Molecular Biology

FOURTH EDITION

Alexander McLennan
Andy Bates
Phil Turner
Mike White

GS
Garland Science

BIOS INSTANT NOTES

Molecular Biology

FOURTH EDITION

BIOS INSTANT NOTES

Molecular Biology

FOURTH EDITION

Alexander McLennan, Andy Bates,
and Phil Turner
Institute of Integrative Biology
University of Liverpool, Liverpool, UK

Mike White
Faculty of Life Sciences
University of Manchester, Manchester, UK

Garland Science
Taylor & Francis Group
NEW YORK AND LONDON

Garland Science
Vice President: Denise Schanck
Editor: Elizabeth Owen
Editorial Assistant: Vicky Noyes
Production Editor: Ioana Moldovan
Copyeditor: Alison Gibbs
Typesetting and illustrations: Phoenix Photosetting, Chatham, Kent
Proofreader: Dawn Booth
Printed by: T. J. International

ISBN 978-0-4156-8416-3

Library of Congress Cataloging-in-Publication

Molecular biology / Alexander McLennan ... [et al.]. — 4th ed.
p. ; cm. — (BIOS instant notes)
Rev. ed. of: Molecular biology / Phil Turner ... [et al.]. 3rd ed. New York, NY : Taylor & Francis, 2005.
Includes bibliographical references and index.
Summary: "Instant Notes in Molecular Biology, Fourth Edition is the perfect text for undergraduates looking for a concise introduction to the subject, or a study guide to use before examinations. Each topic begins with a summary of essential facts, an ideal revision checklist followed by a description of the subject that focuses on core information, with clear, simple diagrams that are easy for students to understand and recall in essays and exams"—Provided by publisher.
ISBN 978-0-415-68416-3 (pbk.)
1. Molecular biology—Outlines, syllabi, etc. I. McLennan, Alexander G. II. Series: BIOS instant notes.
[DNLM: 1. Molecular Biology—Outlines. QH 506]
QH506.I4815 2013
572.8—dc23 2012013042

Published by Garland Science, Taylor & Francis Group, LLC, an informa business, 711 Third Avenue, 8th Floor, New York NY 10017, USA, and 3 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Visit our web site at http://www.garlandscience.com

Preface to the fourth edition

There can be few scientific disciplines that have advanced as rapidly as molecular biology in the years since the last edition. Thus, our hope that the improvements made to the last edition would make future changes simpler was rather naïve. Our revisions and updates to the fourth edition have been major and have involved all Sections and Topics; such has been the rate of progress. The first two Sections of the third edition have been combined and simplified to reduce overlap with other titles in the *Instant Notes* series, whereas other Sections have been reorganized and restructured in a more logical fashion. This has created room for the consideration of advances in topics such as 'next-generation' DNA sequencing and genomics, global gene expression analysis, regulatory RNAs, proteomics, stem cells, systems biology, and many other areas. One difficulty has been judging what to leave out in order to accommodate new material. As knowledge expands and technology advances, and the old ways of doing things fall out of favor, the challenge is to keep the reader excited with new discoveries and with the power of new methods while keeping enough of the traditional background story to allow complete understanding of a topic. We are also well aware that this is an introductory textbook and so we have tried to avoid unnecessary complexity and detail. We hope we have achieved our aims. As always, we are most grateful to those many reviewers who made suggestions for improvements to the previous edition and are particularly indebted to Liz Owen and Vicki Noyes for their patience and understanding during the revision process.

Alexander McLennan, Andy Bates, Phil Turner, and Mike White

November 2011

Contents

A1 Information processing and molecular biology

Key Notes

The 'central dogma'	The central dogma is the original proposal that 'DNA makes RNA makes protein,' which happens via the processes of transcription and translation respectively. This is broadly correct, although a number of examples are known that contradict parts of it. Retroviruses reverse transcribe RNA into DNA, other viruses can replicate RNA directly into an RNA copy, whereas some RNAs can be edited after synthesis so that the resulting sequence is not directly specified by the DNA sequence.
Recombinant DNA technology	The ability to sequence and manipulate the genomes of microorganisms, animals, and plants has led to major advances in our understanding of cellular biology. In addition, transgenic organisms containing DNA from other sources have found many applications in medicine, agriculture, and industry. The ability to synthesize novel genomes will lead to even greater advances in these areas.
Related topics	(A2) Nucleic acid structure and function (A3) Protein structure and function (Section F) Transcription in bacteria (Section H) Transcription in eukaryotes (Section K) The genetic code and tRNA (Section L) Protein synthesis

The central dogma

Molecular biology is the study of the molecular reactions and interactions that underpin biological function. This overlaps considerably with biochemistry and genetics, and so it is often deemed mainly to deal with the structural basis and control of information processing in the cell and the technologies required to investigate these. Through the pioneering experiments of Avery, MacLeod, and McCarty, and Hershey and Chase in the 1940s and 1950s, it became firmly established that the genetic instructions for creating a cell were held in the nucleus within the linear sequence of **bases** contained in the structure of a long chemical polymer, **deoxyribonucleic acid** (**DNA**). Then, in 1953, the famous double helical structure of DNA was proposed by Crick and Watson, which revealed exactly how this information was stored and passed on to subsequent generations. To explain how cells use the instructions encoded in this DNA **genome**, Crick suggested that there was a unidirectional flow of genetic information from DNA through an intermediary nucleic acid, **ribonucleic acid** (**RNA**), to **protein**, i.e. 'DNA makes RNA makes protein.' This became known as the **central dogma** of molecular biology, as it was proposed without much evidence for the individual steps. We now know that the broad

thrust of the central dogma is correct, although a number of modifications have now been made to the original scheme. A diagrammatic version of this information flow is shown in Figure 1. The primary route remains from DNA to RNA to protein, and this is now known to include the DNA in the small, independent genomes of mitochondria and chloroplasts. In all cells, DNA is divided conceptually (though not physically) into discrete coding units (**genes**) that contain the information for individual proteins. This DNA is **transcribed** (Sections F and H) to yield RNA molecules (**messenger RNA, mRNA**) that contain the same sequence information as the DNA, and which can be regarded as working copies of the genes present in the master DNA blueprint. These mRNAs are then **translated** (Section L) into the amino acid sequences of proteins according to the **genetic code** (Section K1). The combination of all the processes that are required to decode the information in DNA to produce a functional molecule is called **gene expression**. We can also include DNA **replication** (Section D) in Figure 1, in which two daughter DNA molecules are formed by duplication of the information in the parent DNA, resulting in information flow and preservation from one generation to the next.

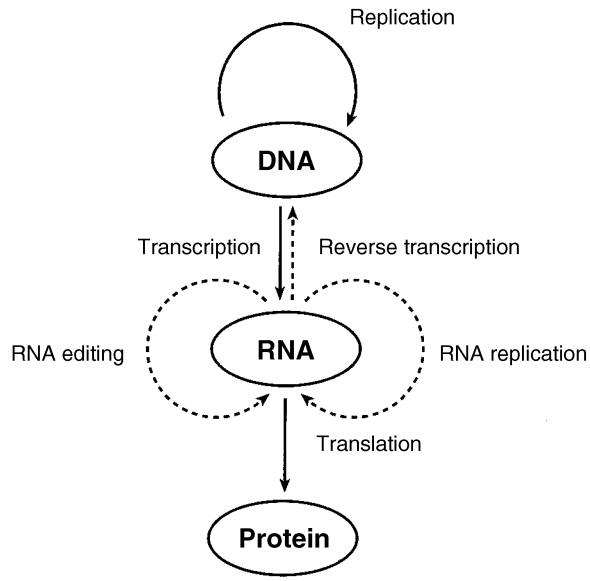

Figure 1. The flow of genetic information.

However, several exceptions to this basic scheme have been identified. Many RNA molecules are not translated into protein, but function as RNAs in their own right (Sections I2, L4, and J). Their genes are known as **RNA genes**. A number of classes of virus have no DNA but contain a genome consisting of one or more RNA molecules (Section M4). In the **retroviruses**, which include **human immunodeficiency virus (HIV)**, the causative agent of acquired immune deficiency syndrome (AIDS), the single-stranded RNA molecule is converted to a double-stranded DNA copy, which is then inserted into the genome of the host cell. This process has been termed **reverse transcription**. There are also a number of viruses known whose RNA genome is copied directly into RNA without the use of DNA as an intermediary (**RNA replication**) (Section M4). Examples include the influenza and hepatitis C viruses. As far as is known, there are no examples of a protein being 'reverse translated' to generate a specific RNA or DNA sequence, so the translation step of the

central dogma does appear to be unidirectional. Finally, one fascinating exception to the dogma that RNA and protein sequences are faithfully encoded in the DNA is the process of **RNA editing**. Examples are known (mainly in eukaryotes) where the base sequence of an RNA is actually altered after it is transcribed from the DNA so that it, and any protein product if it is an mRNA, no longer correspond precisely to the DNA (Section J4).

Discussion of these systems in this book is based on the **three-domain system** of biological classification in which the **last universal common ancestor** (**LUCA**) of all life first split into the **bacteria** and the common progenitor of the **archaea** and **eukarya**, which split later. Bacteria and archaea are both **prokaryotic**, in that they lack a nucleus, but in many aspects of information processing, archaea have more in common with the nucleated **eukaryotes**. Most examples are drawn from bacteria and eukaryotes.

Recombinant DNA technology

Major advances in molecular biology became possible in the late 1970s with the development of **recombinant DNA technology** (**genetic engineering**). This has allowed genes to be isolated, sequenced, modified, and transferred from one organism to another and has been of paramount importance in advancing our understanding of how cells work. Furthermore, **transgenic** microorganisms produced in this way are now routinely used to produce human therapeutics on a large scale while transgenic animals and plants have huge potential to increase the range of useful products as well as leading to improved growth, disease resistance, and models for human disease, etc. (Section S5). The permanent correction of genetic disease by **gene therapy** is also now a realistic possibility. In recent years, the technology behind determining DNA base sequences has progressed and the costs have fallen so rapidly that it will soon become feasible to sequence an individual's entire genome to determine disease susceptibility as part of a routine health care program. New genes can now even be chemically synthesized and assembled into complete genomes. In 2010, J. Craig Venter and colleagues recreated the complete chromosomal DNA of a small mycoplasma bacterium and inserted it into an 'empty' cell, denuded of its own chromosome, thus recreating the living organism (Section S7). This DNA molecule also had some novel features, paving the way to the future possibility of creating truly **synthetic life** – 'designer' organisms with artificial genomes, able to carry out novel biochemical functions not seen in the natural world, with the aim of producing new medicines, fuels, and other products. Thus, molecular biology and the technologies that have been created around it have played a central role in the development of human and animal medicine, agriculture, and the biotechnology industry, and are now set to meet the challenges of global health, environmental change, and food security that we face in the 21st century.

A2 Nucleic acid structure and function

Key Notes

Bases

In DNA, there are four heterocyclic bases: adenine (A) and guanine (G) are purines; cytosine (C) and thymine (T) are pyrimidines. In RNA, thymine is replaced by the structurally very similar pyrimidine, uracil (U).

Nucleosides

A nucleoside consists of a base covalently bonded to the 1′-position of a pentose sugar molecule. In RNA, the sugar is ribose and the compounds are ribonucleosides, or just nucleosides, whereas in DNA it is 2′-deoxyribose, and the nucleosides are named 2′-deoxyribonucleosides, or just deoxynucleosides. Base+sugar=nucleoside.

Nucleotides

Nucleotides are nucleosides with one or more phosphate groups covalently bound to the 3′-, 5′-, or, in some ribonucleotides, the 2′-position. Base+sugar+phosphate =nucleotide. The nucleoside 5′-triphosphates, NTPs and dNTPs, are the building blocks of polymeric RNA and DNA respectively.

Phosphodiester bonds

In nucleic acid polymers, the ribose or deoxyribose sugars are linked by a phosphate between the 5′-position of one sugar and the 3′-position of the next, forming a 3′,5′-phosphodiester bond. Hence, nucleic acids consist of a directional sugar-phosphate backbone with a base attached to the 1′-position of each sugar. The repeat unit is a nucleotide. Nucleic acids are highly charged polymers with a negative charge on each phosphate.

DNA/RNA sequence

The nucleic acid sequence is the sequence of bases A, C, G, T/U in the DNA or RNA chain. The sequence is conventionally written from the free 5′- to the free 3′-end of the molecule, for example 5′-ATAAGCTC-3′ (DNA) or 5′-AUAGCUUGA-3′ (RNA).

DNA double helix

DNA most commonly occurs as a double helix. Two separate and antiparallel chains of DNA are wound around each other in a right-handed helical (coiled) path, with the sugar-phosphate backbones on the outside and the bases, paired by hydrogen bonding and stacked on each other, on the inside. Adenine pairs with thymine; guanine pairs with cytosine. The two chains are complementary; one specifies the sequence of the other.

A, B, and Z helices	As well as the 'standard' DNA helix discovered by Watson and Crick, known as the B-form, and believed to be the predominant structure of DNA *in vivo,* nucleic acids can also form the right-handed A-helix, which is adopted by RNA sequences *in vivo* and the left-handed Z-helix, which only forms in specific alternating base sequences and is probably not a very important *in vivo* conformation.
RNA secondary structure	Most RNA molecules occur as a single strand, which may be folded into a complex conformation, involving local regions of intramolecular base pairing and other hydrogen bonding interactions. This complexity is reflected in the varied roles of RNA in the cell.
Modified nucleic acids	Covalent modifications of nucleic acids have specific roles in the cell. In DNA, these are mostly restricted to methylation of adenine and cytosine bases, but the range of modifications of RNA is much greater.
Nucleic acid function	DNA acts only as a carrier of expressible genetic information. However, the more versatile RNAs have numerous structural and functional roles in the mechanisms and regulation of information storage, flow, and processing.
Related topics	(B1) Chemical and physical properties of nucleic acids (B2) Spectroscopic and thermal properties of nucleic acids (B3) DNA supercoiling (Section C) Prokaryotic and eukaryotic chromosome structure

Bases

The **bases** of DNA and RNA are heterocyclic (carbon- and nitrogen-containing) aromatic rings, with a variety of substituents (Figure 1). Adenine (A) and guanine (G) are **purines**, bicyclic structures with two fused rings, whereas cytosine (C), uracil (U), and thymine (T) are monocyclic **pyrimidines**. In DNA, the uracil base of RNA is replaced by thymine. Thymine differs from uracil only in having a methyl group at the 5-position, i.e. thymine is 5-methyluracil.

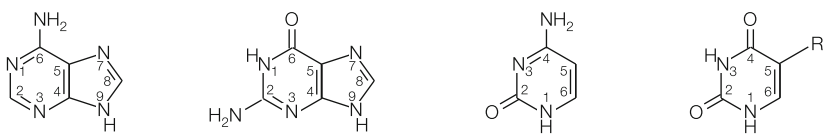

Figure 1. Nucleic acid bases.

Nucleosides

In nucleic acids, the bases are covalently attached to the 1′-position of a pentose sugar ring, to form a **nucleoside** (Figure 2). In RNA, the sugar is **ribose**, and in DNA it is **2′-deoxyribose**, in which the hydroxyl group at the 2′-position is replaced by a hydrogen. The point of attachment to the base is the 1-position (*N*-1) of the pyrimidines and the 9-position (*N*-9) of the purines (Figure 1). The numbers of the atoms in the ribose ring are designated 1′-, 2′-, etc., merely to distinguish them from the base atoms. The bond between the bases and the sugars is the **glycosylic** (**or glycosidic**) **bond**. If the sugar is ribose, the nucleosides (technically **ribonucleosides**) are adenosine, guanosine, cytidine, and uridine. If the sugar is deoxyribose (as in DNA), the nucleosides (**2′-deoxyribonucleosides**) are deoxyadenosine, etc. The terms thymidine and deoxythymidine may be used interchangeably.

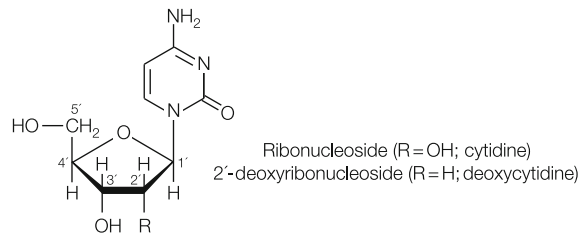

Figure 2. Nucleosides.

Nucleotides

A **nucleotide** is a nucleoside with one or more phosphate groups bound covalently to the 3′-, 5′-, or (in some **ribonucleotides** only) the 2′-position. If the sugar is deoxyribose, then the compounds are termed **2′-deoxyribonucleotides**, or just **deoxynucleotides** (Figure 3). Chemically, the compounds are phosphate esters. In the case of the 5′-position, up to three phosphates may be attached, to form, for example, adenosine 5′-triphosphate, or deoxyguanosine 5′-triphosphate, commonly abbreviated to ATP and dGTP respectively. In the same way, we have deoxycytidine triphosphate (dCTP), uridine triphosphate (UTP) and deoxythymidine triphosphate (dTTP; also just called TTP). 5′-mono and -diphosphates are abbreviated as, for example, AMP and dGDP. Nucleoside 5′-triphosphates (NTPs), or deoxynucleoside 5′-triphosphates (dNTPs) are the building blocks of the polymeric nucleic acids. In the course of DNA or RNA synthesis, two phosphates are split off as pyrophosphate to leave one phosphate per nucleotide incorporated

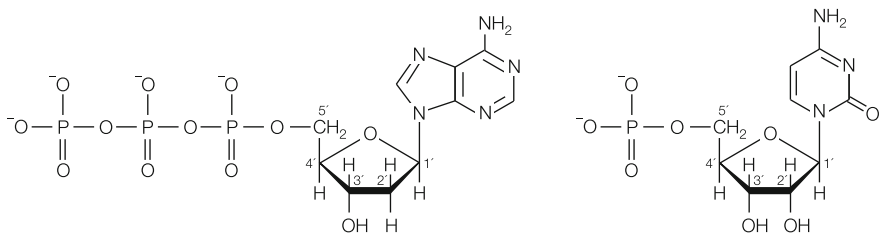

Figure 3. Nucleotides.

into the nucleic acid chain (Sections D1 and F1). Hence, the repeat unit of a DNA or RNA chain is a nucleotide.

Phosphodiester bonds

In a DNA or RNA molecule, deoxyribonucleotides or ribonucleotides respectively are joined into a polymer by the covalent linkage of a phosphate group between the 5′-hydroxyl of one ribose and the 3′-hydroxyl of the next (Figure 4). This kind of bond or linkage is called a **phosphodiester bond**, since the phosphate is chemically in the form of a diester. Thus, a nucleic acid chain can be seen to have a direction, or **polarity**. Any nucleic acid chain, of whatever length (unless it is circular, Section B3), has a free 5′-end, which may or may not have any attached phosphate groups, and a free 3′-end, which is most likely to be a free hydroxyl group. At neutral pH, each phosphate group has a single negative charge. This is why nucleic acids are termed acids; they are the anions of strong acids. Nucleic acids are thus **highly negatively charged polymers**.

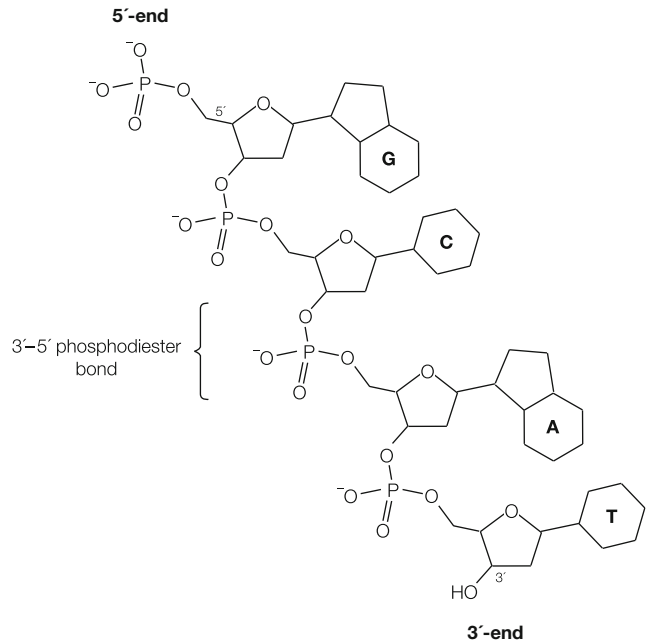

Figure 4. Phosphodiester bonds and the covalent structure of a DNA strand.

DNA/RNA sequence

Conventionally, the repeating monomers of DNA or RNA are represented by their single letters A, T, G, C, or U. In addition, there is a convention to write the sequences with the 5′-end to the left. Hence a stretch of DNA sequence might be written 5′-ATAAGCTC-3′, or even just ATAAGCTC. An RNA sequence might be 5′-AUAGCUUGA-3′. Note that the directionality of the chain means that, for example, ATAAG is not the same as GAATA.

DNA double helix

DNA most commonly occurs in nature as the well-known **double helix**. The basic features of this structure were deduced by James Watson and Francis Crick in 1953. Two

separate chains of DNA are wound around each other, each following a helical (coiling) path, resulting in a **right-handed** double helix (Figure 5a). The negatively charged sugar-phosphate backbones of the molecules are on the outside, and the planar bases of each strand stack one above the other in the center of the helix (Figure 5b). Between the backbone strands run the **major** and **minor grooves**, which also follow a helical path. The strands are joined noncovalently by hydrogen bonding between the bases on opposite strands, to form **base pairs** (**bp**). There are around 10 bp/turn in the DNA double helix. The two strands are oriented in opposite directions (**antiparallel**) in terms of their 5′→3′ direction and, most crucially, the two strands are **complementary** in terms of sequence. This last feature arises because the structures of the bases and the constraints of the DNA backbone dictate that the bases hydrogen-bond (Section A3) to each other as purine–pyrimidine pairs, which have very similar geometry and dimensions (Figure 6). Guanine pairs with cytosine (three H-bonds) and adenine pairs with thymine (two H-bonds). Hence, any sequence can be accommodated within a regular double-stranded DNA structure. The sequence of one strand uniquely specifies the sequence of the other, and Watson and Crick were quick to realize that this fact implies an obvious mechanism for the replication of DNA (Section D1). Of course, it also underlies the mechanism of transcription of DNA sequence into RNA (Section F1).

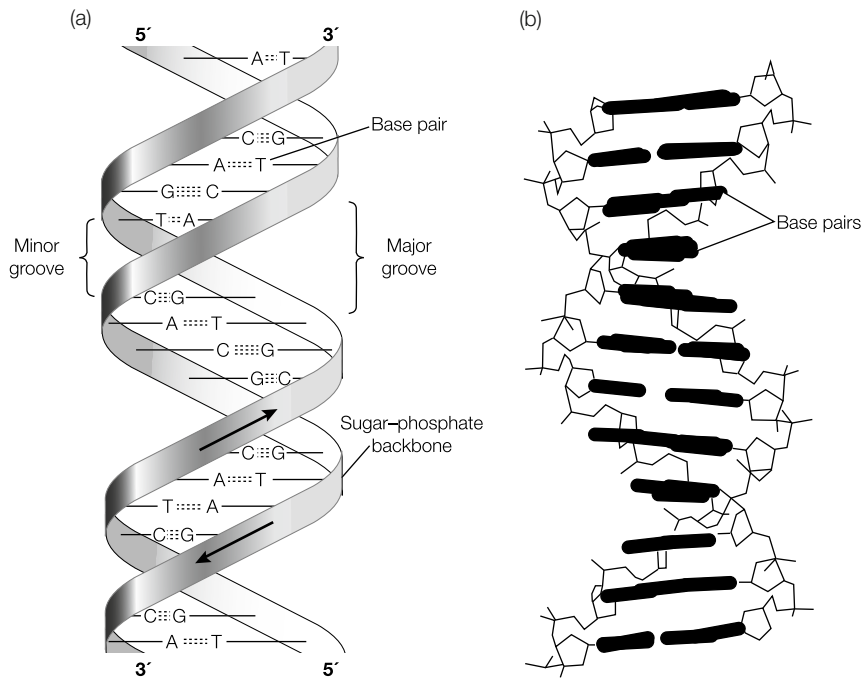

Figure 5. The DNA double helix. (a) A schematic view of the structure; (b) a more detailed structure, highlighting the stacking of the base pairs (in bold).

A, B, and Z helices

In fact, a number of different forms of nucleic acid double helix have been observed and studied, all having the basic pattern of two helically wound antiparallel strands. The structure identified by Watson and Crick, and described above, is known as **B-DNA** (Figure 7a),

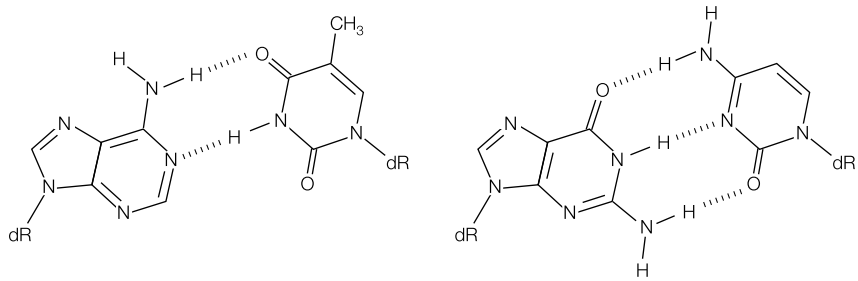

Figure 6. The DNA base pairs. Hydrogen bonds are shown as dashed lines; dR=deoxyribose.

and is believed to be the idealized form of the structure adopted by virtually all DNA *in vivo*. It is characterized by a helical repeat of 10 bp/turn, by the presence of base pairs lying on the helix axis and almost perpendicular to it, and by having well defined, deep major and minor grooves. Actually, real DNA sequences have a helical repeat closer to 10.5 bp/turn and have a variety of other structural distortions that depend on the exact base sequence.

DNA can be induced to form an alternative helix, known as the **A-form** (Figure 7b) under conditions of low humidity. The A-form is right-handed, like the B-form, but has a wider, more compressed structure in which the base pairs are tilted with respect to the helix axis, and actually lie off the axis (seen end-on, the A-helix has a hole down the middle). The helical repeat of the A-form is around 11 bp/turn. Although it may be that the A-form, or something close to it, is adopted by DNA *in vivo* under unusual circumstances, the major importance of the A-form is that it is the helix formed by RNA (see below), and

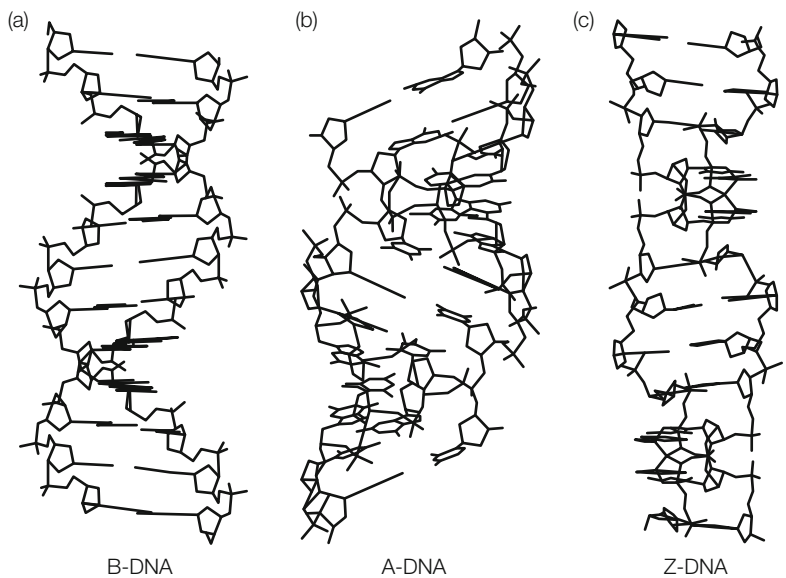

Figure 7. The alternative helical forms of the DNA double helix.

by DNA–RNA hybrids; it turns out that it is impossible to fit the 2′-OH of RNA into what would otherwise be the more stable B-form structure.

A further unusual helical structure can be formed by DNA. The left-handed **Z-DNA** (Figure 7c) is stable in synthetic double-stranded DNA consisting purely of alternating pyrimidine–purine sequence (such as 5′-CGCGCG-3′, with the same in the other strand, of course). This is because, in this structure, the pyrimidine and the purine nucleotides adopt very different conformations, unlike in A- and B-form, where each nucleotide has essentially the same conformation and immediate environment. In particular, the purine nucleotides in the Z-form adopt the *syn* conformation, in which the purine base lies directly above the deoxyribose ring (imagine rotating the base through 180° around the glycosylic bond in Figure 3; the nucleotides shown there are in the alternative *anti* conformation). The pyrimidine nucleotides in Z-DNA and all nucleotides in the A- and B-forms adopt the *anti* conformation. The Z-helix has a zigzag appearance, with 12 bp/turn, although it probably makes sense to think of it as consisting of six 'dimers of base pairs' per turn; the repeat unit along each strand is really a dinucleotide. Z-DNA does not easily form in normal DNA, even in regions of repeating CGCGCG, since the boundaries between the left-handed Z-form and the surrounding B-form would be very unstable. The Z-form is probably not a common feature of DNA (or RNA) *in vivo*, although it has been proposed to play a role in the dissipation of torsional stress (Section B3) in DNA during transcription in some specific cases. A comparison of the A, B, and Z helices is shown in Table 1.

Table 1. Summary of the major features of A, B, and Z nucleic acid helices

	A-form	B-form	Z-form
Helical sense	Right handed	Right handed	Left handed
Diameter	~2.6 nm	~2.0 nm	~1.8 nm
Base pairs per helical turn (*n*)	11	10	12 (6 dimers)
Helical twist per bp (= 360/*n*)	33°	36°	60° (per dimer)
Helix rise per bp (*h*)	0.26 nm	0.34 nm	0.37 nm
Helix pitch (= *nh*)	2.8 nm	3.4 nm	4.5 nm
Base tilt to helix axis	20°	6°	7°
Major groove	Narrow/deep	Wide/deep	Flat
Minor groove	Wide/shallow	Narrow/deep	Narrow/deep
Glycosylic bond	*anti*	*anti*	*anti* (pyr) *syn* (pur)

RNA secondary structure

RNA normally occurs as a single-stranded molecule, and hence it does not adopt a long regular helical structure like double-stranded DNA. RNA instead forms relatively globular conformations, in which local regions of helical structure are formed where one part of the RNA chain is complementary to another by **intramolecular** hydrogen bonding and base stacking within the single nucleic acid chain to form **hairpin** and **stem-loop** structures (Sections F1, Figure 3 and K2, Figure 2). This conformational variability is reflected in the more diverse roles of RNA in the cell, when compared with DNA (see below).

Modified nucleic acids

The chemical modification of bases or nucleotides in nucleic acids is widespread, and has a number of specific roles. In cellular DNA, the modifications are restricted to the methylation of the *N*-6 position of adenine, and the *N*-4 and 5-positions of cytosine (Figure 1), although more complex modifications occur in some phage DNAs. These methylations have a role in restriction modification (Section O3), base mismatch repair (Section E3) and eukaryotic genome structure and expression (Sections C2 and C3). A much more diverse range of modifications occurs in RNA after transcription, which again reflects the different roles of RNA in the cell. These are considered in more detail in Sections J3 and K2.

Nucleic acid function

DNA functions exclusively as a carrier of genetic information from generation to generation and, in that role, as a template for the synthesis of complementary RNA species. In contrast, although they can function as genomes or templates themselves (e.g. **RNA viruses** and **telomerase RNA**; Sections M4 and D3) and act as intermediates in the flow of information from DNA to protein (**mRNA**; Section A1), RNA molecules are less reliable as permanent stores of information due to their inherent chemical instability (Section B1). However, as they can achieve a wide range of tertiary structures and base-pair to DNA, many RNAs have additional functions similar to proteins. These highly abundant RNAs are called **noncoding RNAs** (**ncRNAs**) as they are not translated into protein, and their genes are known as **RNA genes**. Many are structural and functional components of the **pre-mRNA** processing and protein synthesis machineries (e.g. **snRNA, tRNA, rRNA**, and **7SL RNA**; Sections J and L), whereas some, known as **ribozymes**, have catalytic activity (e.g. **rRNA** again, and **RNase P**; Sections J2 and L2). Recently, it has become clear that a surprisingly large part of the eukaryotic genome encodes further ncRNAs that are essential for the control of gene expression. These are divided into **lncRNAs** (>200 nt), which are primarily involved in transcriptional control (Section I2), and the smaller (<200 nt) **miRNAs, siRNAs,** and **piRNAs** that are mainly involved in translational control (Section L4), although the size and functional distinctions are not absolute. As RNA has the ability to store genetic information and also catalyze and control chemical reactions, life based on RNA may have predated the existing system of life based on DNA, RNA, and proteins.

A3 Protein structure and function

Key Notes

Amino acid structure

The 20 common amino acids found in proteins have a chiral α-carbon atom linked to a proton, amino and carboxyl groups, and a specific side chain that confers different physical and chemical properties. These side chains may be basic (positively charged), acidic (negatively charged), hydrophobic (both aliphatic and aromatic) or possess other specific functional groups, e.g. hydroxyls, amides or thiols. They behave as zwitterions in solution. With two notable exceptions, nonstandard amino acids in proteins are formed by post-translational modification.

Protein sizes and shapes

Globular proteins, including most enzymes, behave in solution like compact, roughly spherical particles. Fibrous proteins have a high axial ratio and are often of structural importance, for example fibroin and keratin. Sizes range from a few thousand to several million Daltons. Some proteins have associated nonproteinaceous material, for example lipid or carbohydrate or small cofactors.

Primary structure

Amino acids are linked by peptide bonds between α-carboxyl and α-amino groups. The resulting polypeptide sequence has an N-terminus and a C-terminus. Polypeptides commonly have between 100 and 1500 amino acids linked in this way.

Noncovalent interactions

A large number of weak interactions maintain the three-dimensional structure of proteins. Charge–charge, charge–dipole and dipole–dipole interactions involve attractions between fully or partially charged atoms. Hydrogen bonds and hydrophobic interactions that exclude water are also important.

Secondary structure

Polypeptides can fold into a number of regular structures. The right-handed α-helix has 3.6 amino acids per turn and is stabilized by hydrogen bonds between peptide N–H and C=O groups three residues apart. Parallel and antiparallel β-pleated sheets are stabilized by hydrogen bonds between different portions of the polypeptide chain.

Tertiary structure

The different sections of secondary structure and connecting regions fold into a well-defined tertiary structure, with hydrophilic amino acids mostly on the surface and hydrophobic ones in the interior. The structure is stabilized by noncovalent interactions and, sometimes, disulfide

	bonds. Denaturation leads to loss of secondary and tertiary structure.
Quaternary structure	Many proteins have more than one polypeptide subunit. Hemoglobin has two α and two β chains. Large complexes such as microtubules are constructed from the quaternary association of individual polypeptide chains. Allosteric effects usually depend on subunit interactions.
Prosthetic groups	Some proteins have associated nonprotein molecules (prosthetic groups) that provide additional chemical functions to the protein. Small prosthetic groups include nicotinamide adenine dinucleotide (NAD^+), heme, and metal ions, for example Zn^{2+}.
Domains, motifs, families, and evolution	Domains form semi-independent structural and functional units within a single polypeptide chain. New proteins can evolve through new combinations of domains. Motifs are groupings of primary or secondary structural elements often found in related members of protein families. Protein families arise through gene duplication and subsequent divergent evolution of the new genes.
Protein function	Proteins have a wide variety of functions. They can act as enzymes, antibodies, structural components inside and outside the cell, receptors and transporters for chemical ligands, regulators, and nutritional stores.
Related topics	(A4) Macromolecular assemblies (Section L) Protein synthesis

Amino acid structure

Proteins are polymers of L-amino acids. Apart from **proline**, all of the 20 amino acids found in proteins have a common structure in which a carbon atom (the α-carbon) is linked to a carboxyl group, a primary amino group, a proton and a **side chain** (R) that is different in each amino acid (Figure 1). Except in **glycine**, the α-carbon atom is asymmetric – it has four chemically different groups attached. Thus, amino acids can exist as pairs of optically active stereoisomers (D- and L-). However, only the L-isomers are found in proteins. As glycine, the simplest amino acid, has a hydrogen atom in place of a side chain, it is optically inactive. Amino acids are dipolar ions (**zwitterions**) in aqueous solution and behave as both acids and bases (they are **amphoteric**). The side chains differ in size, shape, charge, and chemical reactivity, and are responsible for the differences in the properties of different proteins (Figure 2). Many proteins also contain nonstandard

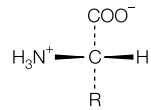

Figure 1. General structure of an L-amino acid. The R group is the side chain.

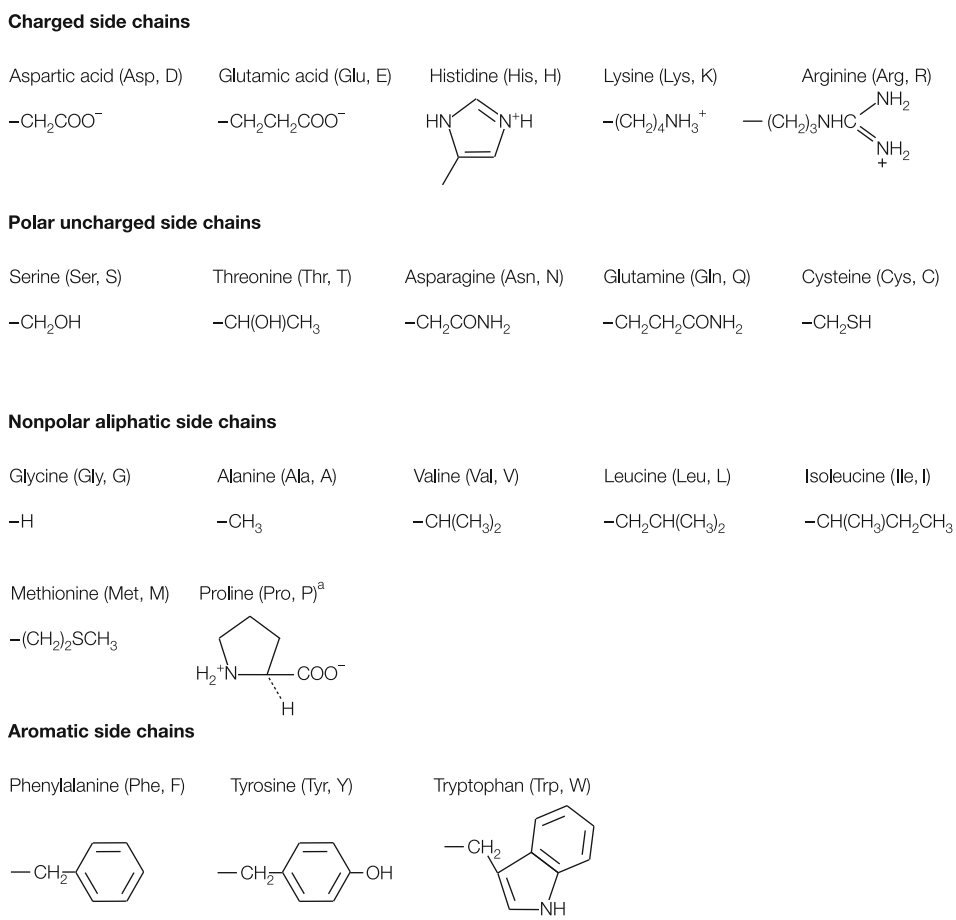

Figure 2. Side chains (R) of the 20 common amino acids. The standard three-letter abbreviations and one-letter code are shown in brackets. [a]The full structure of proline is shown as it is a secondary amino acid.

amino acids, such as 4-hydroxyproline and 5-hydroxylysine in collagen. These are mostly formed by **post-translational modification** of the parent amino acids, e.g. proline and **lysine**, in the newly synthesized protein (Section L4). However, **selenocysteine** (found in a number of enzymes), in which Se replaces the S of cysteine, and **pyrrolysine** (a modified lysine found only in certain archaeal proteins) are both incorporated into growing protein chains by a subtle manipulation of the genetic code (Section K1) and are regarded by some as the 21st and 22nd 'standard' amino acids.

Taking pH 7 as a reference point, several amino acids have ionizable groups in their side chains that provide an extra positive or negative charge at this pH. The 'acidic' amino acids, **aspartic acid** and **glutamic acid**, have additional carboxyl groups that are usually ionized (negatively charged). The 'basic' amino acids have positively charged groups – **lysine** has a second amino group attached to the ε-carbon atom while **arginine** has a guanidino group. The imidazole group of **histidine** has a pK_a near neutrality. Reversible protonation of this group under physiological conditions contributes to the catalytic mechanism of many enzymes. Together, acidic and basic amino acids can form important salt bridges in proteins.

Polar, uncharged side chains contain groups that form hydrogen bonds with water. Together with the charged amino acids, they are often described as **hydrophilic** ('water-loving'). **Serine** and **threonine** have hydroxyl groups that can be reversibly phosphorylated by protein kinases (see below) while **asparagine** and **glutamine** are the amide derivatives of aspartic acid and glutamic acid. **Cysteine** has a **thiol** (sulfhydryl) group, which often oxidizes to **cystine**, in which two cysteines form a structurally important disulfide bond.

Phenylalanine, **tyrosine** (which can also be phosphorylated), and **tryptophan** have bulky **hydrophobic** ('water-hating') side chains that participate in **hydrophobic interactions** in protein structure (see below). The aromatic structures of tyrosine and tryptophan account for most of the ultraviolet (UV) absorbance of proteins, which absorb maximally at 280 nm. The phenolic hydroxyl group of tyrosine can also form hydrogen bonds. Other nonpolar, hydrophobic side chains include the aliphatic alkyl groups of **alanine**, **valine**, **leucine**, **isoleucine**, and **methionine**, which contains a **sulfur** atom in a thioether link, and the cyclic ring of proline, which is unusual in being a secondary amino (or **imino**) acid.

Protein sizes and shapes

Two broad classes of protein may be distinguished. **Globular proteins** are folded compactly and behave in solution more or less as spherical particles; most enzymes are globular in nature. **Fibrous proteins** have very high axial ratios (length/width) and are often important structural proteins, for example silk fibroin and keratin in hair and wool. Molecular masses can range from a few thousand Daltons (Da), e.g. the hormone insulin with 51 amino acids and a molecular mass of 5734 Da (5.7 kiloDaltons, kDa), to nearly 4 million Da (4 MDa) in the case of the muscle protein titin. Some proteins contain bound **nonproteinaceous** material, either in the form of small **prosthetic groups**, which may act as cofactors in enzyme reactions, or as large associations (e.g. the lipids in **lipoproteins** or the carbohydrate in **glycoproteins**, Section A4).

Primary structure

The α-carboxyl group of one amino acid is covalently linked to the α-amino group of the next amino acid by an amide bond, commonly known as a **peptide bond** when in proteins. When two amino acid **residues** are linked in this way the product is a **dipeptide**. Many amino acids linked by peptide bonds form a **polypeptide** (Figure 3). The repeating sequence of α-carbon atoms and peptide bonds provides the structural **backbone** of the polypeptide while the different amino acid **side chains** confer functionality on the protein. The amino acid at one end of a polypeptide has an unattached α-amino group while the one at the other end has a free α-carboxyl group. Hence, polypeptides are directional, with an **N-terminus** and a **C-terminus**. Sometimes the N-terminus is **blocked** with, for example, an acetyl group. The sequence of amino acids from the N- to

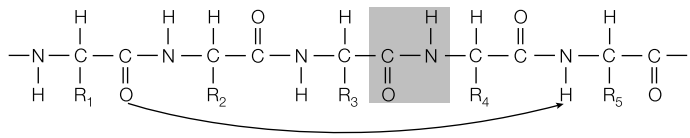

Figure 3. Section of a polypeptide chain. The peptide bond is boxed. In the α-helix, the CO group of amino acid residue n is hydrogen-bonded to the NH group of residue *n*+4 (arrowed).

the C-terminus is the **primary structure** of the polypeptide. Typical sizes for single polypeptide chains are within the range 100–1500 amino acids, although shorter and longer ones exist, e.g. titin has around 34,000.

Noncovalent interactions

The three-dimensional structure of proteins and large, protein-containing assemblies (Section A4) is maintained by many different noncovalent interactions. Electrostatic **charge–charge** interactions (**salt bridges**) operate between ionizable groups of opposite charge at physiological pH, e.g. between positive lysine and arginine side chains and negative glutamic acid and aspartic acid side chains or the negative phosphates of DNA in DNA-binding proteins such as histones (Section C2). **Charge–dipole** and **dipole–dipole** interactions are weaker and form when either or both of the participants is a dipole because of the asymmetric distribution of charge in the molecule (Figure 4a). Even uncharged groups like methyl groups can attract each other weakly through transient dipoles arising from the motion of their electrons (**dispersion forces**).

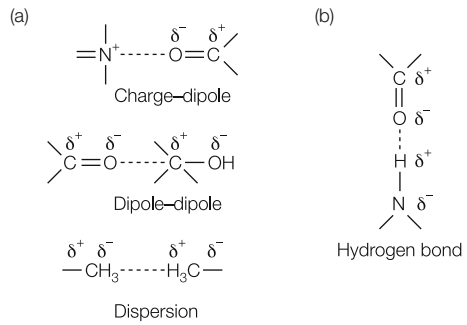

Figure 4. Examples of (a) van der Waals forces and (b) a hydrogen bond.

Noncovalent associations between electrically neutral molecules are known collectively as **van der Waals forces**. **Hydrogen bonds** are of great importance. They form between a covalently bonded hydrogen atom on a donor group (e.g. –O-H or –N-H) and a pair of nonbonding electrons on an acceptor group (e.g. :O=C– or :N–) (Figure 4b). Hydrogen bonds and other interactions involving dipoles are directional in character and so help define macromolecular shapes and the specificity of molecular interactions. The presence of uncharged and nonpolar substances, e.g. lipids, in an aqueous environment tends to force a highly ordered structure on the surrounding water molecules. This is energetically unfavorable, as it reduces the entropy of the system. Hence, nonpolar molecules and functional groups such as the aliphatic and aromatic amino acid side chains tend to clump together, reducing the overall surface area exposed to water. This attraction is termed a **hydrophobic** interaction and is a major stabilizing force in protein–protein and protein–lipid interactions and in nucleic acids (Section A2).

Secondary structure

The highly polar nature of the C=O and N–H groups of the peptide bonds gives the C–N bond partial double bond character. This makes the peptide bond unit rigid and planar, though there is free rotation between adjacent peptide bonds. This polarity also favors hydrogen bond formation between appropriately spaced and oriented peptide bond

units. Thus, polypeptide chains are able to fold into a number of regular structures that are held together by these hydrogen bonds. The best-known **secondary structure** is the **α-helix** (Figure 5a). The polypeptide backbone forms a right-handed helix with 3.6 amino acid residues per turn such that each peptide N–H group is hydrogen bonded to the C=O group of the peptide bond three residues away (Figure 3). Sections of α-helical secondary structure are often found in globular proteins and in some fibrous proteins. The rarer **3_{10}-helix** is similar, but with different dimensions. The **β-pleated sheet** (β-sheet) is formed by hydrogen bonding of the peptide bond N–H and C=O groups to the complementary groups of another section of the polypeptide chain (Figure 5b). Several sections of polypeptide chain may be involved side-by-side, giving a sheet structure with the side chains (R) projecting alternately above and below the sheet. If these sections run in the same direction (e.g. N-terminus→C-terminus), the sheet is **parallel**; if they alternate N→C and C→N, then the sheet is **antiparallel**. **Mixed β-sheets** comprising both orientations are also found. β-Sheets are strong and rigid and are important in structural proteins, for example silk fibroin. The connective tissue protein **collagen** has an unusual **triple helix** secondary structure in which three polypeptide chains are intertwined, making it very strong.

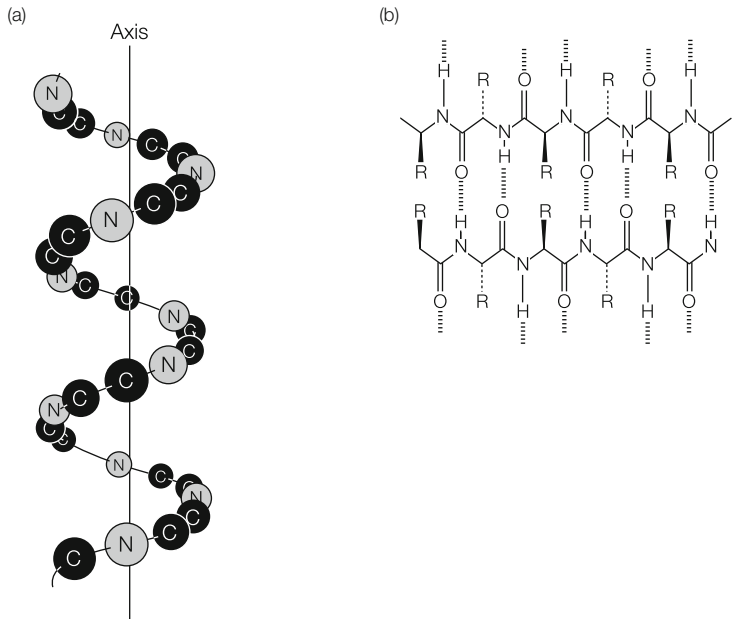

Figure 5. (a) α-Helix secondary structure. Only the α-carbon and peptide bond carbon and nitrogen atoms of the polypeptide backbone are shown for clarity. (b) Section of a β-sheet secondary structure.

Tertiary structure

The way in which the different sections of α-helix, β-sheet, other minor secondary structures and connecting, unstructured loops fold in three dimensions is the **tertiary structure** of the polypeptide (Figure 6). The nature of the tertiary structure is inherent in the primary structure and, given the right conditions, most polypeptides will fold

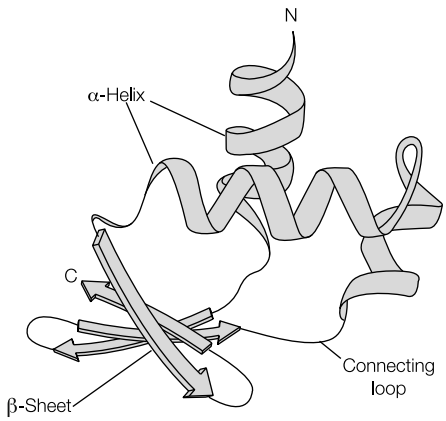

Figure 6. Schematic diagram of a section of protein tertiary structure.

spontaneously into the correct tertiary structure, as it is generally the lowest energy conformation for that sequence. However, proper folding *in vivo* is usually promoted by proteins called **chaperones,** which help prevent misfolding of new polypeptides before their synthesis (and primary structure) is complete. Folding is such that amino acids with hydrophilic side chains locate mainly on the exterior of the protein, where they can interact with water or solvent ions, whereas the hydrophobic amino acids become buried in the interior, from which water is excluded. This gives overall stability to the structure. Various types of noncovalent interaction between side chains hold the tertiary structure together: van der Waals forces, hydrogen bonds, electrostatic salt bridges between oppositely charged groups, and hydrophobic interactions between nonpolar side chains. In addition, covalent disulfide bonds can form between two cysteine residues that may be far apart in the primary structure but close together in the folded tertiary structure. Disruption of secondary and tertiary structure by heat or extremes of pH leads to **denaturation** of the protein and formation of a **random coil** conformation. Any associated biological activity is usually lost and the denatured proteins tend to clump into insoluble aggregates, as their exposed hydrophobic interiors interact to exclude water.

The importance of correct protein folding is illustrated by the fact that many neurodegenerative disorders such as **Alzheimer's disease** are associated with the accumulation of insoluble protein aggregates in neurons called **amyloid fibrils**, which contain extensive β-sheet structure. Also, the fatal diseases **scrapie** (sheep), **bovine spongiform encephalopathy** (**BSE**, 'mad cow disease') and **Creutzfeldt–Jakob disease** (**CJD**, humans) are caused by an infectious agent called a **prion,** which consists solely of a misfolded protein. When the prion enters a neuron it binds to a related cellular protein causing it to misfold, thus setting off a chain reaction of misfolding and amyloid formation, leading to loss of cell function.

Quaternary structure

Many proteins are composed of two or more polypeptide chains (**subunits**) forming **oligomers** – a few subunits, or **multimers** – many subunits. The subunits may be identical (**homomers**), or different (**heteromers**). For example, **hemoglobin** has two α-globin and two β-globin chains ($\alpha_2\beta_2$). The same forces that stabilize tertiary structure hold subunits together, including in some cases disulfide bonds between cysteines on

separate polypeptides. This level of organization is known as the **quaternary structure** and has certain consequences. First, it allows very large protein structures to be made, e.g. the **microtubules** of the cytoskeleton (Section A4, Figure 1). Secondly, it can provide greater functionality to a protein by combining different activities into a single entity, as in DNA polymerase III holoenzyme and the replisome (Section D2). Often, the interactions between the subunits are modified by the binding of small molecules and this can lead to the **allosteric** effects seen in enzyme regulation. There are also many examples of transient protein complexes, particularly in cell signaling pathways, where a post-translational modification (such as phosphorylation, Section L4) of one protein causes it to briefly associate with another, often resulting in a conformational change in the second protein that switches its function on or off.

Prosthetic groups

Many proteins contain covalently or noncovalently attached small molecules called **prosthetic groups** that give structural or chemical functionality to the protein that the amino acid side chains cannot provide. Many of these are **cofactors** in enzyme-catalyzed reactions. Examples are nicotinamide adenine dinucleotide (NAD^+) in many dehydrogenases, pyridoxal phosphate in transaminases, heme in hemoglobin and cytochromes, metal ions such as Zn^{2+}, and fatty acyl groups that can anchor proteins in cell membranes through hydrophobic interactions. Such proteins are termed **conjugated** proteins and the protein without its prosthetic group is known as an **apoprotein**. Other conjugated proteins contain associated macromolecules in large complexes such as carbohydrate (**glycoproteins**), lipid (**lipoproteins**), or nucleic acid (**nucleoproteins**) (Section A4).

Domains, motifs, families, and evolution

Many individual polypeptides are composed of structurally independent units, or **domains**, that are connected by sections with limited higher order structure within the same polypeptide. The connections can act as hinges to permit the individual domains to move in relation to each other, and breakage of these connections by limited proteolysis can often separate the domains, which can then behave like independent globular proteins. The active site of an enzyme is sometimes formed in a groove between two domains, which wrap around the substrate. Domains can also have a specific function such as binding a commonly used molecule, for example ATP. When such a function is required in many different proteins, the same domain structure is often found. In eukaryotes, domains are often encoded by discrete parts of genes called **exons** (Section J3). Therefore, it has been suggested that during evolution, new proteins were created by the duplication and rearrangement of domain-encoding exons in the genome to produce new combinations of binding sites, catalytic sites, and structural elements in the resulting new polypeptides. In this way, the rate of evolution of new functional proteins may have been greatly increased.

Structural motifs (also known as **supersecondary structures**) are groupings of secondary structural elements that frequently occur in globular proteins. They often have functional significance and can represent the essential parts of binding or catalytic sites that have been conserved during the evolution of protein families from a common ancestor. Alternatively, they may represent the best solution to a structural–functional requirement that has been arrived at independently in unrelated proteins. A common example is the βαβ **motif**, in which the connection between two consecutive parallel strands of a β-sheet is an α-helix (Figure 7). Two overlapping βαβ motifs (βαβαβ) form a dinucleotide (e.g. NAD^+) binding site in many otherwise unrelated proteins. **Sequence motifs** consist

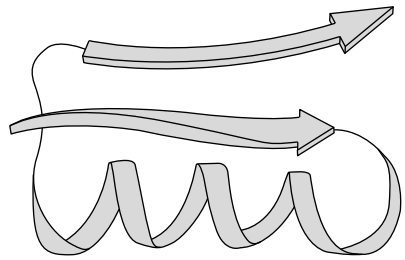

Figure 7. Representation of a βαβ motif. The α-helix is shown as a coiled ribbon and the β-sheet segments as flat arrows.

of linear sequences of conserved, functionally important amino acids, i.e. primary structure, rather than supersecondary structure. They can also represent binding or catalytic sites.

Protein families arise through successive duplications and subsequent **divergent evolution** of an ancestral gene. Myoglobin, the oxygen-carrying protein in muscle, the α- and β-globin chains and the minor δ-(delta) chain of adult hemoglobin and the γ- (gamma), ε- (epsilon) and ζ- (zeta) globins of embryonic and fetal hemoglobins are all related polypeptides within the **globin family** (Figure 8). Their genes, and the proteins, are said to be **homologs**. Family members in different species that have retained the same function and carry out the same biochemical role (e.g. rat and mouse myoglobin) are **orthologs** while those that have evolved different but often related functions (e.g. α-globin and β-globin) are **paralogs** (Section S6). The degree of similarity between the amino acid sequences of orthologous members of a protein family in different organisms depends on how long ago the two organisms diverged from their common ancestor and on how important conservation of the sequence is for the function of the protein. This function, whether structural or catalytic, is inherently related to its structure. As indicated above, similar structures and functions can also be achieved by **convergent evolution** whereby unrelated genes evolve to produce proteins with similar structures or catalytic activities. A good example is provided by the proteolytic enzymes **subtilisin** (bacterial) and **chymotrypsin** (animal).

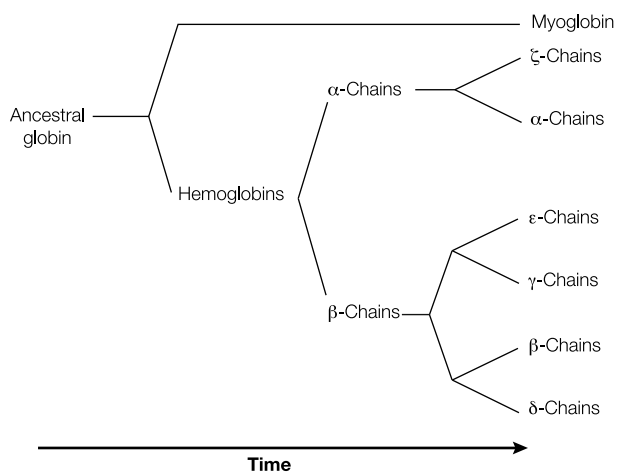

Figure 8. Evolution of globins from an ancestral globin gene.

Even though their amino acid sequences are very different and they are composed of different structural motifs, they have evolved the same spatial orientation of the **catalytic triad** of active site amino acids – serine, histidine, and aspartic acid – and use exactly the same catalytic mechanism to hydrolyse peptide bonds. Such proteins are termed **functional analogs.** Where similar structural motifs have evolved independently, the resulting proteins are **structural analogs**.

Protein function

- **Enzymes:** Apart from a few catalytically active RNA molecules (Section J2), all enzymes are proteins. These can enhance the rate of biochemical reactions by several orders of magnitude. Binding of the **substrate** involves various noncovalent interactions with specific amino acid side chains, including van der Waals forces, hydrogen bonds, salt bridges, and hydrophobic forces. Specificity of binding can be extremely high, with only a single substrate binding (e.g. glucose oxidase binds only glucose), or it can be group-specific (e.g. hexokinase binds a variety of hexose sugars). Side chains can also be directly involved in catalysis, for example by acting as nucleophiles, or proton donors, or abstractors.
- **Signaling:** Receptor proteins in cell membranes can bind **ligands** (e.g. hormones) from the extracellular medium and, by virtue of the resulting conformational change, initiate reactions within the cell in response to that ligand. Ligand binding is similar to substrate binding but the ligand usually remains unchanged. Some hormones are themselves small proteins, such as insulin and growth hormone. **Protein kinases**, which modify the properties of other proteins by adding a phosphoryl group from ATP to them, are extremely important enzymes in intracellular signaling.
- **Transport and storage: Hemoglobin** transports oxygen in the red blood cells while **transferrin** transports iron to the liver. Once in the liver, iron is stored bound to the protein **ferritin**. Dietary fats are carried in the blood by **lipoproteins**. Many other molecules and ions are transported and stored in a protein-bound form. This can enhance solubility and reduce reactivity until they are required.
- **Structure and movement: Collagen** is the major protein in skin, bone, and connective tissue, whereas hair is made mainly from **keratin**. There are also many structural proteins within the cell, for example in the **cytoskeleton**. The major muscle proteins **actin** and **myosin** form sliding filaments, which are the basis of muscle contraction.
- **Nutrition: Casein** and **ovalbumin** are the major proteins of milk and eggs, respectively, and are used to provide the amino acids for growth of developing offspring. Seed proteins also provide nutrition for germinating plant embryos.
- **Immunity: Antibodies**, which recognize and bind to bacteria, viruses and other foreign material (the **antigen**), are proteins.
- **Regulation: Transcription factors** bind to and modulate the function of DNA. Many other proteins modify the functions of other molecules by binding to them.

A4 Macromolecular assemblies

Key Notes

Large protein complexes

Proteins can associate with each other in extremely large structures. The eukaryotic cytoskeleton consists of various such complexes including microtubules (made of tubulin), microfilaments and muscle fibers (containing actin and myosin), cilia, and flagella. These organize the shape and movement of cells and subcellular organelles.

Conjugated proteins

Glycoproteins and proteoglycans (mucoproteins) are proteins with covalently attached carbohydrate and are generally found on extracellular surfaces and in extracellular spaces. Lipoproteins are used to transport lipids in aqueous environments. Mixed macromolecular complexes such as these provide a wider range of functions than the component parts.

Nucleoproteins

Bacterial 70S ribosomes comprise a large 50S subunit, with 23S and 5S ribosomal RNA (rRNA) molecules and 31 proteins, and a small 30S subunit, with a 16S rRNA molecule and 21 proteins. Eukaryotic 80S ribosomes have 60S (28S, 5.8S, and 5S rRNAs) and 40S (18S rRNA) subunits. Chromatin contains DNA and the basic histone proteins. Viruses are also nucleoprotein complexes.

Membranes

Membrane phospholipids and sphingolipids form bilayers with the polar groups on the exterior surfaces and the hydrocarbon chains in the interior. Membrane proteins may be peripheral or integral and act as receptors, enzymes, transporters, or mediators of cellular interactions.

Related topics

(A3) Protein structure and function
(C2) Chromatin structure
(J1) rRNA processing and ribosomes
(Section M) Bacteriophages and eukaryotic viruses

Large protein complexes

Few macromolecules work in isolation as monomers. They generally associate with other macromolecules of the same or a different class in stable or weak, transient complexes to carry out their functions. Some of these can be extremely large. For example, many of the major structural and locomotory elements of the cell consist of large protein complexes. The **cytoskeleton** is an array of protein filaments that organizes the shape and motion of cells and the intracellular distribution of subcellular organelles. **Microtubules** are 200- to 25,000-nm-long polymers of tubulin, a 110-kDa globular protein, which is itself a dimer

of distinct α and β subunits (Section A3) (Figure 1). These are a major component of the cytoskeleton, the **mitotic spindle** (Section C3), and of eukaryotic **cilia** and **flagella**, the hair-like structures on the surface of many cells that whip to move the cell or to move fluid across the cell surface. Cilia also contain the proteins nexin and dynein. **Microfilaments** consisting of the protein **actin** form huge contractile assemblies with the protein **myosin** to cause cytoplasmic motion. Actin and myosin are also major components of muscle fibers.

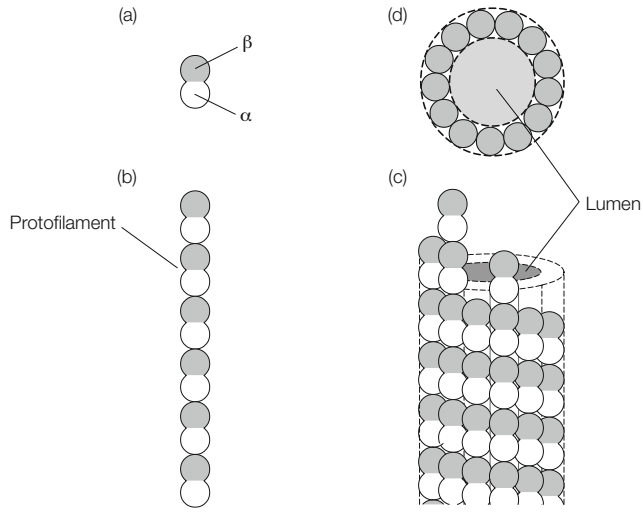

Figure 1. The structure of a microtubule: (a) tubulin consists of α- and β-subunits; (b) a tubulin protofilament consisting of many adjacent subunits; (c) the microtubule is formed from 13 protofilaments aligned in parallel; (d) cross-section of the hollow microtubule. From Hames and Hooper (2011) *Instant Notes Biochemistry*, 4th edn, Garland Science.

Conjugated proteins

Some conjugated proteins (Section A3) comprise associations of proteins with one or more of the other major classes of large biomolecules – carbohydrates, lipids, and nucleic acids. This can greatly increase the functionality or structural capabilities of the resulting complex.

Glycoproteins contain both protein and carbohydrate (between <1% and >90% of the weight) components; **glycosylation** is the commonest form of **post-translational modification** of proteins (Section L4). The carbohydrate is always covalently attached to the surface of the protein, never the interior, and is often variable in composition, causing microheterogeneity (Figure 2). This has made glycoproteins difficult to study. Glycoproteins have functions that span the entire range of protein activities, and are usually found extracellularly, either as secreted proteins or embedded in the plasma membrane (see below) where they can mediate cell–cell recognition or function as **receptors**. **Antibodies** and several protein hormones are glycoproteins.

Proteoglycans (**mucoproteins**) are large complexes (>10^7 Da) of protein and **mucopolysaccharides** (**glycosaminoglycans**) and are important components of the **extracellular matrix**, the material that binds and organizes cells in tissues. Their sugar units often have sulfate groups (e.g. **chondroitin sulfate**, **heparan sulfate**), which makes them highly charged and hydrated. This, coupled with their lengths (>1000 units), produces solutions

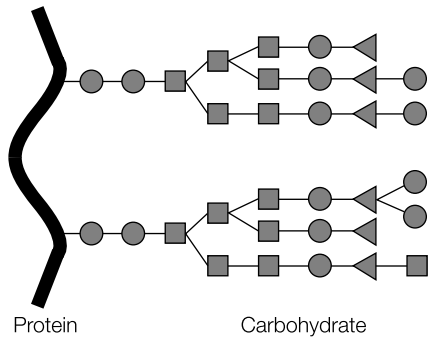

Figure 2. Glycoprotein structure. The different symbols represent different monosaccharide units (e.g. galactose, *N*-acetylglucosamine).

of high viscosity. In connective tissue, proteoglycans act as lubricants and shock absorbers in the extracellular space.

In **lipoproteins**, the lipids and proteins are linked noncovalently. Because lipids are poorly soluble in water, they are transported in the blood as lipoproteins. These are basically large particles of triglycerides and cholesterol esters coated with a layer of phospholipids, cholesterol and protein and vary in diameter and protein composition from 500 nm and 1–2% protein (**chylomicrons**) to 10 nm and 35% protein (**high-density lipoprotein, HDL**). The structures of the **apolipoproteins** (protein part without lipid) are such that their hydrophobic amino acids face towards the lipid interior of the particles while the charged and polar amino acids (Section A3) face outwards into the aqueous environment. This renders the particles soluble.

Nucleoproteins

Nucleoproteins, comprising both nucleic acid and protein, provide particularly important examples of conjugated proteins in molecular biology. Small **ribonucleoproteins** include **telomerase** (Sections C3 and D3) and the ribozyme **ribonuclease P** (Section J2), both of which contain RNA. **Ribosomes** are much larger ribonucleoprotein complexes and are the sites of protein synthesis in the cytoplasm (Section L). Bacterial 70S ribosomes have large (50S) and small (30S) subunits with a total mass of 2.5×10^6 Da. (The **S value**, e.g. 50S, is the numerical value of the **sedimentation coefficient**, *s*, and describes the rate at which a macromolecule or particle sediments in a centrifugal field. It is determined by both the mass and shape of the molecule or particle; hence S values are not additive.) The 50S subunit contains 23S and 5S ribosomal RNA (**rRNA**) molecules and 31 different proteins while the 30S subunit contains a 16S rRNA and 21 proteins. Under the correct conditions, mixtures of the rRNAs and proteins will self-assemble in a precise order into functional ribosomes *in vitro*. Thus, all the information for ribosome structure is inherent in the structures of the components. The rRNAs are not simply frameworks for the assembly of the ribosomal proteins, but participate in both the binding of the messenger RNA and in the catalysis of peptide bond synthesis (Section L2).

Chromatin is the material from which eukaryotic chromosomes are made. It is a **deoxyribonucleoprotein** complex made up of roughly equal amounts of DNA and small, basic proteins called **histones** that together form a discrete, repeating unit called a **nucleosome** (Section C2). Histones neutralize the repulsion between the negative charges of the DNA

sugar–phosphate backbone and allow the DNA to be tightly packaged within the chromosomes but are also of great functional importance in the control of gene expression. **Bacteriophages** and **viruses** are another example of nucleoprotein complexes. They are discussed in Section M.

Membranes

When placed in an aqueous environment, phospholipids and sphingolipids naturally form a **lipid bilayer** with the polar groups on the outside and the nonpolar hydrocarbon chains on the inside. This is the structural basis of all biological membranes. Such membranes form cellular and organellar boundaries and are selectively permeable to uncharged molecules. The precise lipid composition varies from cell to cell and from organelle to organelle. Proteins are also a major component of cell membranes (Figure 3). **Peripheral** membrane proteins are loosely bound to the outer surface or are anchored via a lipid or **glycosyl phosphatidylinositol** anchor and are relatively easy to remove. **Integral membrane proteins** are embedded in the membrane and cannot be removed without destroying the membrane. Some protrude from the outer or inner surface of the membrane while **transmembrane proteins** span the bilayer completely and have both extracellular and intracellular **domains** (Section A3). The transmembrane regions of these proteins contain predominantly hydrophobic amino acids. Many membrane proteins are also glycoproteins and have a variety of functions, for example:

- Receptors for signaling molecules such as hormones and neurotransmitters
- Enzymes for degrading extracellular molecules before uptake of the products
- Pores or channels for the selective transport of small, polar ions and molecules
- Mediators of cell–cell interactions (mainly glycoproteins)

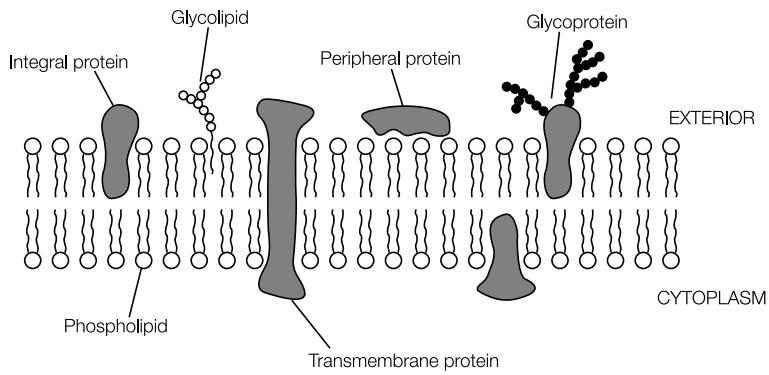

Figure 3. Schematic diagram of a plasma membrane showing the major macromolecular components.

A5 Analysis of proteins

Key Notes

Recombinant protein production

Purification of a protein from its native source for analysis is often inefficient and laborious and so they are now generally produced by inserting the gene encoding the protein into an expression vector and expressing in a suitable host cell. Sequence tags added to the recombinant protein simplify purification and detection, and aid solubility. Correct post-translational modifications may require a eukaryotic host–vector system.

Protein sequencing

After breaking a polypeptide down into smaller peptides using specific proteases, the peptides can be sequenced by chemical methods or by mass spectrometry. The original sequence is recreated from the overlaps produced by cleavage with proteases of different specificities. However, prediction of a protein sequence from the sequence of its gene or complementary DNA (cDNA) is simpler.

Biophysical methods

Many proteins can be crystallized and their three-dimensional structures determined by X-ray diffraction. The structures of small proteins in solution can also be determined by multi-dimensional nuclear magnetic resonance, particularly if the normal ^{12}C and ^{14}N are substituted by ^{13}C and ^{15}N. Many other techniques are available to study protein structures and interactions.

Mass determination and mass spectrometry

Approximate molecular masses can be obtained by gel electrophoresis in the presence of sodium dodecyl sulfate. Mass spectrometry using electrospray ionization gives masses that are accurate to within 0.01%. Mass spectrometry also detects post-translational modifications.

Applications for antibodies

Antibodies are proteins produced by the immune system of vertebrates in response to a foreign agent (the antigen), such as an injected protein. Their high binding affinities and specificities for the protein antigens make them useful laboratory tools for the detection and analysis of proteins by immunofluorescence, western blotting and immunoprecipitation.

Functional analysis

Functional analysis of a protein involves its isolation and study in vitro combined with a study of the behavior of a mutant organism in which the protein has been rendered nonfunctional by mutation or deletion of its gene. The function of a new protein can sometimes be predicted by comparing its sequence and structure to those of known proteins.

Related topics	(A2) Nucleic acid structure and function (A3) Protein structure and function	(S3) Proteomics (S6) Bioinformatics

Recombinant protein production

A typical eukaryotic cell may contain thousand of different proteins, some abundant and some present in only a few copies. In order to analyze the structure and function of a protein, it must be separated from other proteins and nonprotein molecules and purified in sufficient quantity. Classical protein purification methods include ion-exchange, gel filtration, hydrophobic interaction, and affinity chromatography, as well as various forms of electrophoresis, but these are beyond the scope of this book. Such methods tend to give low yields and <100% purity, and so isolation of a protein from its native source has now been largely supplanted by **recombinant protein** production. A full appreciation of this process requires an understanding of the cloning procedures described in Sections O and P.

Briefly, the gene or complementary DNA (cDNA) encoding the protein is inserted into an **expression vector**, usually a modified plasmid or viral DNA (Section P), which is then introduced into a host cell such as the bacterium *Escherichia coli*, which uses its transcription and translation machinery to synthesize the protein. By engineering the appropriate control sequences into the vector, such as a strong promoter (Section F3), the recombinant protein can be synthesized to form up to 30% of the total protein of the cell. Purification can be simplified by adding a **purification tag** sequence, such as six histidine codons (Section K1), to the 5′- or 3′-end of the cloned gene. In this case, the protein is synthesized with six histidine residues at the N- or C-terminus, which allows one-step purification on an **affinity column** containing immobilized metal ions such as Ni^{2+} or Co^{2+}, which bind the histidines. Untagged proteins pass straight through the column and the pure, recombinant protein can then be eluted from the column with a solution of histidine. Fortunately, the tag usually has little effect on the structure and function of the protein.

Often, a eukaryotic protein will not express well in *E. coli* because of **codon usage bias** (Section K1) and so problematic codons may need to be changed to the favored *E. coli* codons by **site-directed mutagenesis** (Section R5) or a specialized host strain of *E. coli* used that expresses a more eukaryotic pattern of transfer RNAs (tRNAs). 'Foreign' proteins expressed in *E. coli* are frequently insoluble, either because they lack the required post-translational modifications, or because the host chaperones are overloaded or inappropriate (Sections A3 and L4). They may form cytoplasmic aggregates called **inclusion bodies**, from which they can sometimes be resolubilized using urea. Solubility can often be improved by expressing the recombinant as a **fusion protein** with a more soluble partner. This involves placing the coding sequences of the two proteins next to each other in the vector and deleting the termination codon of the upstream partner so that both are translated as a single polypeptide (Section L1). Common fusion partners are **thioredoxin, maltose-binding protein** (**MBP**) and the enzyme **glutathione-S-transferase** (**GST**). GST and MBP can also be used as purification tags using affinity columns containing immoblized **glutathione**, the tripeptide substrate of GST, or maltose for MBP. Other tags that may be included in the protein by encoding them in the vector include an **epitope tag** for antibody-based detection (see below) and a tag encoding the

recognition sequence for a proteolytic enzyme such as Factor Xa, which cleaves proteins specifically after the sequence IleGluGlyArg. If placed between the two partners of a fusion protein, this tag allows them to be separated by proteolysis after purification.

Sometimes expression in *E. coli* just fails, and many eukaryotic proteins require expression in a eukaryotic host–vector system to achieve the correct folding and post-translational modifications (Sections P3 and L4). As well as producing proteins for laboratory analysis, recombinant methods are now routinely used to produce therapeutic proteins such as insulin and blood clotting factors, with over 130 different proteins currently approved for treatment.

Protein sequencing

An essential requirement for understanding how a protein works is a knowledge of its primary structure. The amino acid composition of a protein can be determined by hydrolyzing all the peptide bonds with strong acid and separating the resulting amino acids by chromatography. This indicates how many glycines and serines, etc. there are but does not give the actual sequence. Early methods of sequence determination involved splitting the protein into a number of smaller peptides using specific proteolytic enzymes or chemicals that break only certain peptide bonds. For example, trypsin cleaves only after lysine or arginine, and V8 protease only after glutamic acid. Cyanogen bromide cleaves polypeptides after methionine residues. Each peptide is then subjected to sequential **Edman degradation** in an automated protein sequencer. Phenylisothiocyanate reacts with the N-terminal amino acid, which, after acid treatment, is released as the phenylthiohydantoin (PTH) derivative, leaving a new N-terminus. The PTH-amino acid is identified by chromatography by comparison with standards and the cycle repeated to identify the next amino acid, and so on. The order of the peptides in the original protein can be deduced by sequencing peptides produced by proteases with different specificities and looking for the overlapping sequences. However, this method is both laborious and expensive, and it is now much easier to 'sequence' proteins indirectly by sequencing the DNA of the gene or cDNA (Section R2) and deducing the protein sequence using the genetic code (Sections K1 and S6). However, this misses post-transcriptional (e.g. mRNA editing, Section J4) and post-translational (Section L4) modifications, and so there is still a need for direct protein sequencing. This is now achieved by mass spectrometry (see below). In organisms whose genome has not been fully sequenced, partial protein sequencing can be used to provide information for the construction of an oligonucleotide **probe** (Section Q3), which is then used to find the corresponding gene or cDNA, from which the full protein sequence can then be deduced.

Biophysical methods

Several methods are available to determine the secondary and tertiary structures, and the physical properties of proteins. **Circular dichroism**, a form of UV spectroscopy, determines the relative proportions of different secondary structures and is useful for measuring changes in protein conformation (shape) under different conditions. Because they have such well-defined tertiary structures, many globular proteins have been crystallized and so the tertiary structure can be determined by **X-ray crystallography**. X-rays interact with the electrons in the matter through which they pass. By measuring the pattern of diffraction of a beam of X-rays as it passes through a crystal, the positions of the atoms in the crystal can be calculated. By crystallizing an enzyme in the presence of its substrate, the precise intermolecular interactions responsible for binding and catalysis can be seen. The power and resolution of modern X-ray crystallography

using high intensity **synchrotron radiation** is such that the detailed structures of large macromolecular assemblies like nucleosomes and ribosomes have now been determined.

The structures of small globular proteins in solution can also be determined by two- or three-dimensional **nuclear magnetic resonance** (**NMR**) **spectroscopy**. In NMR, the relaxation of protons is measured after they have been excited by the radiofrequencies in a strong magnetic field. The properties of this relaxation depend on the relative positions of the protons in the molecule. The multi-dimensional approach is required for proteins to spread out and resolve the overlapping data produced by the large number of protons. Substituting ^{13}C and ^{15}N for the normal isotopes ^{12}C and ^{14}N in the protein also greatly improves data resolution by eliminating unwanted resonances. In this way, the detailed structures of proteins up to about 40 kDa in size can be deduced, while partial information on larger proteins can also be obtained. NMR is particularly useful for proteins that do not crystallize readily because they contain unstructured, flexible regions. **Solid state NMR** is used to analyze insoluble membrane proteins. Where both X-ray and NMR methods have been used to determine the structure of a protein, the results usually agree well. This suggests that the measured structures are the true *in vivo* structures.

Cryo-electron microscopy is performed at extremely low temperatures and has lower resolution than X-ray and NMR techniques but can provide structures of much larger entities, e.g. viruses and organelles, whereas **isothermal titration calorimetry** and **surface plasmon resonance** are used to measure the thermodynamic and kinetic parameters of protein–ligand interactions, including protein–protein and protein–DNA.

Mass determination and mass spectrometry

The mass of individual polypeptide chains can be determined by electrophoresis through a polyacrylamide gel in the presence of the ionic detergent sodium dodecyl sulfate (**SDS polyacrylamide gel electrophoresis**, **SDS-PAGE**). SDS binds to, denatures, and imparts a negative charge to polypeptides, so all move towards the anode during electrophoresis at a rate that depends on their mass. Masses are determined by reference to known standards. Denaturation disrupts quaternary structure, so multimeric proteins are split into individual subunits. SDS-PAGE is cheap and easy though not particularly accurate (5–20% error). **Mass spectrometry** (**MS**) offers an extremely accurate method. A **mass spectrometer** consists of an **ion source** that generates characteristic, multiply-charged ionic fragments in the gas phase from the sample molecule, a **mass analyzer** that measures the mass-to-charge ratio (m/z) of the ionized sample, and a **detector** that counts the numbers of ions of each m/z value. For small molecule analysis, samples are traditionally vaporized and ionized by a beam of Xe or Ar atoms. The degree of deflection of the various ions in an electromagnetic field is mass dependent and can be measured, giving a **mass spectrum** (or '**fingerprint**') that identifies the original molecule. However, such methods have an upper mass limit of only a few kDa and are too destructive for protein analysis, so non-destructive ionization techniques are necessary to extend the mass range.

An **electrospray ionization** (**ESI**) ion source creates positively charged (protonated) ions of individual protein molecules by creating then vaporizing a fine spray of highly charged droplets of the protein solution, whereas a **matrix-assisted laser desorption/ionization** (**MALDI**) ion source generates gas-phase ions by the laser vaporization of the sample contained in a solid bed of one of several chemicals (the 'matrix'). These ion sources can be coupled to a variety of different mass analyzers and detectors, each suited to a different purpose. An ESI source is commonly attached to a **quadrupole** analyzer. This uses a set of four parallel rods to produce a time-varying electric field that filters ions of different

m/z values, allowing them to arrive and be counted individually at the detector. A MALDI source is commonly coupled to a **time-of-flight** (**TOF**) analyzer. The ions are accelerated along a flight tube and separate according to their velocities. The detector then counts the different ions as they arrive. A system set up in this way is called **MALDI-TOF** and is most commonly used for protein identification by **peptide mass fingerprinting** of individual peptides generated by prior proteolysis of the protein of interest (Section S3).

For mass determination, the highest quality data are obtained from an **ESI-Q-TOF** mass spectrometer with combined quadrupole and TOF mass analyzers. Because proteins have multiple sites to carry a proton (all lysine, arginine, and histidine residues), they acquire multiple positive charges, and in a slightly unpredictable fashion. Thus a protein with 20 basic residues might be charged with between eight and 18 protons. Because mass spectrometers analyze ions according to the *m/z* ratio, each differently protonated variant creates a different signal (the mass, *m*, stays almost the same but the charge, *z*, increases in integral values of +1, +2, etc.). From the profile of *m/z* values, it is possible to calculate the molecular weight of the protein. The precision of this method is around 0.01%, so the measured mass of a 50-kDa protein would be accurate to about ±5 Da. This method has also been used to study protein complexes in the MDa range.

Applications for antibodies

Antibodies are useful molecular tools for investigating protein structure and function. Antibodies are themselves glycoproteins and are generated by the immune system of higher animals when they are injected with a macromolecule (the **antigen**) such as a protein that is not native to the animal. Their physiological function is to bind to antigens on the surface of invading viruses and bacteria as part of the animal's response to kill and eliminate the infectious agent. Antibodies fall into various classes, but all have the same Y-shaped structure comprising two **heavy chains** and two **light chains**, linked by disulfide bonds (Figure 1). The most useful are the immunoglobulin G (IgG) class, produced as soluble proteins by B lymphocytes. IgGs have a very high affinity and specificity for the corresponding antigen and can be used to detect and quantitate the antigen in cells and cell extracts. The specificity lies in the variable region of the molecule, which is generated by recombination (Section E4). This region recognizes and binds to a short sequence of

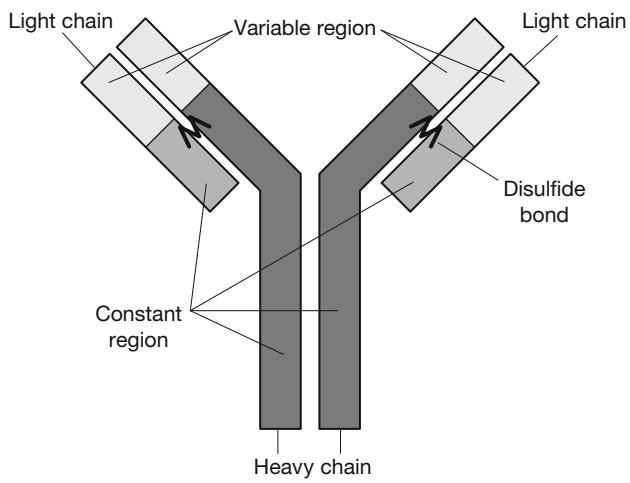

Figure 1. Structure of an antibody molecule.

5-8 amino acids on the surface of the antigen (an **epitope**). Usually, a single antigen elicits the production of several different antibodies by different B cell clones, each of which recognizes a different epitope on the antigen (a **polyclonal** antibody mixture). However, it is possible to isolate the B cell clones (usually from mice) and grow them in culture after fusion to cancer cells; the resulting **hybridoma** cell lines secrete **monoclonal** antibodies, which bind individual epitopes and so tend to be more specific than polyclonals. Antibodies can be used to detect specific proteins in cells by **immunocytochemical** techniques, particularly **immunofluorescence** (Section S4). They can also be used to detect proteins in cell extracts after separation by SDS-PAGE, which separates polypeptides according to their molecular mass (Figure 2). (Note that the detergent SDS splits multimeric proteins into their individual polypeptide subunits.) After separation, the polypeptides are transferred from the gel to a membrane in a procedure similar to Southern blotting (Section R1). This so-called **western blot** is incubated with an antibody that is specific for and binds to the polypeptide of interest (the **primary** antibody), followed by a **second antibody,** several molecules of which recognize and bind the first antibody (the use of a labeled second antibody has several practical advantages, including signal amplification). As the second antibody has attached to it an enzyme or chemical that can generate a color or a light signal, the position of the polypeptide on the blot can be visualized using a suitable detector. This can show, for example, if a protein is present or absent, or increases or decreases in a cell under particular conditions, e.g. hormone stimulation. If it moves position, this could indicate a modification, perhaps partial degradation.

In theory, detection of different proteins requires a different antibody in each case, However, recombinant proteins can be engineered and expressed with an **epitope tag** at one end. This is a short sequence of extra amino acids (an epitope) that is recognized by just one antibody, thus allowing this same antibody to be used to detect any protein

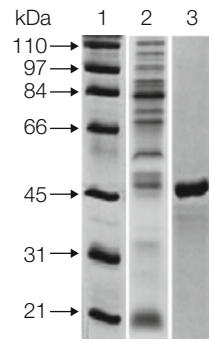

Figure 2. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and western blot. A set of molecular weight standards (lane 1) and a sample of rat liver extract (lane 2) were boiled in buffer containing SDS, which denatures the proteins and imparts a mass-dependent negative charge to them. The proteins were then separated according to their mass by electrophoresis in a polyacrylamide gel and visualized by staining the gel with the dye Coomassie Blue. The masses of the standards are show in kiloDaltons (kDa). An unstained liver sample identical to lane 2 was blotted on to a nitrocellulose membrane and incubated with a rabbit IgG antibody specific for the enzyme nucleoside triphosphatase (NTPase). This was then incubated with a goat anti-rabbit immunoglobulin G (IgG), which had previously been covalently linked to the enzyme peroxidase. Finally, the blot was treated with a peroxidase substrate that generates a visible color. The presence of the 46-kDa NTPase in the liver extract can clearly be seen (lane 3).