions, many of which are weak and transient, but nonetheless important for the expression of protein function. The best-characterized protein–protein, protein–nucleic acid, or protein–lipid complexes are those which are most stable and have survived the harsh cell disruption procedures used during complex isolation.

The study of the interaction with other molecules can involve the identification of molecules which interact with the protein of interest or characterization of the effect of these molecules on protein activity. Typically, the identification of novel effectors involves the immobilization of the protein of interest which then serves as a bait to bind potential effectors. A great deal of care must be exercised when establishing the conditions which will promote and maintain protein–effector interactions as these interactions are highly sensitive to pH, ionic strength and the presence of other effectors. Failure to optimize conditions can result in false positive results, i.e. certain conditions can promote non-specific interactions, or

produce false negative data in which interactions are not identified due to weakened binding. The influence of 'other molecules' on protein activity can be determined using activity assays (section 5.3) and saturation curve analyses (Chapter 4, section 4.3). The binding of 'other molecules' may be accompanied by structural changes which can be monitored as outlined in the previous section, providing structural data to complement activity measurements.

## 5.8 Use of bioinformatics

**KEY CONCEPTS**

- Appreciating the range of databases and bioinformatic tools available to assist protein characterization
- Understanding the theoretical basis of the tools used to calculate properties of a protein from its sequence

The study of proteins is no longer an exclusively laboratory-based activity pursued by biochemists; instead it is possible to use computer-based methods to examine proteins in detail. The field of bioinformatics has harnessed the exponential growth in the amount of information relating to nucleotide sequences, protein sequences, and biomolecular structures. As a result, all of this information has been organized into web-based databases which can be accessed and analysed using a range of computer programs. In this section, we shall consider some of these databases and how they can be employed to explore protein structure and function. In addition, we shall look at a range of programs which can be used to analyse database information to assist protein characterization.

### 5.8.1 Web resources and databases

Nucleotide sequences, amino acid sequences, and protein structures are collected within a number of web-based databases. The aim of each of these databases is not only to collect information but also to present it in an annotated form to help biologists understand better the significance of the data. More recently, there has been an effort to integrate data sets (e.g. linking individual nucleotide/protein sequences to related three-dimensional structures, metabolic pathway databases, enzyme databases, disease databases, organism-specific databases, two-dimensional gel databases, and associated references), allowing researchers to characterize more fully the structural and functional properties of proteins.

The major databases for RNA and DNA sequences are EMBL, Genbank, and DDBJ.

Sequencing technologies have enabled the completion of increasing numbers of genome-sequencing projects, which in turn has maintained a very large growth

The International Nucleotide Sequence Database Collaboration (http://www.insdc.org/) combines the efforts of European Molecular Biology Lab (EMBL) in Hinxton, UK, GenBank in Bethesda, Maryland, USA and DNA Data Bank of Japan (DDBJ) in Mishima, Japan.

of these databases, with the EMBL Nucleotide Sequence Database containing 130 million sequence entries, comprising in excess of $2 \times 10^{11}$ nucleotides in 2008. Search tools permit retrieval of relevant database entries which can be further analysed using a suite of molecular biology tools (e.g. restriction site identification, sequence comparisons, and translation to identify open reading frames) or cross-referenced with other databases.

One of the main protein sequence databases is SwissProt (http://www.ebi.ac.uk/swissprot/), which contains sequences which have been generated either directly by amino acid sequencing or indirectly from translating open reading frames of genes within TrEMBL (a database containing translated coding regions of EMBL/GenBank/DDBJ nucleotide databases). The SwissProt database is linked with about 50 other databases, allowing extensive characterization of the protein of interest.

An increasing number of whole genome-specific databases have become available over the past decade. Each genome is annotated to identify individual genes and these are cross-referenced with protein sequence and three-dimensional structure databases. The genomes characterized to date reflect their importance either as biological models (e.g. *E. coli, C. elegans, S. cerevisiae*) or as commercial organisms (e.g. rice, maize, chicken, salmon) or as disease-related organisms (human, *Helicobacter pylori, Anopheles gambiae*). Fig. 5.18 displays the NCBI websites for the *C. elegans* and the *A. gambiae* genomes.

The bacterium *H. pylori*, although isolated over 100 years ago, has only recently been shown to cause stomach or duodenal ulcers.

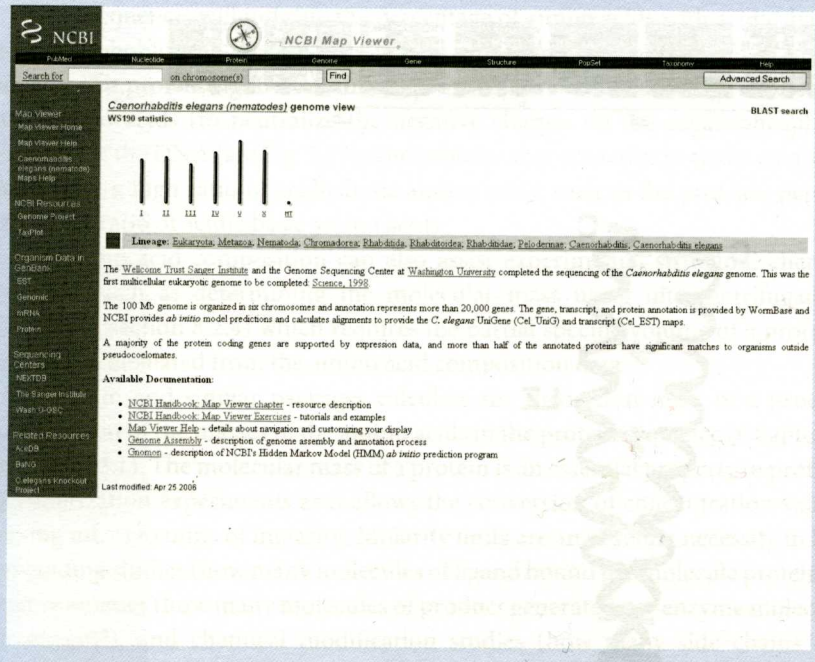
*A. gambiae*, the mosquito, carries the malaria parasite and is the main vector of malaria in Africa, where it is estimated that a million children die each year from this disease. Characterization of the genomes of such disease agents will enhance the development of disease prevention, detection, and treatment.

The PDB is currently maintained by the Research Collaboratory for Structural Bioinformatics (RSCB), involving Rutgers (NJ), San Diego (CA) and Madison (WI), and is located at http://www.rcsb.org/pdb. Linked sites are maintained by the European Bioinformatics Institute in Cambridge, UK (http://www.ebi.ac.uk/msd/) and in Osaka, Japan (PDBj; see http://www.pdbj.org).

All three-dimensional biomolecular structures are deposited in the PDB database. A search of the PDB database using the name of a protein or its PDB entry code, will provide an image of the protein structure together with information about the protein and its source, reference to the paper describing structure determination, the amino acid sequence, ligands present in the structure, secondary structure composition, and the atomic coordinates. The atomic coordinates can be downloaded from the PDB site and analysed in detail using molecular viewer software such as Rasmol (http://www.openrasmol.org) or its derivative Protein Explorer (http://www.umass.edu/microbio/rasmol). Such tools facilitate an exploration of protein structure, including ligand binding sites, catalytic residues, key structural residues, and protein–protein interactions.

### 5.8.2 Sequence analysis

Protein sequences, derived from experimental data or database entries, can generate a wealth of information that can assist protein characterization. Empirically derived information will be considered in this section; sequence comparisons



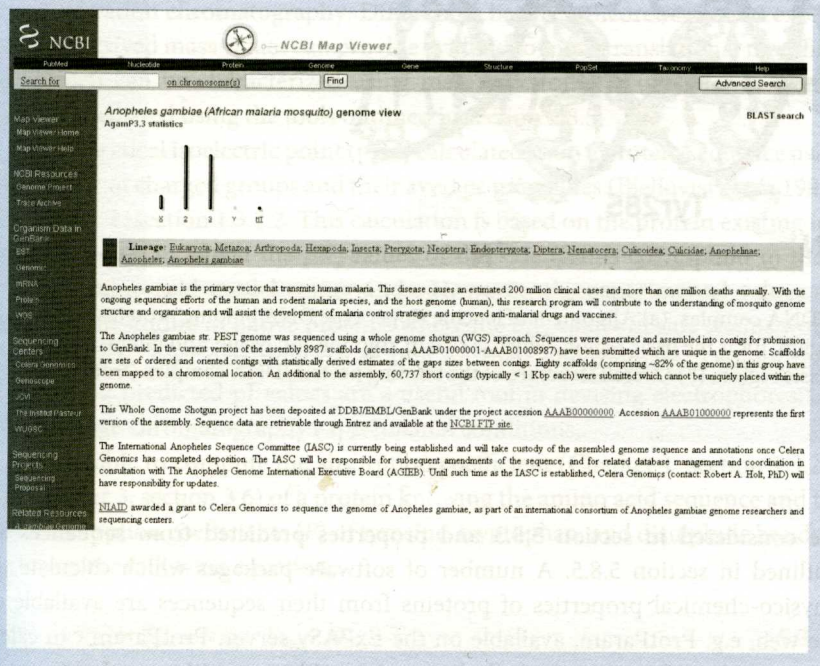




**Fig. 5.18** NCBI website entries for the complete genomes of *C. elegans* and *A. gambiae*. All genome sequencing projects, complete and incomplete, can be accessed via http://www.genomesonline.org/.

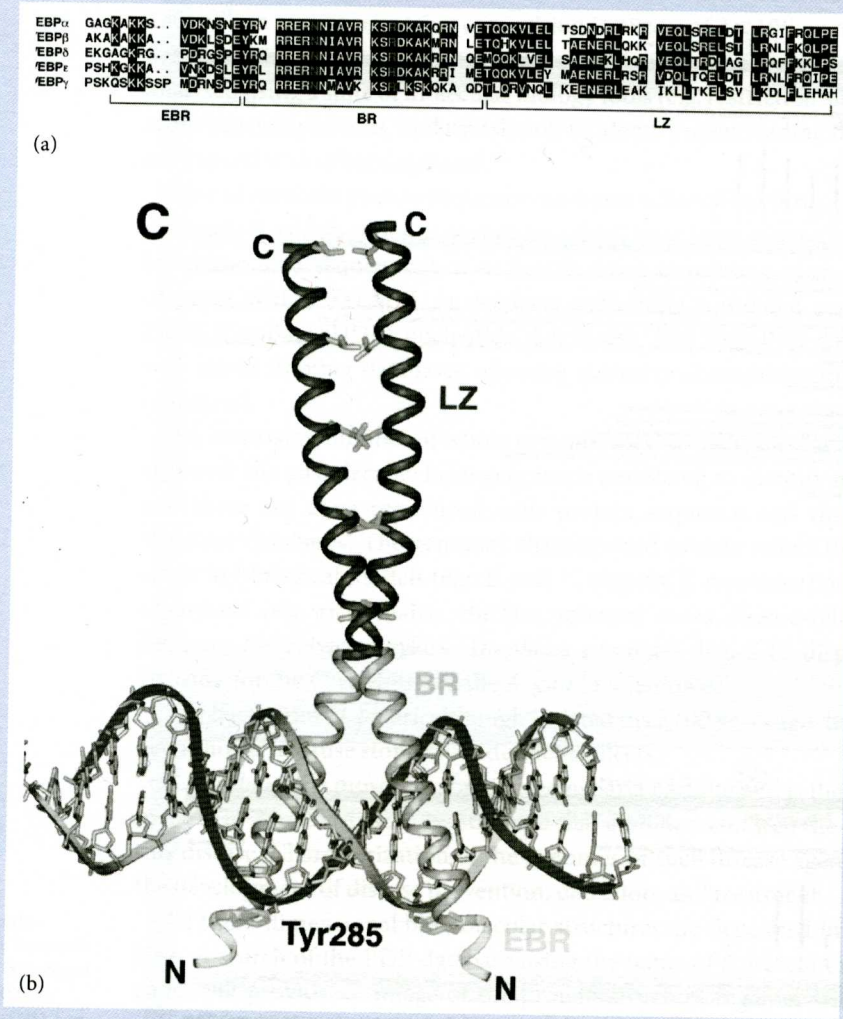



**Fig. 5.19** DNA binding protein/DNA complex. (a) Alignment of leucine zipper type DNA binding proteins, where LZ is the leucine zipper region, BR is the basic region, which is rich in arginine and lysine, and interacts with DNA, and EBR is the extended basic region. (b) Overall structure of the binding protein/DNA complex (from Miller *et al.*, 2003).

are considered in section 5.8.3 and properties predicted from sequences are outlined in section 5.8.5. A number of software packages which calculate the physico-chemical properties of proteins from their sequences are available on the web, e.g. ProtParam, available on the ExPASy server. ProtParam can calculate a number of parameters, including amino acid composition, molecular mass, pI (isoelectric point), and absorption coefficient, from complete sequences or selected stretches of sequence.

The ExPASy (Expert Protein Analysis System) web address is http://www.expasy.org/.

The amino acid composition derived from a protein sequence, or selected stretch of sequence, will indicate the percentage of amino acids with acidic, basic, and hydrophobic side chains. This information alone can provide clues to the function of a protein e.g. DNA-binding proteins have stretches which are rich in basic amino acids (to neutralize the negative charges on the sugar–phosphate backbone of the DNA (see Fig. 5.19)) and proteins that are active in acidic environments have a high ratio of acidic:basic amino acids, such as the protease pepsin with ~11:1 ratio of acidic:basic amino acids.

The amino acid composition can also assist experimental structural characterization such as determining the molecular mass using ultracentrifugation (Chapter 8, section 8.2.4) which requires the partial specific volume of a protein; this can be calculated from the amino acid composition data.

ProtParam and similar packages calculate the theoretical mass of a protein from the sum of the masses of the amino acids in the protein sequence (Chapter 1, section 1.3.2.1). The molecular mass of a protein is an essential property in protein characterization experiments as it allows the conversion of concentration values (e.g. mg $mL^{-1}$) to units of molarity. Molarity units are an absolute necessity in ligand binding studies (how many molecules of ligand bound per molecule protein?), enzyme kinetics (how many molecules of product generated per enzyme molecule per second?), and chemical modification studies (how many side chains are modified per molecule of protein?). In addition, the molecular mass gives an indication of the relative mobility of the protein on polyacrylamide gel electrophoresis and gel filtration chromatography. Differences between theoretically and experimentally derived mass values can provide evidence of post-translational modification, which can be characterized using mass spectrometry (Chapter 8, section 8.2.5) or predicted using the tools outlined in section 5.8.5.

The theoretical isoelectric point (pI) is calculated from a protein sequence using the number of charged groups and their average $pK_a$ values (Bjellqvist *et al.*, 1993), see Chapter 1, section 1.3.2.2. This calculation is based on the protein existing in a denatured state in which the $pK_a$ values of the side chains correspond to those of small amino acid model compounds. However, the experimental values are derived from studies of native proteins where the $pK_a$ values of side chains would be expected to vary markedly depending on their environment. Despite these reservations, predicted pI values are a useful tool in devising electrophoresis or ion-exchange chromatography experimental conditions.

It is possible to compute the absorption coefficient (Chapter 1, section 1.3.2.3, and Chapter 3, section 3.6) of a protein knowing the amino acid sequence and the molar absorption coefficients ($E$) of tyrosine, tryptophan, and disulphide bonds at $A_{280}$ using the following equation:

$$E_{Prot} = N_{Tyr} \times E_{Tyr} + N_{Trp} \times E_{Trp} + N_{Cys} \times E_{Cys} \qquad 5.3$$

This equation is taken from Pace *et al.* (1995). It applies to those proteins known to contain disulphide bonds; if the protein is intracellular and therefore does not contain disulphide bonds, the term in $N_{Cys}$ is omitted.

where $N_X$ is the number of amino acid $X$ per polypeptide chain. The values of $E_{Tyr}$, $E_{Trp}$ and $E_{Cys}$ are 1490, 5500 and 62.5 $M^{-1}$ $cm^{-1}$, respectively. Table 5.2 presents a

**Table 5.2 Sample calculations of the absorption coefficients for a range of proteins.**

| Protein | Molecular mass (Da) | $N_{Tyr}$ | $N_{Trp}$ | $N_{Cys}$ | $E_{Prot}$ (calc) (1 M)* | $E_{Prot}$ (calc) (1 mg mL$^{-1}$)* | $E_{Prot}$ (exp) (1 mg mL$^{-1}$) |
|---|---|---|---|---|---|---|---|
| Insulin (bovine) | 5 734 | 4 | 0 | 6 | 5 960 (6 335) | 1.04 (1.10) | 0.97 |
| Lysozyme (hen) | 14 314 | 3 | 6 | 8 | 37 470 (37 970) | 2.62 (2.65) | 2.63 |
| Chymotrypsinogen (bovine) | 25 666 | 4 | 8 | 10 | 49 960 (50 585) | 1.95 (1.97) | 1.98 |
| Phosphoglycerate kinase (yeast) | 44 607 | 7 | 2 | 1 | 21 430 (21 430) | 0.48 (0.48) | 0.49 |
| Pyruvate kinase (rabbit muscle) | 57 917 | 9 | 3 | 9 | 29 910 (30 410) | 0.52 (0.53) | 0.54 |
| Serum albumin (bovine) | 66 296 | 20 | 2 | 35 | 40 800 (42 925) | 0.62 (0.65) | 0.66 |

*The numbers in brackets assume that all (or the maximum number) of Cys residues form disulphide bonds. Disulphide bonds would be expected in extracellular proteins such as insulin, lysozyme, chymotrypsinogen, and serum albumin, but not in intracellular proteins such as phosphoglycerate kinase or pyruvate kinase.

number of examples of proteins with calculated absorption coefficients, as well as the experimentally measured values.

Division of $E_{Prot}$ by the molecular mass of the polypeptide chain of the protein (in Da) gives the absorption coefficient for a 1 mg mL$^{-1}$ solution of the protein.

The calculation of $E_{Prot}$ refers to a protein in its native state and is based on average values of the absorption coefficients for the amino acids in native proteins. There are other equations which allow the $E_{Prot}$ of the denatured protein to be calculated with reasonable accuracy; however, the value obtained must then be adjusted by an experimentally determined factor to give the correct value for the native protein. These calculations will not be reliable if the protein contains prosthetic groups which absorb in the near UV portion of the spectrum.

The absorption coefficient at $A_{280}$ is essential for simple and accurate estimates of the concentration of protein present in most purification and characterization techniques (Chapter 6, section 6.1.5). In the case of proteins that contain very few tyrosine and tryptophan residues, e.g. collagen, it may be possible to estimate the concentration of protein present using absorbance measurements at $A_{205}$ due to the peptide bond absorbance (Chapter 6, section 6.1.6).

### 5.8.3 Sequence comparisons

The term 'homologous' in terms of sequence comparisons implies sequences are descended from a common ancestor. As the name implies, 'similarity' is simply a measure of how similar sequences are, irrespective of the evolutionary route, i.e. could arise from divergence, convergence, gene duplication, or a combination of these.

Sequence comparisons (Chapter 1, section 1.3.2.7) have proved increasingly informative in protein characterization as a result of the growth of sequence databases. Whilst pairwise alignments provide a measure of the similarity of two sequences, the information arising from multiple sequence alignments, using more than two sequences from a protein family, can be used to:

*Suggest the function of unknown proteins*: The rationale of this approach is to relate the sequence of an unknown protein to the multiple alignment of a family

of similar known proteins. This approach assumes that sequence features common to the multiple alignment and the unknown protein will confer similar function and structure. The growth of sequence databases alongside improved software packages has greatly improved the accuracy of this approach. Most packages use sequence and structure information to generate multiple alignments of families of known proteins, which can be used to identify unknown protein function, e.g. PSI-BLAST, Gapped BLAST and PSI-BLAST (Altschul *et al.*, 1997). The significance of the similarity of an unknown protein to multiple alignments identified by programs such as BLAST is given a score. This score provides a measure of the probability that similarities between sequences have occurred by chance. In the case of PSI-BLAST an *E*-score is given, ranging from 0 to the number of sequences in the database: *E*-scores less than 0.2 are indicative of homology and values greater than 1.0 indicate that the similarities are just as likely to have arisen by chance. If we consider the serine proteases, we find that the similarity between trypsin (EC 3.4.21.4) and chymotrypsin (EC 3.4.21.1) produces an *E*-score of $\sim 1 \times 10^{-40}$, whereas the similarity between trypsin and subtilisin (EC 3.4.21.62) produces an *E*-score >1.0. Structural studies have established that trypsin and chymotrypsin are closely related and are both members of the trypsin-like superfamily (section 5.8.4). However, while trypsin and subtilisin have similar catalytic residues (the charge relay residues Asp, His and Glu, see Chapter 9, section 9.8.2), the non-catalytic residues are not similar. Structural studies have revealed that trypsin is composed mainly of β-sheet and is a member of the trypsin-like superfamily, whereas subtilisin is an α-helix/β-sheet class of protein and is a member of the subtilisin-like superfamily.

BLAST is an acronym for Basic Local Alignment Search Tool.

*Identify functionally and/or structurally important residues*: Multiple alignments of related protein sequences can be used to identify residues that are identical across all sequences. Absolute identity is often indicative of functionally and/or structurally important residues, e.g. residues essential for the catalytic mechanism of enzymes or residues essential for prosthetic group interactions. A good example of this is provided by haemocyanin, the oxygen transporter in many chelicerates and arthropods, which contains two copper-binding sites, copper A and copper B. Each copper is coordinated to three histidine residues which are highly conserved across all species (see Fig. 5.20), highlighting the importance of these residues.

As well as complete identity, comparisons of sequences can show that certain amino acids may be functionally conserved (e.g. always non-polar such as Leu, Ile, or Val), or may be freely variable (e.g. Lys, Asp, Gly, or Phe). Clearly freely variable amino acids are unlikely to play any crucial role in either the structure or the function of the protein concerned.

*Improve structure prediction tools*: It is possible to predict the three-dimensional structure of a protein by modelling its sequence onto a homologous structure of a known protein (parent). This approach has become more accurate with the availability of multiple structures of homologous proteins that reveal the highly conserved structural features of the family.

*Detect and characterize evolutionary relationships*: The classical approach to phylogeny, based on macroscopic observations of form and function, has been replaced with the comparison of molecular information, including protein sequences. Sequence alignments can be used to infer the evolutionary relationship

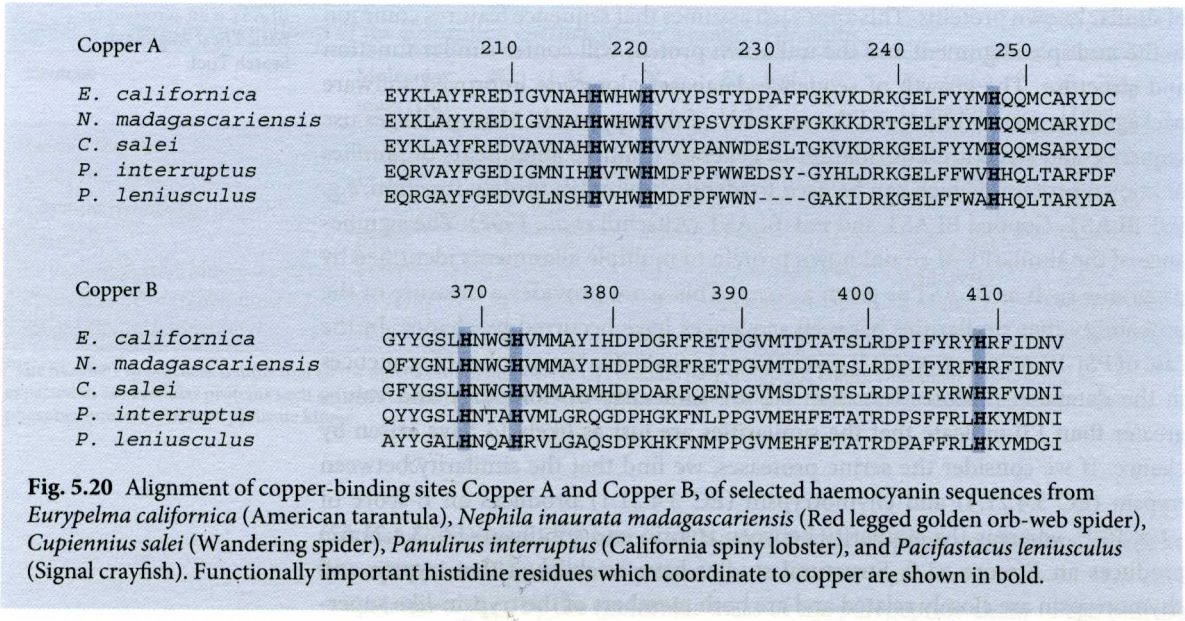



**Fig. 5.20** Alignment of copper-binding sites Copper A and Copper B, of selected haemocyanin sequences from *Eurypelma californica* (American tarantula), *Nephila inaurata madagascariensis* (Red legged golden orb-web spider), *Cupiennius salei* (Wandering spider), *Panulirus interruptus* (California spiny lobster), and *Pacifastacus leniusculus* (Signal crayfish). Functionally important histidine residues which coordinate to copper are shown in bold.

between organisms (by comparing the same protein from different organisms) or between proteins (by comparing related proteins from the same organism). Protein sequences greater than 100 amino acids in length and which share 30% identity are assumed to be homologous, i.e. are members of the same family, whereas those sharing less than 15–20% identity are similar but are not necessarily derived from a common evolutionary ancestor. In such cases additional information such as three-dimensional structures may be needed to confirm homology (Rost, 1999).

### 5.8.4 Structure comparisons

Protein structure comparisons are powerful tools in establishing structure, function, and evolutionary relationships, particularly when comparing distantly related proteins with low levels of sequence identity. The development of databases which group proteins with related structures in a hierarchal manner, e.g. SCOP (Structural Classification of Proteins; http://scop.mrc-lmb.cam.ac.uk/scop/) and CATH (Class Architecture Topology Homologous superfamily; http://www.cathdb.info/latest/cath_info.html), has enabled the rapid identification of the close relatives of any particular protein. In order of decreasing hierarchy, the SCOP database organizes protein domains according to family (homologous with respect to structure inferring a common evolutionary ancestor), superfamily (similar structure/function suggesting a probable relationship between proteins),

and fold (similar secondary structure arrangement but distantly related function and overall structure). The CATH database organizes protein domains into homologous superfamilies (sharing a common ancestor), topology (similar shape and connectivity of secondary structure), architecture (similar secondary structure arrangement), and class (similar secondary structure composition).

Having identified structures related to the protein of interest, it is then possible to use molecular visualization tools (section 5.8.1) to overlay related structures onto the protein of interest to characterize protein function, identify residues of structural and functional importance, and infer evolutionary relationships, in a manner analogous to sequence comparisons.

### 5.8.5 Predicting possible functions of proteins

The classical experimental approach to determining the function of a novel protein requires structure characterization (section 5.2.2), determination of the factors influencing activity (section 5.2.3), and identification of ligands (section 5.2.4). However, the work by Max Perutz in the 1960s on myoglobin and haemoglobin indicated that the three-dimensional structure of a protein, and hence its function, is determined by its amino acid sequence. The subsequent expansion of sequence and structure databases has enabled the development of tools to predict structure and function from amino acid sequences.

Max Perutz and John Kendrew were awarded the Nobel Prize in chemistry in 1962 for their studies on the structures of globular proteins, including haemoglobin and myoglobin. Perutz noted that the tertiary structures of haemoglobin and myoglobin subunits were similar and proposed that this was as a result of similar amino acid sequences. The subsequent determination of the amino acid sequences of these proteins confirmed they shared a substantial degree of identity (of the order of 30%) that conferred similar structures and functions.

In many cases, it will be possible to identify homologues of a novel protein using tools such as *PSI-BLAST*. Assuming the complete sequence of the novel protein shares a high degree of identity with a family of well-characterized homologous proteins and contains key functional residues, it will be possible to assign function. More functional information can be obtained using comparative tools to build a high-quality structure model of the novel protein, e.g. details of an active site can be used to infer substrate specificity.

Folded proteins often consist of a number of structural units known as domain folds, with each fold often associated with a particular function. In the absence of a family of well-characterized homologues, fold recognition methods can be used to match the novel sequence to known folding patterns. This process involves aligning the query sequence to a number of similar sequences of known fold, e.g. nucleotide binding fold, globin-like fold, or immunoglobulin fold (see Chapter, section 1.5.2). Once aligned, approximate models of the novel protein are constructed and evaluated to identify the most accurate model.

Secondary structure prediction methods are outlined in this chapter, section 5.2.2 and in Chapter 8, section 8.4.

In addition to suggesting function from the predicted structure, there are a number of bioinformatic tools that can predict the subcellular location of proteins from their amino acid sequences. Transmembrane helices and their topology can be predicted with a high degree of accuracy, based on sequence length and composition. For example, the predicted transmembrane helices of bovine rhodopsin are in good agreement with the experimentally determined structure of this

seven transmembrane helix protein (bovine rhodopsin PDB entry 1F88). Likewise there are a number of methods that can identify N-terminal pre-sequences which determine organelle destination, e.g. mitochondrial transfer peptides (25–45 residues), nuclear location signals (4–6 residues), signal peptides for extracellular proteins (20–30 residues), and peroxisomes target signals (either the C-terminal signal Ser–Lys–Leu as in catalase or a 30 residue N-terminal signal sequence, as occurs in thiolase). Examples of other functional motifs are outlined in Chapter 1, section 1.3.2.6.

NetPhos http://www.cbs.dtu.dk/services/NetPhos/

NetNGlyc http://www.cbs.dtu.dk/services/NetNGlyc/

CASTp http://sts.bioengr.uic.edu/castp/

PONDR http://www.pondr.com

Factors affecting protein activity, including post-translational modifications and ligand interactions, can also be predicted using bioinformatic tools. Post-translational modification sequence motifs (see Chapter 1, section 1.3.2.6) can be identified by a number of pattern recognition programs such as NetPhos to predict serine, threonine and tyrosine phosphorylation sites in eukaryotic systems and NetNGlyc to predict N-glycosylation sites in human proteins. Analysis of the surface of protein structures, using programs such as CASTp, can identify and characterize accessible pockets such as ligand binding sites and active sites. As a final example, regions of proteins that do not adopt stable three-dimensional structure even under native conditions can be predicted using PONDR (Prediction of Natural Disordered Regions); in many proteins such regions can play important roles in adjusting the specificity of ligand recognition.

Prediction of protein structure, factors influencing activity, and ligand binding sites using bioinformatic tools can reveal the possible function of a protein and provide valuable information to complement classical experimental approaches.

## 5.9 Experimental design

**KEY CONCEPTS**

- Understanding the process of successful experimental design
- Appreciating the need to comply with relevant safety and ethical requirements

A sound understanding of the goals and methods employed to characterize proteins can be best put into practice alongside good experimental design. Whether you are aiming to study the function, structure, or the nature of effectors of a particular protein, it is important to consider some essential, if somewhat obvious, aspects of experimental design. A thorough literature review coupled with use of a range of bioinformatic tools (e.g. sequence data of a similar protein, the calculated molecular mass, isoelectric point, and absorption coefficient at 280 nm) will yield a wealth of information to assist experimental design. In this section, we shall consider the design of two types of experiments to demonstrate how to obtain relevant and specific information to address key questions being asked about a protein.

*Example 1*

Suppose that we wish to characterize the quaternary structure of a signal-transducing G-protein and explore the binding of a GTP analogue to it. Prior to characterization studies, it is essential to test the purity and structural and functional integrity of the G-protein. SDS-PAGE and mass spectrometry can be employed to check that possible contaminating proteins are absent. SDS-PAGE combined with specific activity measurements (GTPase assay) will provide a good indication of the structural and functional integrity of the protein. To maintain this integrity in later experimental procedures, it is essential to determine the optimal conditions for storage and activity measurements of the G-protein. This will require exploring a range of possible storage and assay conditions; however, guidance may be available from previous studies on similar systems.

All protein characterization experiments hinge on two important methods: an accurate method to determine the concentration of the protein and a specific assay to determine the activity of the protein. As each protein concentration determination method has its limitations, it would be worthwhile determining the concentration of the sample of G-protein using a number of different methods (see Chapter 6, section 6.1) to ensure the reliability of the estimate. A reliable assay of activity requires high-quality reagents and a number of control measurements to validate the procedure. In the case of a G-protein, a simple colourimetric assay involving malachite green to measure phosphate release is used to monitor GTPase activity. This assay includes a suitable buffer, salts, GTP, malachite green, and G-protein. Omission of each assay component in turn will indicate the specificity of the assay. It is also important to determine whether the assay measurements are proportional to the amount of G-protein added, i.e. doubling the amount of G-protein added to the assay should generate a two-fold increase in activity. If this is not the case, it may be that some assay component is present in limiting quantities and the concentrations of the various components may need to be adjusted to overcome the problem.

A number of experimental approaches will be required to determine the quaternary structure of the G-protein including gel filtration chromatography (Chapter 8, section 8.2.3), SDS-PAGE (Chapter 8, sections 8.2.1 and 8.2.2), and mass spectrometry (Chapter 8, section 8.2.5). All three methods require a range of suitable standard proteins, of known molecular mass, to calibrate the procedure. Analysis of the G-protein using each of these methods may require a buffer exchange step; mass spectrometry and SDS-PAGE are sensitive to high salt concentrations. On the other hand, poor gel filtration data may be obtained if the ionic strength of the buffer is too low, since non-specific interactions between protein and chromatographic media can occur. It is also important to consider the amount of protein analysed using each of these techniques, as too much or too little may compromise the quality of the data.

To explore the effects of a GTP analogue on the G-protein, activity assays and binding assays should be conducted. Prior to gathering data in the presence of

the G-protein, it will be important to assess the impact of the presence of the GTP analogue on the activity assay and binding assay, e.g. is it soluble under the assay conditions, will it cause a change in pH or a change in absorption? Preliminary measurements in the presence of the GTP analogue will be required to establish the correct concentration range to use. Ideally, we should be aiming to cover the concentration range from 0.1 to 10 times the $K_d$ for the GTP analogue in the binding assays. It is worthwhile exploring more than one binding assay method to confirm our findings, e.g. a direct binding measurement using equilibrium dialysis and a spectroscopic titration method to monitor ligand-induced structural changes.

Throughout all of these approaches, it is good practice to test the outcome of empirical calculations using experimentally derived data. For example, dilution of a stock solution of G-protein should result in a suitable $A_{280}$ value or introduction of a calculated amount of G-protein to an assay should generate a predictable activity measurement.

### *Example 2*

Use of the artificial substrate MUNANA leads to a much more convenient and sensitive assay for the enzyme compared with the natural substrates (removal of sialic acid groups from oligosaccharides, glycoproteins and glycolipids).

Suppose that we wish to obtain kinetic parameters for the enzyme neuraminidase (EC 3.2.1.18) from the virus H5N1 (the causative agent of bird flu) acting on the fluorogenic substrate 2′ (4-methylumbelliferyl)-α-D-*N*-acetylneuraminic acid (MUNANA). Neuraminidase is crucial for the degradation of glycan structures at the surfaces of viral-infected cells and its action is required for the release of viral particles. We also wish to study the effects of the inhibitor 5-*N*-acetyl-3-(1-ethylpropyl)-1-cyclohexene-1-carboxylic acid (Tamiflu) (Fig. 5.21), on neuraminidase activity.

As in the previous example, preliminary experiments should be undertaken to establish the purity and quality of the neuraminidase, determine optimal storage conditions, and test the ability of the assay to provide reliable and accurate results. To obtain data that will establish the kinetic parameters of neuraminidase, the concentration of the substrate MUNANA should ideally be varied over the range from 0.1 to 10 times the $K_m$. The success of this approach requires that the MUNANA remains soluble in the assay over this concentration range and does

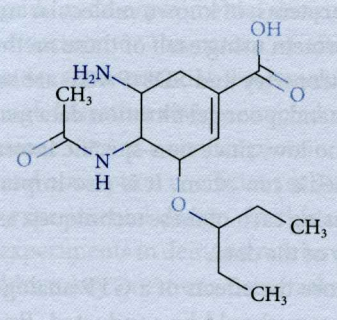



**Fig. 5.21** Structure of the neuraminidase inhibitor, 5-N-acetyl-3-(1-ethylpropyl)-1-cyclohexene-1-carboxylic acid (Tamiflu).

not alter the properties of the assay (e.g. change pH or spectroscopic properties). Repeating this range of activity measurements in the presence of increasing concentrations of Tamiflu will provide a measure of the type and strength of neuraminidase inhibition. A series of preliminary experiments will be required to establish the concentration ranges of substrate and inhibitor required and to check that Tamiflu, in the absence of neuraminidase, has no effect on the assay procedure.

Two final considerations in the design of any protein characterization experiment are safety and ethical issues. All experiments should meet the safety standards required by national health and safety regulatory bodies and institutional safety guidelines. Likewise, experiments involving animals or human subjects (or materials derived from them) are required to adhere to national ethical legislation and institutional requirements.

## References for Chapter 5


Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al.* (1997) *Nucl. Acids Res.* **25**, 3389–402.

Armache, K.-J., Mitterweger, S., Meinhart, A., and Cramer, P. (2005) *J. Biol. Chem.* **280**, 7131–34.

Bjellqvist, B., Hughes, G.J., Pasquali, C., *et al.* (1993) *Electrophoresis* **14**, 1023–31.

Boxer, D.H., Zhang, H., Gourley, D.G., *et al.* (2004) *Org. Biomol. Chem.* **2**, 2829–37.

Chang, A.C.Y, Nunberg, J.H., Kaufman, R.J., *et al.* (1978) *Nature* **275**, 617–24.

Cheng, Y. and Prusoff, W.H. (1973) *Biochem. Pharmacol.* **22**, 3099–108.

Chou, P.Y. and Fasman, G.D. (1974) *Biochemistry* **13**, 222–45.

Chou, W.Y. and Matthews, K. S. (1989) *J Biol Chem* **264**, 18314–19.

Cowan, D.A. (1995) *Essays Biochem.* **29**, 193–207.

Engel, P.C. (1981) *Enzyme Kinetics*, 2nd edition, Chapman & Hall, London.

Evrard, C., Fastrez, J., and Declercq, J.-P. (1998) *J. Mol. Biol.* **276**, 151–64.

Groch, N., Schindelin, H., Scholtz, A. S., *et al.* (1992) *Eur. J. Biochem.* **207**, 677–85.

Groll, M., Koguchi, Y., Huber, R. and Kohno, J. (2001) *J. Mol. Biol.* **311**, 543–48.

Gu, Y.Q., Holzer, F.M. and Walling, L.L. (1999) *Eur. J. Biochem.* **263**, 726–35.

Kelly, S.M., Jess, T.J., and Price, N.C. (2005) *Biochim. Biophys. Acta* **1751**, 119–39.

Magliery, T.J. (2005) *Med. Chem. Rev.* **2**, 303–23.

Miller, M. Shuman, J.D., Sebastian, T., *et al.* (2003) *J. Biol. Chem.*, **278**, 15178–84.

Pace, C.N., Vajdos, F., Fee, L., *et al.* (1995) *Protein Sci.* **11**, 2411–23.

Rost, B. (1996) *Methods Enzymol.* **266**, 525–39.

Rost, B. (1999) *Protein Eng.* **12**, 85–94.

Sander, C. and Schneider, R. (1994) *Nucl. Acids Res.* **22**, 3597–99.

Theerasilp, S. and Kurihara, Y. (1988) *J. Biol. Chem.* **263**, 11536–39.

Viles, J.H., Cohen, F.E., Prusiner, S.B., *et al.* (1999) *Proc. Natl. Acad. Sci. USA* **96**, 2042–47.

Wang, L., Xie, J., Deniz, A.A. and Schultz, P.G. (2003) *J. Org. Chem.* **68**, 174–76.

Wlodawer, A. and Vondrasek, J. (1998) *Annu. Rev. Biophys. Biomol. Struct.*, 27, 249–84.

# Appendix

## Appendix 5.1 Derivation of the Cheng–Prusoff equation (Cheng and Prusoff, 1973)

According to the Michaelis–Menten equation (Chapter 4, eqn. 4.26), the rate of a reaction ($v_1$) at a substrate concentration [$S$] is:

$$v_1 = \frac{V_{max} \times [S]}{K_m + [S]}$$

In the presence of a competitive inhibitor at a concentration [$I$] with an inhibitor constant $K_{EI}$, the $K_m$ is increased by a factor $\left(1 + \frac{[I]}{K_{EI}}\right)$ (Chapter 4, section 4.3.4), so the new rate ($v_2$) is given by:

$$v_2 = \frac{V_{max}[S]}{K_m\left(1 + \frac{[I]}{K_{EI}}\right) + [S]} \qquad \textbf{5.4}$$

If $[I] = IC_{50}$, $v_2 = 0.5 \times v_1$, i.e. $v_1 = 2 \times v_2$

$$\frac{V_{max} \times [S]}{K_m + [S]} = \frac{2 \times V_{max} \times [S]}{K_m\left(1 + \frac{IC_{50}}{K_{EI}}\right) + [S]}$$

1) Dividing both sides by $V_{max} \times [S]$, and then cross-multiplying we obtain:

2) $2 \times K_m + 2 \times [S] = K_m + \frac{K_m \times IC_{50}}{K_{EI}} + [S]$

3) Collecting terms in $K_m$ and [$S$], we obtain:

4) $K_m + [S] = \frac{K_m \times IC_{50}}{K_{EI}}$

5) Rearranging we obtain:

6) $K_{EI} = \dfrac{K_m \times IC_{50}}{K_m + [S]}$

7) Dividing the top and bottom terms on the right-hand side by $K_m$, we obtain:

8) $$K_{EI} = \frac{IC_{50}}{1 + \dfrac{[S]}{K_m}} \qquad \textbf{5.5}$$

which is the Cheng–Prusoff equation.