12

Protein structure and a molecular approach to medicine

Introduction

The introduction to this book made clear that one major reason for studying protein structure and function is to obtain insight into diseases afflicting modern civilizations. These diseases range in diversity from cancer to cholera, from human immunodeficiency virus (HIV) to hepatitis, from malaria to myocardial infarction to name a few. One potential impact of improved structural knowledge is the ability to treat many diseases with highly specific drugs. For bacterial and viral infections immunization has been refined to allow the successful introduction of live, but attenuated, biological samples or components of the active protein. Inoculation elicits an immune reaction that confers lifelong or substantial immunity and has played a major role in improved health care. In this manner diseases such as smallpox have been eradicated from all parts of the world whilst polio and measles have been drastically limited in distribution.

For diseases such as cholera treatments have been devised that limit routes of infection. Cholera spreads via a faecal–oral infection route and the presence of the bacterium *Vibrio cholerae*. By establishing clean water supplies infection routes via contaminated water or food is limited. Infection leads to colonization of the intestines and the production of toxin that results in chronic diarrhea which will, if left untreated, result in death as a result of fluid loss leading to shock and acidosis. Victims die of dehydration unless water and salts are replaced.

Cholera represents one example of how disease origin is traced back to the action of one protein in this case a toxin. By studying the structure and function of the cholera toxin we gain insight into the molecular events that underpin disease. The toxin is a hexameric protein of 87 kDa (subunit composition of AB_5) that binds to a ganglioside known as GM1 via the B subunits and transfers the A subunit across membranes via receptor mediated endocytosis. Inside the cell reductive cleavage of a disulfide bond releases a 195 residue fragment from the A subunit that is responsible for pathogenesis.

The fragment catalyses the transfer of ADP-ribose from NAD^+ to the side chain of Arg187 in the G_α subunit of the heterotrimeric G_s protein; a process known as ADP-ribosylation (Figure 12.1). GTP binding leads to dissociation into G_α and $G_{\beta\gamma}$ each of which activates further cellular components. G protein activation is usually transitory because G_α is a GTPase that hydrolyses GTP to GDP and promotes G protein re-association. G_α-GTP activates adenylate cyclase with the cAMP acting as a second messenger and regulating the activity of cAMP dependent

Proteins: Structure and Function by David Whitford
© 2005 John Wiley & Sons, Ltd

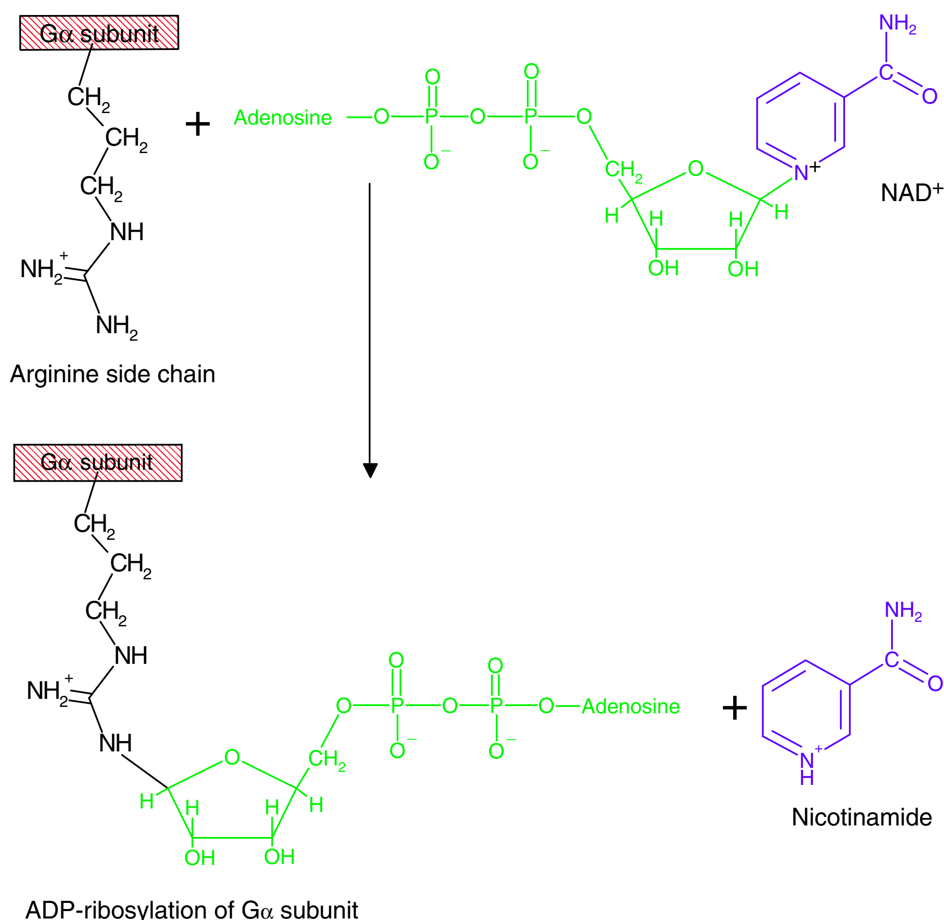

Figure 12.1 Cholera toxin promotes ADP-ribosylation of the G_α subunit of heterotrimeric G proteins

protein kinase. ADP ribosylation of the G_α subunit has a critical effect; it continues to activate adenylate cyclase whilst inhibiting the GTPase activity of G_α. Adenylate cyclase remains in a permanently activated state along with G_α and cellular cAMP levels increase dramatically. Intestinal cells respond to increased cAMP levels by activating sodium pumps and the secretion of Na^+. To compensate chloride, bicarbonate and water are also secreted but the net effect is the loss of enormous quantities of fluid and dehydration.

A molecular approach to the study of cholera elucidated the basis of toxicity and showed that dehydration is counteracted by replenishment of fluids containing salts. Significantly, other bacterial toxins such as *E. coli* heat stable enterotoxins and the pertussis toxin from the bacterium *Bordetella pertussis* act via similar mechanisms. Pertussis is responsible for whooping cough in young babies and is effectively countered in economically advantaged countries by vaccination of young babies around 6 months of age using an inactivated fragment of the whole toxin.

The fundamental events contributing to two diseases have been elucidated through their common mode of action and effective therapeutic strategies have been devised – rehydration and vaccination. However, this approach is not always possible. Increasingly the techniques of molecular and cell biology

are providing insights into the underlying causes of disease giving rise to the new field of molecular medicine. Molecular medicine became a possibility with the completion of genome sequencing projects and the identification of many inherited diseases arising from mutated genes. Gene screening is now established as an important diagnostic tool with DNA extracted from biological samples and 'searched' for defects. Defective genes can be as small as a simple base change or may be larger involving missing, duplicated or translocated segments of genes.

The availability of 'predictive' gene tests has increased dramatically in the last decade. Currently genetic tests are available for the most common diseases such as cystic fibrosis, neurodegenerative disorders such amylotrophic lateral sclerosis (Lou Gehrig's disease), Tay–Sachs disease, Huntington's disease, familial hypercholesterolaemia (a disease causing catastrophically high cholesterol levels and myocardial infarction), childhood eye cancer (retinoblastoma), Wilms' tumour (a kidney cancer), Li–Fraumeni syndrome, familial adenomatous polyposis (an inherited predisposition to form pre-cancerous polyps and associated with colon cancer), and *BRCA*1 (a gene implicated in familial breast cancer susceptibility). The list could be extended for many more pages. Over 3000 mutated genes are known to be involved in inherited disorders or human disease.

Alongside molecular medicine there have been major advances in biotechnology to produce protein-based drugs for the treatment of disease. Today human insulin is given as fast and slow acting forms in the treatment of Type I diabetes arising from damage to the β cells of the islets of Langerhans found in the pancreas. Human insulin is produced by recombinant DNA methods and avoids many problems faced previously where insulin was extracted from the pancreas of pigs or cows.

Molecular medicine will occupy a pivotal role in clinical practice throughout the 21st century. This chapter deals with a selective group of proteins implicated in disease states where there is a growing understanding of the biological problem. As such it represents a series of case studies, by no means complete, described to provide a flavour of modern approaches to molecular medicine.

Sickle cell anaemia

The discovery of Ingram in 1956 that sickle cell anaemia arose from the substitution of glutamate by valine at position 6 of the β subunit of haemoglobin was the first example of a molecular defect contributing to human disease. It represents a key event, perhaps even the origin, of molecular medicine. Mutations in haemoglobin were identified because the protein is obtained from individuals in sufficient quantities for peptide sequencing; this situation rarely extends to other diseases where samples are normally only obtained via invasive biopsies or post-mortem.

Substitution of valine by glutamate alters the surface properties of the β subunit leading to aggregates of protein in the deoxy form of haemoglobin *S* (HbS). The hydrophobic side chain of Val6 projects out into solution fitting precisely into a small hydrophobic pocket on the surface of a neighbouring β subunit formed by Phe85 and Leu88. The pocket favours aggregation whilst oxygenation removes this pocket by conformational change. In normal haemoglobin the charged side chain of glutamate is not accommodated within a non-polar surface region. Suddenly the molecular basis of sickle cell anaemia is clear.

Long polymers or fibres of deoxy HbS ∼20 nm in diameter extend the length of the erythrocyte deforming a normal biconcave cell into a sickle shape where distortion leads to cell rupture (Figure 12.2). The greatest danger of aggregation occurs in the capillaries where oxygen is 'off loaded' to tissues and the concentration of deoxy-HbS is higher. Precipitation prevents blood flow through the capillaries causing tissue necrosis and life threatening complications.

Sickle cell anaemia occurs when an individual carries two mutated forms of the β subunit gene. A single copy leads to the sickle cell trait and is generally asymptomatic. Generally, 'harmful' mutations are present at very low frequency and it is unusual for a mutant gene to be present at high frequency unless it confers a specific evolutionary advantage. In certain regions of West and Central Africa approximately 25 percent of the population exhibit heterozygosity, or the sickle cell trait. This is unexpected since harmful mutations are normally selectively weeded out by

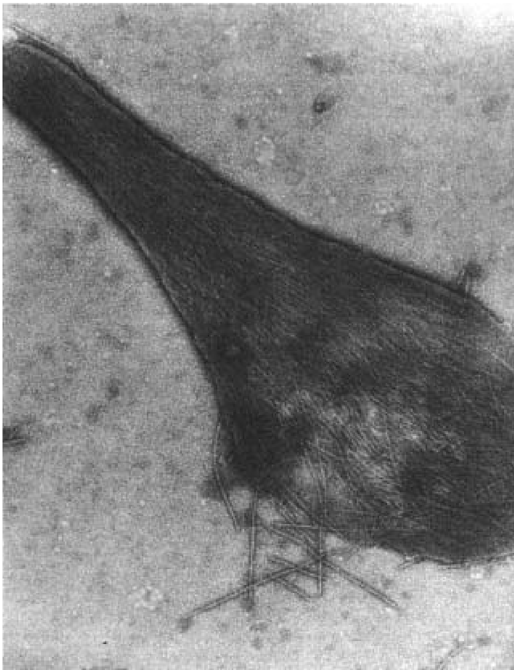

Figure 12.2 Electron micrograph of deoxy-HbS fibres spilling out from a ruptured erythrocyte

evolution. It is clear that the sickle cell trait correlates closely with the prevalence of malaria. Malaria kills over 1 million individuals each year mostly young children – a staggering and regrettable statistic – and is caused by the mosquito-borne protozoan parasite *Plasmodium falciparum*. The parasite resides in the erythrocyte during its life cycle and causes adhesion of red blood cells to capillary walls. Individuals with the sickle cell trait possess red blood cells containing abnormal haemoglobin and these tend to sickle when infected by the parasite. The result is that infected cells flow through the spleen and are rapidly removed from circulation. In other words individuals with sickle cell trait deal with the parasite more effectively than normal individuals – the consequence is that the mutant gene persists in malaria-endemic regions.

The gene encoding the β subunit is located on chromosome 11 – a region containing many other globin genes and known as the β globin cluster. Polymorphisms in this region are equated with disease severity and several distinct haplotypes are identified in sickle cell anaemia and given names reflecting their origins within Africa, such as Senegal haplotype, the most benign form of sickle cell disease, the Benin haplotype and the Central African Republic haplotype−one of the severest forms of sickle cell anaemia.

Screening of individuals is routinely performed using electrophoresis, where the sickle cell haemoglobin shows altered migration properties when compared with the normal protein. Tests are cheap, quick and reliable and a positive result points towards further diagnostic investigation. Tests provide prospective parents with realistic information on the inheritance of the trait. When both parents carry the sickle cell trait there is a 25 percent chance of the offspring having sickle cell anemia, a 50 percent chance of the sickle cell trait and a 25 percent chance of a normal individual. Despite good screening procedures the treatment of sickle cell anaemia is problematic and highlights a major dilemma of molecular medicine. Diseases can be identified but frequently there is no simple 'cure'. Before the introduction of antibiotics individuals with sickle cell anemia frequently died in childhood. Modern care involves treatment with penicillins and significantly reduces infant mortality.

Viruses and their impact on health as seen through structure and function

Viral infections are well known to all of us since they are responsible for the common cold as well as many more serious diseases. The term virus is derived from the Latin word for poison and seems entirely appropriate in view of the devastating illnesses caused in humans from viral infections. The diseases attributed to viruses are vast and include influenza (flu), measles, mumps, smallpox, yellow fever, rabies, poliomyelitis, as well as acquired immune deficiency syndrome (AIDS), haemorrhagic fever and certain forms of cancers.

The existence of sub-microscopic infectious agents was suspected by the end of the 19th century when the Russian botanist Dimitri Ivanovsky showed that sap from tobacco plants infected with mosaic disease contained an agent capable of infecting other tobacco plants even after passing through filters known to retain bacteria. Plant and animal cells are not the only

living cells subject to viral attack. In 1915 Frederick Twort isolated filterable entities that destroyed bacterial cultures and produced cleared areas on bacterial lawns. The entities were bacteriophages, often abbreviated to phage, that attach to bacterial cell walls introducing genetic material. The result is rapid replication of new viral progeny with cell lysis releasing new infectious particles.

Crystallization of tobacco mosaic virus (TMV) by W.M. Stanley in 1934 represented one of the first systems to be crystallized but also emphasized the high inherent order present in viruses. Subsequently TMV has been a popular experimental subject. One important conclusion from studying TMV was that viruses were constructed using identical subunits packed efficiently around helical, cubic or icosahedral patterns of symmetrical organization. To clarify the terminology for virus components a number of terms were introduced. The capsid denotes the protein shell that encloses the nucleic acid and is based on a subunit structure. The subunit is normally the smallest functional unit making up the capsid. The capsid together with its enclosed nucleic acid (either DNA or RNA) constitutes the nucleocapsid. In some viruses the nucleocapsid is coated with a lipid envelope that can be derived from host cells. All of these structures form part of the virion – the complete infective virus particle.

Viruses lack normal cellular structure and do not perform typical cell-based functions such as respiration, growth and division. Instead they are viewed as ‘parasites’ that at one stage of a life cycle are free and infectious but once entered into a living cell are able to sequester the host cell’s machinery for replication. The conversion of viral genetic material into new viruses or the incorporation of the viral DNA into the host genome can occur via a variety of different routes and depends on the organization of the viral genome.

The Baltimore classification (Table 12.1) is based on nuclei acid type and organization and uses the relationship between viral mRNA and the nucleic acid found in the virus particle. In this system viral mRNA is designated as the plus (+) strand and a RNA or DNA sequence complementary to this strand is designated as the minus (−) strand. Production of mRNA requires the use of either DNA or RNA (−) strands as templates and viruses can exist as double-stranded DNA, single-stranded + or − DNA, double-stranded RNA, and either (−) or (+) single strands of RNA. Almost all viruses can be classified in this scheme. RNA-based viruses form DNA using the enzyme reverse transcriptase and represent retroviruses. Retroviruses are involved in many diseases, transforming cells into cancerous states by the conversion of genetic material (RNA) into DNA followed by integration into the host cell genome (proviral state) and the production of transcripts and translation products via host cell machinery.

The two most important retroviruses in terms of their effect on man are the human influenza virus (flu) and HIV. These viruses are responsible for the two greatest pandemics affecting humans over the last 100 years.

Viruses, like bacterial infective agents, act as antigens in the body and elicit antibody formation in infected individuals. As a result vaccines can be developed against viruses most frequently based on a component of the virus coat or an inactivated form of the virus. Although diseases such as smallpox have been effectively eliminated through the use of vaccines this is not always an effective treatment route. For influenza and HIV this is particularly difficult and lies at the heart of the pandemics caused by these agents. These retroviruses are of considerable relevance to health care in the modern era and are described as a result of their importance in two major diseases affecting the world today.

HIV and AIDS

HIV has been the subject of intense research since 1983 when Luc Montagnier, Robert Gallo and colleagues identified it as the causative agent contributing to a complex series of illnesses grouped under the term ‘acquired immune deficiency syndrome’ (AIDS). The disease was originally recognized in the United States at the beginning of the 1980s as one affecting primarily homosexual men and led to a systematic decrease in the ability of the body to fight infection arising from a progressive loss of lymphocytes. The decline in the immune response led to the name AIDS. Lymphocyte loss rendered the body defenseless against infection, with individuals succumbing to rare

Table 12.1 A scheme of classifying viruses according to their mode of replication and genomic material sometimes called Baltimore classification

Genome organization	Examples	Details of replication
ds DNA	Adenoviruses Herpesviruses Poxviruses	Adenoviruses replicate in the nucleus using cellular proteins. Poxviruses replicate in the cytoplasm making their own enzymes for replication
ss (+) sense DNA	Parvoviruses	Replication occurs in the nucleus, involving the formation of a (−) sense strand, which serves as a template for (+) strand RNA and DNA synthesis
ds RNA	Reoviruses Birnaviruses	Viruses have segmented genomes. Each genome segment is transcribed separately to produce monocistronic mRNAs
ss (+) sense RNA	Picornaviruses Togaviruses	(a) Polycistronic mRNA, e.g. picornaviruses means naked RNA is infectious. Translation results in the formation of a polyprotein product, which is cleaved to make mature proteins (b) Complex transcription, e.g. togaviruses. Two or more rounds of translation are necessary to produce the genomic RNA
ss (−) sense RNA	Orthomyxoviruses Rhabdoviruses	Must have a virion particle RNA directed RNA polymerase (a) Segmented, e.g. orthomyxoviruses. First step in replication is transcription of the (−) sense RNA genome by the virion RNA-dependent RNA polymerase to produce monocistronic mRNAs, which also serve as the template for genome replication (b) Non-segmented, e.g. rhabdoviruses. Replication occurs as above and monocistronic mRNAs are produced
ss (+) sense RNA with DNA intermediate in life-cycle	Retroviruses	Genome is (+) sense but does not act as mRNA. Instead acts as template for reverse transcription
ds DNA with RNA intermediate	Hepadnaviruses	Viruses also rely on reverse transcription. Unlike retroviruses this occurs inside the mature virus particle. On infection the first event to occur is repair of gapped genome, followed by transcription

ds = double stranded, ss = single stranded.

opportunistic infections or diseases such as pneumonia and tuberculosis (TB). In some cases rare forms of skin cancer (Kaposi's sarcoma) were associated with advanced HIV infection along with other cancers.

In the immune system a defining event in HIV-mediated disease is a fall in the number of circulating helper T cells and this feature is used as a marker for immunodeficiency. Once the absolute T-cell count falls below a threshold of 200 cells per mm^3 in the peripheral blood individuals are vulnerable to AIDS-defining opportunistic infections. The virus preferentially infects helper T cells, as well as other cells such as macrophages and dendritic cells, because of the presence of CD4, a cell surface protein that acts as a high affinity receptor for the virus. The interaction between CD4 and virus mediates entry of the virus into the host cell and is a critical step in the overall process of HIV pathogenesis.

Subsequently, the disease was shown to exist in other populations; it was not restricted to homosexuals nor was it confined to the United States (Figure 12.3). HIV infection is responsible for significant mortality in *all* sexually active populations and is present throughout the world. In many countries the absence of good health care systems has meant that AIDS has become the fourth leading cause of mortality throughout the world and is responsible for major decreases in life expectancy. Infection with HIV also occurs from intravenous drug use and is also observed as a result of blood transfusions using infected (unscreened) donated samples.

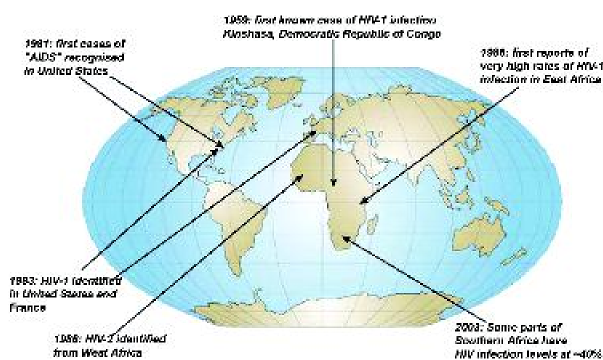

Figure 12.3 The sites of key events in the development of the HIV epidemic. This world map highlights the first known cases of infection with HIV-1 and HIV-2 in West and sub-Saharan Africa

The spectre of progressive HIV infection spread predominantly through sexual intercourse has dominated health care issues in all areas of the world. In 2003 it was estimated by the World Health Organization that over 40 million individuals were affected by HIV infection – a sombre and immensely problematic issue. There is little doubt that HIV infection and its sequelae will be responsible for massive increases in mortality rates, especially in sexually active populations and will also affect the young as a result of fetal or neonatal transmission. The pandemic nature of the disease means that dealing with HIV-infected individuals is a major health care issue in all areas of the world.

The HIV genome

HIV is a member of the lentivirus genus that includes retroviruses with complex genomes. An early objective in studying HIV was to sequence the genome to understand the size, function and properties of the virally encoded proteins. The HIV genome is small (~9.4 Kb) and encodes three structural proteins (MA, CA, and NC), two envelope proteins (gp41, gp120), three enzymes (RT, protease and integrase), and six accessory proteins (Tat, Rev, Vif, Vpr, Nef & Vpu) (Figures 12.4 and 12.5).

These viruses exhibit cone-shaped capsids covered with a lipid envelope derived from the membrane of the host cell. Puncturing the exposed surface are glycoproteins called gp120 – a trimeric complex of proteins anchored to the virus via interactions with a transmembrane protein called gp41. Both gp120 and gp41 are derived from the *env* gene product. Inside this layer is a shell, often called the matrix, composed of the matrix protein (MA). The matrix protein lines the inner surface of the viral membrane and a conical capsid core forms from the major capsid protein (CA). This complex is at the centre of the virus and surrounds two copies of the unspliced viral genome that are stabilized by interaction with the nucleocapsid protein (NC). One major region of the HIV genome is the *pol* gene (Figure 12.5); its product is a polyprotein of three essential enzymes. These enzymes are protease, reverse transcriptase and integrase.

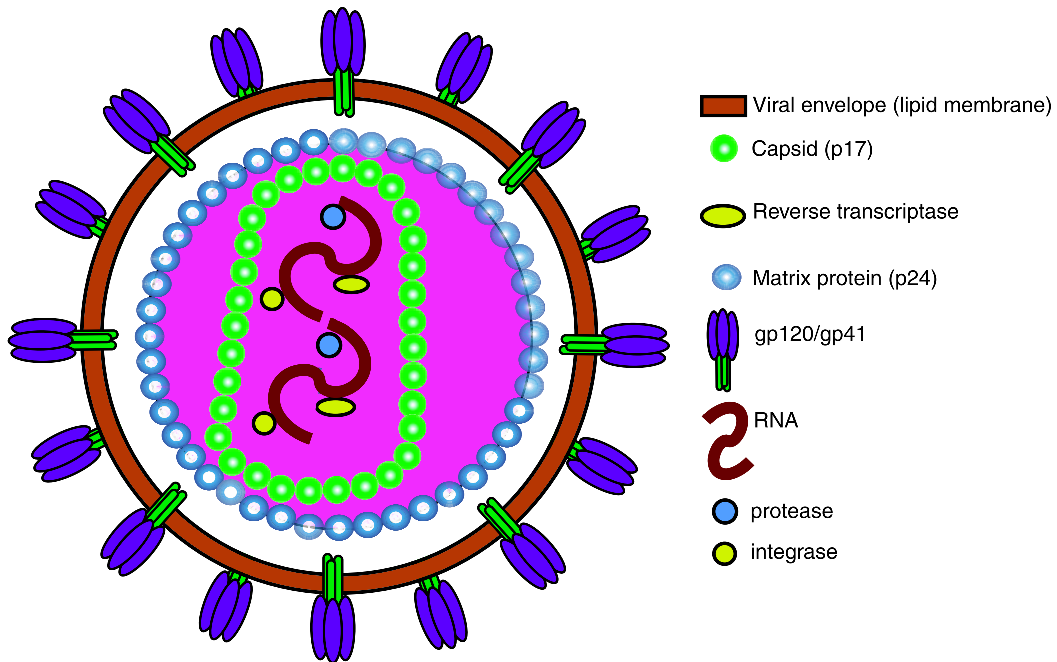

Figure 12.4 A picture of the organization of protein and RNA within the intact human immunodeficiency virus

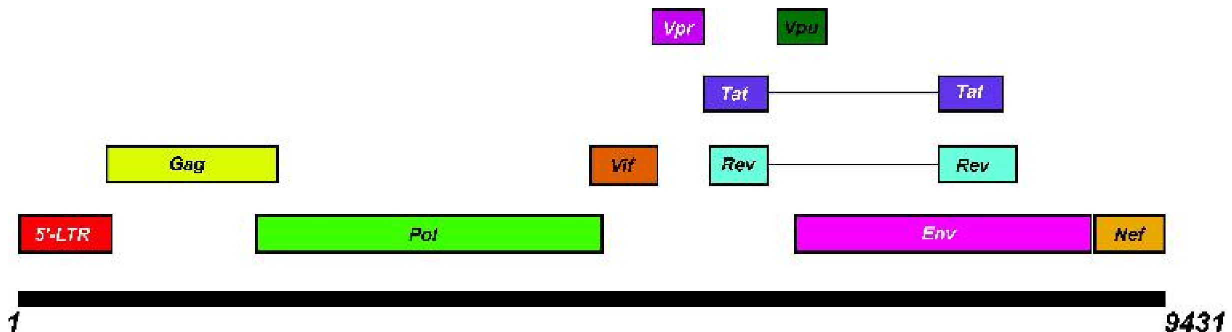

Figure 12.5 Organization of the HIV genome and the proteins derived from the genome

Of the accessory proteins three function within the host cell by facilitating transcription (Tat), assisting in transport of viral DNA to the cell nucleus (Rev), and acting as a positive factor for viral expression (Vif). These proteins are not found within the virion. The other three proteins are packaged with the virus and play important roles in regulating viral promoters (Vpr), acting as a negative regulatory factor for viral expression (Nef) and promoting viral release and CD4 degradation (Vpu) (Table 12.2).

The first phase of infection results in virus binding to $CD4^+$ lymphocytes. The helper T cells are the primary

Table 12.2 Functional role of HIV encoded gene products

Protein	Gene	Role
MA (p17)	*Gag*	Matrix protein
CA (p24)	*Gag*	Matrix protein
NC (p7)	*Gag*	Nucleocapsid protein
gp120	*Env*	Extracellular glycoprotein binds to CD4 receptor
gp41	*Env*	Transmembrane protein
Reverse transcriptase (RT)	*Pol*	Conversion of RNA > DNA
Integrase	*Pol*	Insertion of DNA into host genome.
Protease	*Pol*	Processing of translated polyprotein
Rev	*Rev*	Regulate expression of HIV proteins
Nef	*Nef*	Negative regulatory factor
Vpr	*Vpr*	Viral (HIV) promoter regulator
Vpu	*Vpu*	Viral release and CD4 degradation
Vif	*Vif*	Viral infectivity factor
Tat	*Tat*	Transcriptional activator

The three viral encoded enzymes play a major role in the HIV replication cycle and the formation of new virions and have formed the basis of most drug-based treatments of HIV infection. In part this has arisen from the highly resolved structural models available for reverse transcriptase, the integrase and protease of HIV.

target and binding leads to a series of intracellular events that causes their destruction with a half-life of less than two days. During initial stages of infection the viral genome becomes integrated into the chromosome of the host cell and is followed by a later phase of regulated expression of the proviral genome leading to new virion production. The stages of viral infection require the concerted involvement of viral proteins; this includes, for example, gp120-mediated binding to $CD4^+$ cells, conversion of viral RNA into DNA by reverse transcriptase, integration of viral DNA into the host chromosome by integrase, nuclear localization of viral DNA by Vpr, and processing of viral mRNAs by Tat, Rev and Nef.

The structure and function of Rev, Nef and Tat proteins

Nef is a single polypeptide chain of 206 residues originally recognized as a 'negative factor' where its absence caused low viral loads in cells. After viral infection the protein is expressed in high concentrations and inoculation of Rhesus monkeys with a Nef-deletion strain in the closely related SIV (simian immunodeficiency virus) did not lead to disease. Nef has at least two distinct roles during infection; it enhances viral replication, and by an unknown mechanism it reduces the number of CD4 receptors found on the surface of infected lymphocytes. The protein is unusual in having a myristoylated N-terminal residue indicating membrane association; removal of this modification inhibits activity. The structure of the core domain of Nef (residues 71–205) has been determined by NMR and crystallographic methods, with both techniques suggesting a fold containing a helix-turn-helix (HTH) motif with three α helices, a five-stranded antiparallel β sheet, an unusual left-handed poly-L-proline type-II helix and a 3_{10} helix.

The role of Rev is to regulate expression of HIV proteins by controlling the export rate of mRNAs from the nucleus to the cytoplasm where translation occurs. The Rev protein contains 116 residues and within this sequence are nuclear localization and export sequences

(NLS and NES) consistent with its role in mRNA export, but a structure has not been determined.

The final non-structural protein synthesized by the virus and packaged within the core is Tat–a transcriptional activator that enhances transcription of integrated proviral DNA. Transcription starts at the HIV promoter–a region of DNA located within the 5′ long terminal repeat region with the promoter binding RNA polymerase II and other transcription factors. Transcription frequently terminates prematurely due to abortive elongation and the Tat protein overcomes this problem by binding to stem–loop sites on the nascent RNA transcript. Description of the Tat protein is complicated by translation via two exons leading to a protein containing between 86 and 101 residues; the first exon encodes residues 1–72 and the second exon translates as residues 73–101. Structural studies have utilized an 86-residue fragment despite most HIV strains containing full-length protein, but have yet to yield a detailed picture for Tat. The protein is divided into four distinct regions; an N-terminal cysteine-rich domain, a core region, a glutamine-rich segment and a basic zone. The basic region contains the sequence $R_{49}KKRRQRRR_{57}$ essential for recognition and binding to RNA.

The role of Vif, Vpr and Vpu

A similar lack of structural information pertains to Vif, Vpr and Vpu. These three proteins range in size from 15 to 23 kDa and structures are unavailable, although considerable functional data has been acquired. It is known that Vpr, a 96-residue accessory protein present early in the viral life cycle, plays an important regulatory role during replication. Vpr localizes to the nucleus and is implicated in cell cycle arrest at the G_2/M interface. Vpr also enhances transcription from the HIV-1 promoter structure and cellular promoters.

Vif stands for viral infectivity factor and is a basic protein (23 kDa) whose precise function is unknown. Naturally occurring strains of HIV strains bearing mutations in Vif are known to replicate at significantly lower levels when compared with the wild-type virus. In some instances this is associated with low viral loads, high CD4 T-cell counts and slower progression to advanced forms of the disease. These results emphasize the importance of Vif and homologous proteins have been recognized in other lentiviruses from a highly conserved – SLQSLA motif located between residues 144–149. Vif is required during later stages of virus formation since virions produced in the absence of Vif cannot complete proviral DNA integration into host chromosomes. Unlike Vpr, Vif is primarily found in the cytoplasm as a membrane-associated protein.

Vpu is a small integral membrane protein that regulates virus release from the cell surface and degradation of CD4 in the endoplasmic reticulum. These two distinct biological activities represent independent mechanisms attributed to the presence of distinct structural domains within Vpu. A transmembrane domain is associated with virus release from the cell whereas CD4 degradation involves interaction with the cytoplasmic domain.

In contrast to the above proteins considerable understanding of HIV has come from structural studies of the remaining components of the virion particularly the enzymes and envelope proteins. Molecular biology has allowed individual proteins, or more frequently domains derived from whole proteins, to be cloned, expressed and subjected to detailed structural analysis. These studies assist in providing a picture of their roles during the viral life cycle and understanding the interdependence of structure and function in these proteins represents a key effort in the attempts to combat the disease. Alongside a large body of data for the reverse transcriptase and protease there also exists structural information for gp120, the integrase, domains of the transmembrane protein gp41 (the TM domain) and the matrix, capsid and nucleocapsid proteins.

Structural proteins of HIV

Structural proteins are derived from the *Gag* gene product by proteolytic cleavage to yield the MA, CA and NC proteins (Figure 12.6). With a combined mass of ~55 kDa the Gag polyprotein yields a matrix (MA) protein of 17 kDa, the capsid protein (CA) of 24 kDa and the nucleocapsid protein (NC) of ~7 kDa. The structure of the MA protein consists of five α helices, two short 3_{10} elements and a three-stranded β sheet. Although monomeric in solution several lines of evidence indicate trimeric structures represent ‘building blocks’ for the matrix shell within a mature virion.

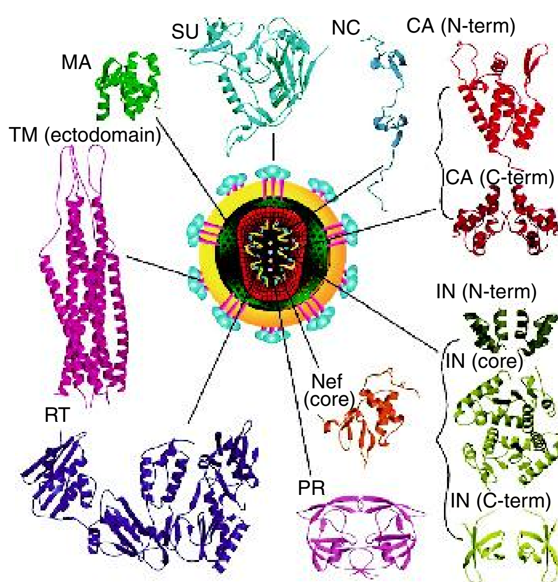

Figure 12.6 The structure of HIV. Drawing of the mature HIV virion surrounded by ribbon representations of the structurally characterized viral proteins and fragments. The protein structures have been drawn to approximately the same scale whilst the TM ectodomain shown is that determined for the closely related SIV (reproduced with permission from Turner, B.G. & Summers, M.F. *J. Mol. Biol.* 1999, **285**, 1–32. Academic Press)

The capsid protein occurs at ∼2000 copies per mature virion forming a cone-shaped, dense, structure at the virus centre that encapsulates RNA as well as enzymes, NC protein and accessory proteins. Two copies of unspliced RNA are packaged in a ribonucleoprotein complex formed with the NC protein. The CA protein has proved difficult to isolate and no structure is available for the intact protein. In an N-terminal domain seven helices, two β hairpins and an exposed partially ordered loop are observed with the domain forming an arrowhead shape. This domain associates with similar regions as part of the condensation of the capsid core.

The nucleocapsid or NC protein is involved in recognition and packaging of the viral genome. Retroviral NC proteins contain the Cys-X_2-Cys-X_4-His-X_4-Cys array characteristic of Zn finger proteins, and in HIV the NC protein contains two such elements. The structure of this small protein is defined by these two Zn finger elements where reverse turns form a metal binding centre that leads via a loop to a C terminal 3_{10} helix. In the NC protein the two metal-binding centres may interact although this is thought to be a flexible, dynamic interaction and both play a role in binding RNA via conserved hydrophobic pockets.

Reverse transcriptase and viral protease

Two of the three enzymes, reverse transcriptase and protease, are amenable targets for drug therapy largely as a result of detailed structural analysis. Reverse transcriptase is responsible for the conversion of RNA to DNA whilst the protease catalyses processing of the initial polyprotein (Gag-Pol) product into separate functional proteins.

HIV protease The protease was the first HIV-1 protein to be structurally characterized and is translated initially as a Gag-Pol fusion protein; its role involves proteolytic cleavage of the polyprotein to release active components. In an autocatalytic reaction analogous to that seen in zymogen processing the protease liberates subunits found in the mature virus. The Gag precursor forms the structural proteins whilst Pol produces enzymes – reverse transcriptase, integrase, and protease.

As a protease the HIV enzyme is unique is catalysing bond scission between a tyrosine/phenylalanine and a proline residue with no human enzyme showing a similar specificity. Despite unique specificity the HIV protease is structurally similar to proteases of the pepsin or aspartic protease family and this has assisted characterization. Historically, pepsin was the second enzyme to be crystallized (after urease) and structures are available for pepsins from different sources. All reveal bilobal structures with two aspartate residues located close together in the active site. These residues activate a bound water molecule for use in proteolytic cleavage of substrate.

The bilobal structure of pepsin is based around a single polypeptide chain of ∼330 residues where one lobe contains the first 170 residues and the remaining residues are located in a second lobe. Each lobe is related by an approximate two-fold axis of symmetry

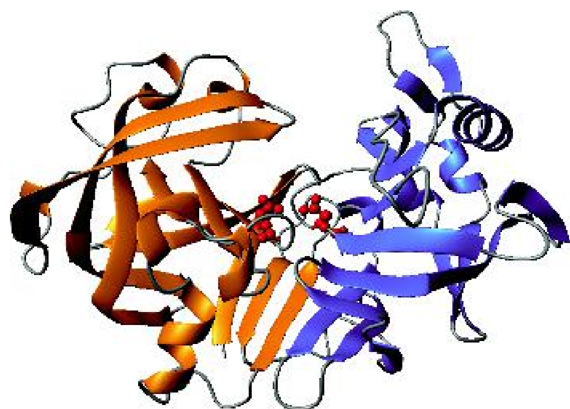

Figure 12.7 The overall topology of pepsin showing strand-rich secondary structure arranged within two lobes. Secondary structure elements for residues 1–170 of the first lobe are shown in orange with the remaining secondary structure shown for the second lobe in blue. Asp32 and 215 of the active site are shown in red

and at the centre, between each lobe, is the active site (Figure 12.7). It seems probable that this topology arose by gene duplication because each lobe contributes a single Asp side chain to the active site found within an Asp-Thr/Ser-Gly- motif preceded by two hydrophobic side chains. In pepsin the catalytic Asp residues are at positions 32 and 215. Each side chain interacts via a complex network of hydrogen bonds (Figure 12.8) with at least one molecule of water. The

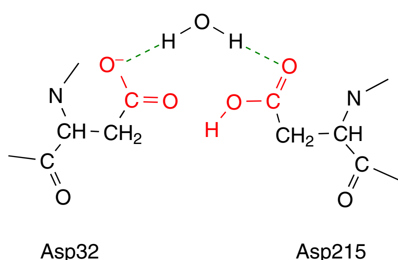

Figure 12.8 The interaction between Asp32 and 215 in pepsin involves a bound water molecule and two protonation states for the Asp side chain. A water molecule bridges between two oxygen atoms, one on each of the carboxyl groups of the side chain

side chains exhibit very different protonation profiles; one aspartate has a low pK value around 1.5 whereas the other exhibits a higher than expected value of 4.8. These values vary from one enzyme to another but the principle of two different pK values for aspartyl side chains is conserved in all proteins.

Pepsin shows optimal activity around pH 2.0 and substrates bind to the active site of pepsin with the C=O group of the scissile bond situated between the Asp side chains with the oxygen atom in a position comparable to that of the displaced water molecule. The unionized side chain of Asp215 protonates the carbonyl group allowing the displaced water molecule, hydrogen bonded to the side chain of Asp32, to initiate a nucleophilic attack on the carbonyl carbon atom by transferring a proton to Asp 32 (Figure 12.9). The resulting proton transfer forms a tetrahedral intermediate that promptly decomposes with protonation of the NH group.

HIV protease uses a slightly different trick to achieve the same catalytic function. The protein is not a single polypeptide chain but is a symmetrical homodimer with each chain containing 99 residues (Figure 12.10). In itself this organization supports a picture of aspartyl proteases arising from duplication of an ancestral gene. Crystallography shows the structure of this dimer as two identical chains each subunit containing a four-stranded β sheet formed by the N- and C-terminal strands together with a small α helix. The enzyme's active site is unusual in that it is formed at the interface of the two subunits. However, again the active site is based around a group of three residues in each subunit with the sequence Asp25-Thr26-Gly27 containing the catalytically relevant acidic side chain that is responsible for bond cleavage. The residues in the second subunit are often denoted as Asp25′, Thr 26′ and Gly27′.

The cavity region is characterized by two flexible flaps at the top of each monomer formed by an antiparallel pair of β strands linked by a four residue β turn (49–52). These flaps are also seen in pepsin and guard the top of the catalytic site with the active site Asp residues located at the bottom of this cleft. These 'flap' regions are flexible, and from studies of HIV protease in the presence of inhibitors it was observed that significant conformational changes occur in this region. Some backbone C_α atoms show

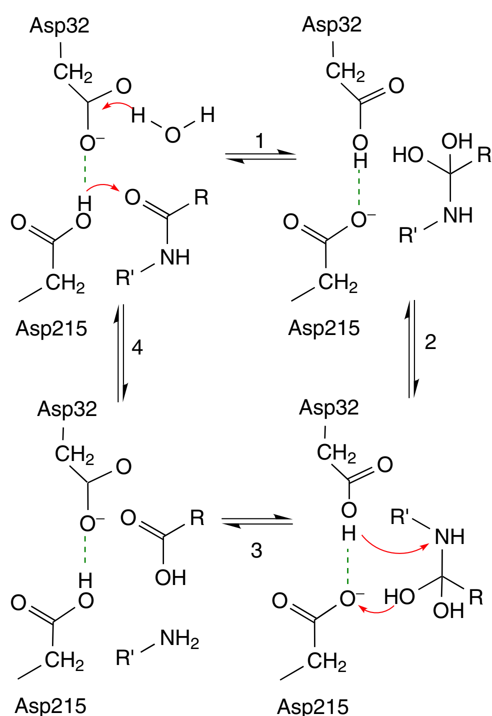

Figure 12.9 The catalytic activity of aspartyl proteases such as pepsin. The reaction mechanism involves the two aspartyl side chains interacting directly with substrate and a bound water molecule. Asp32 accepts a proton from water whilst Asp215 donates a proton to the substrate (1), a tetrahedral intermediate formed breakdown via the donation and acceptance of protons (2) with the resulting product yielding a free amino group and a free carboxyl group within the substrate (3). The dissociation of products allows the pepsin to start the reaction cycle again (4)

displacements of ~0.7 nm when the bound and free states of the enzyme are compared around the active site region.

Catalysis is mediated by twin Asp25 residues characterized by pK values that deviate considerably from a normal value around 4.3. One carboxyl group exhibits an unusually low pK of 3.3 whilst the other displays an unusually high pK of 5.3. The catalytic diad interacts with the amide bond of the target substrate in a manner analogous to that describe for pepsin. Inhibiting the protease represents one potential method of fighting HIV infection since this will prevent processing of the Gag-Pol polyprotein and the generation of proteins necessary for new virus formation. Many crystal structures of HIV-1 protease together with inhibitor complexes are known as part of drug development aimed at catalytic inhibition. In view of the preference of HIV protease for Tyr-Pro and Phe-Pro sequences within target substrates inhibitors have been designed to mimic the transition state complex and to prevent, by competitive inhibition, the turnover of the protease with native substrate (Figure 12.11).

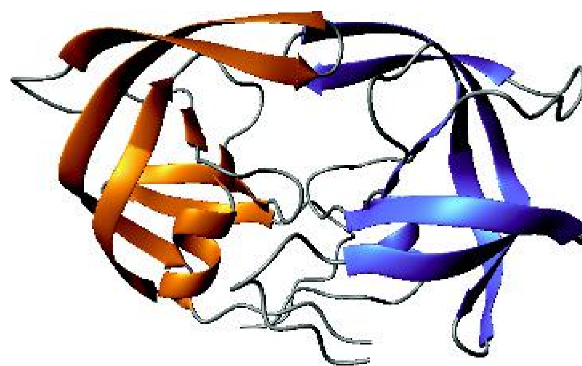

Figure 12.10 The HIV protease homodimer. The two monomers are shown in orange and blue

The protease inhibitors are known as peptidomimetic drugs because of their imitation of natural peptide substrates. These drugs mimic the enzyme's transition state tetrahedral complex and attempt to reproduce the hydrophobic side chain of phenylalanine, the restricted mobility of prolyl side chains and a carbonyl centre that is not easily protonated.

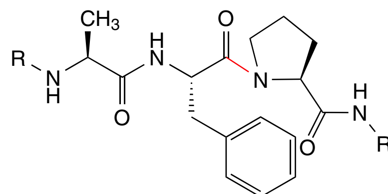

Figure 12.11 Ala-Phe-Pro sequence containing the target peptide bond (substrate) for HIV protease. The bond to be cleaved is shown in red

Figure 12.12 The structure of saquinavir

Several drugs were rapidly brought from the laboratory through clinical trials and into effective use in 1995 and included ritonavir, indinavir, nelfinavir and saquinavir (Figure 12.12). The presence of additional functional groups such as OH groups close to the target peptide bond coupled with aromatic groups results in inhibitors that bind with high affinity ($K_I < 1$ nM) and decrease viral protease activity by acting as competitive inhibitors. Although these drugs are responsible for remarkable increases in life expectancy they do not eliminate virus from the body. In general, the protease inhibitors are given when treatment with drugs aimed against reverse transcriptase fails or may be given in combination with other drug therapies. Unsurprisingly, these drugs have side effects but have proved effective at limiting viral proliferation. A major problem with this approach is resistance due to mutations in HIV protease that prevent the inhibitor from binding with optimal efficiency. Mutations within HIV protease lead to a loss of inhibitor binding but protein that retains catalytic activity. The mutations arise from the high error rate associated with activity of reverse transcriptase.

Reverse transcriptase The other major target for antiviral drug therapy is reverse transcriptase. The viral genome of RNA is transcribed into DNA by the action of reverse transcriptase followed by the integration of viral DNA into host cell chromosomes by the integrase. Reverse transcriptase is derived from the Gag-Pol precursor by proteolytic cleavage producing two homodimers of mass 66 000 (p66). The p66 subunit is bifunctional with a polymerase activity located in the larger of two domains and RNase activity located in a second smaller domain (called RNaseH). The mature, i.e. active reverse transcriptase is unusual in that it requires proteolytic removal of one of the RNase domains from a p66 subunit to form a heterodimer of p66 and p51 subunits. This heterodimer represents the functional enzyme of HIV.

Inhibiting the action of reverse transcriptase should slow down new virus production and by implication disease progression. To understand how this might be achieved it was necessary to understand the structure and function of the enzyme, as well as its interaction with substrates before effective inhibitors could be designed. From the structure of p66 (Figure 12.13) and p51 (Figure 12.14) subunits in the presence of DNA and inhibitors as well as unliganded states a number of important facts have emerged. The polymerase domain is composed of 'finger', 'palm', 'thumb' and 'connecting' regions and this is reminiscent of the organization of other RNA polymerases. Significantly the p66 subunit shows structural similarity to the Klenow fragment of *E. coli* DNA polymerase I suggesting that conservative patterns of structure and organization exist within this class of enzymes.

Studies of both p66 and p51 subunits suggested that binding DNA does not perturb complex structure significantly, but that p51 and p66 differ in their

Figure 12.13 The organization of reverse transcriptase p66 subunit is shown. The finger domain is shown in red, the thumb in dark red, the palm in orange and the connecting region in gold. The active site aspartate residues lie at the base of the cleft between thumb and fingers

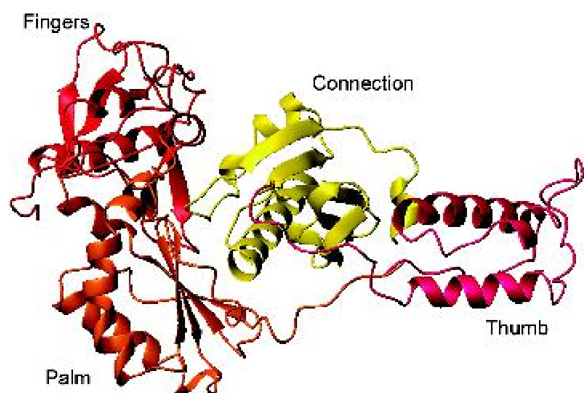

Figure 12.14 The p51 subunit of reverse transcriptase. The different conformations are clearly seen with the relative orientations of sub-domains forming a closed conformation in p51 and an open state in p66

structures largely as a result of the reorientation of domains or sub-domains within each subunit. The 'fingers' sub-domain is composed of β strands and three α helices whilst the 'palm' sub-domain possesses five β strands hydrogen bonded to four β strands at the base of the "thumb" region. The remainder of the 'thumb' region is predominantly α helical and links to a 'connection' sub-domain between the polymerase and RNaseH domains. The 'connection' domain is based around a large extended β sheet with two solitary helices. In p66 subunits the sub-domains are arranged in an open configuration exposing the active site containing three critical Asp side chains at residues 110, 185 and 186. In contrast, structural analysis of the p51 subunit reveals that, despite identical sequences, the polymerase sub-domains are packed differently with the fingers covering the palm and by implication burying the catalytically important active site residues. Unsurprisingly, functional studies reveal the p51 domain to be catalytically inactive although it interacts with the RNaseH domain of the p66 subunit and is important to the enzyme's overall conformation.

Inhibitors of reverse transcriptase fall into two classes. The first group based around the natural nucleotide substrate bind to the active site pocket whilst the second group, generally described as non-nucleoside inhibitors, function by binding at sites remote from the catalytic centre. Taken together these inhibitors represent two very different mechanisms of action. The first, and archetypal, reverse transcriptase inhibitor is 3′-azido-3′ deoxythymidine (AZT) and was the first HIV antiretroviral to be licensed for use. It contains an azido group in place of a 3′ OH group in the deoxyribose sugar. In a comparable fashion to dideoxynucleotides AZT acts as a chain terminator because it lacks a 3′ OH group. When the drug enters cells it is phosphorylated at the 5′ end and incorporated into growing DNA chains by the action of reverse transcriptase on the RNA template (HIV genome). The absence of a 3′ OH ensures that polymerization cannot continue. Although it might be expected that host DNA polymerases would be inhibited with severe consequences it appears fortuitous that host cell polymerases do not bind these inhibitors with high affinity. As a result AZT is effective against viral replication by inhibiting only reverse transcriptase activity. Competitive inhibition has proved extremely effective in the primary treatment of HIV with AZT, introduced in 1987 and complemented by additional nucleoside inhibitors, the accepted treatment option (Figure 12.15).

The non-nucleoside inhibitors (Figure 12.16) bind to a hydrophobic pocket remote from the active site. The effect of binding is to shift three β strands of the palm sub-domain containing the active site aspartyl residues towards a conformation seen in the p51 subunit. In other words, these inhibitors favour

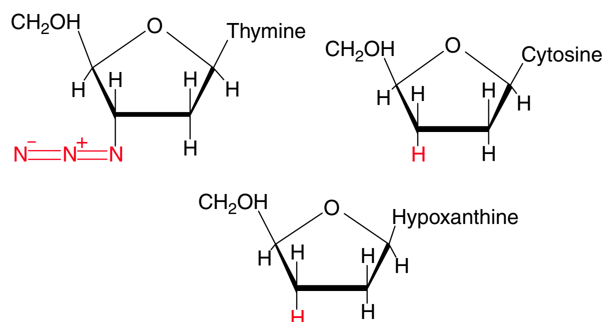

Figure 12.15 Active site (nucleoside inhibitors) of reverse transcriptase. Besides AZT, 2′,3′-dideoxycytidine (ddC) and 2′3′-dideoxyinosine (ddI) are other commonly used nucleoside analogues for fighting HIV infection

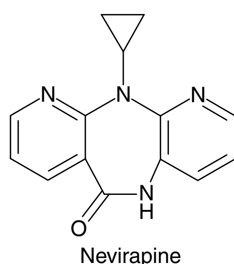

Figure 12.16 Nevirapine, a non-nucleoside inhibitor of HIV reverse transcriptase

formation of inactive states by distorting the catalytic site as the basis for pharmacological action.

To date these drugs represent the most effective route of controlling HIV infection and are particularly important in the continual fight against viral proliferation. The combined use of nucleoside and non-nucleoside inhibitors in a cocktail known as highly active antiretroviral therapy (HAART) drastically reduces viral load (the number of copies of virus found within blood) and is associated with improved health. Unfortunately the spectre of drug resistance means that additional drugs or alternative regimes are needed to effectively combat the virus over a sustained period of time.

The introduction of antiviral drugs has drastically improved the quality of life and life expectancy of individuals infected with HIV but the prohibitive cost of providing these medicines has restricted treatment to countries with advanced forms of health care. This is particularly regrettable since HIV infection rates and deaths attributable to advanced HIV disease are most prevalent in the less economically advantaged areas of the world. Currently, many millions of individuals infected with HIV are unable to obtain effective life extending treatment.

The surface glycoproteins of HIV and alternative strategies to fighting HIV

An alternative approach to fighting HIV infection is to counter the ability of the virus to enter cells via the interaction between surface viral glycoproteins and the host cell receptors. The surface of HIV is coated with spikes that represent part of the envelope glycoprotein gp120. This exposed surface glycoprotein is anchored to the virus via interactions with the transmembrane protein gp41. Both proteins are synthesized initially as a precursor polyprotein – gp160. The envelope proteins gp41 and gp120 have been characterized as part of systematic attempts to understand the mechanism of entry into cells.

Viral entry is initiated by gp120 binding to receptors such as CD4, a protein containing immunoglobulin-like folds, expressed on the surface of a subset of T cells and primary macrophages. The gp120 protein binds to CD4 with high affinity exhibiting a dissociation constant (K_d) of ~4 nM and is a major route of HIV entry into cells although additional membrane proteins (the chemokine receptors CXCR4 and CCR5) can also operate in viral entry.

Structural studies of gp120 are aimed at understanding the molecular basis of protein–protein interactions. However, gp120 has proved a difficult subject because extensive glycosylation has prevented crystallization. Studies have therefore involved partially glycosylated core domains containing N- and C-terminal deletions and truncated loop regions. Immunoglobulin-like variable domains exist within gp120 in the form of five regions (V1–V5) and are found in all HIV strains. These variable zones are interspersed with five conserved regions. Four of these variable regions are located as surface-exposed loops containing disulfide links at their base. Despite substantial deletions the core gp120 binds to CD4 and presents a credible version of the native interaction.

The gp120 core structure has ‘inner’ and ‘outer’ domains linked by a four-stranded β sheet that acts as a bridge. The inner domain contains two α helices, a five-strand β sandwich and several loop regions, whilst the larger outer domain comprises two β barrels of six and seven strands, respectively. The core is often described as heart shaped with the N- and C-terminals located at the inner domain (top left-hand corner, Figure 12.17), the V1 and V2 loop regions located at the base of the inner domain and the V3 loop region located at the base of the outer domain. The V4 and V5 loop regions project from the opposite side of the outer domain.

The inner domain is directed towards the virus and from mutagenesis studies the interaction site between gp120 and CD4 is defined broadly at the interface

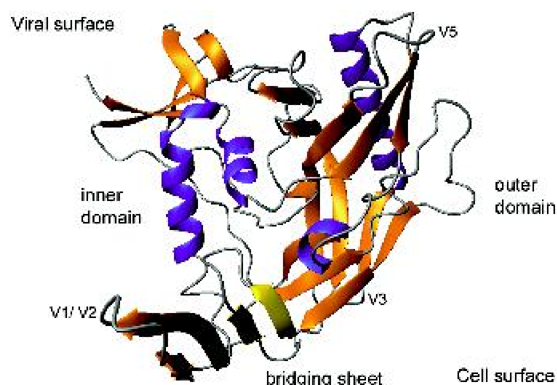

Figure 12.17 The structure of the gp120 core determined in a complex with a neutralizing antibody and a two domain fragment of CD4. Only the gp120 is shown for clarity. The bridging sheet is probably in contact with the cell surface with the N-terminus lying closer to the virus envelope surface (PDB: 1G9N)

of all three domains. This region contains conserved residues required for CD4 binding as well as variable residues. CD4 interacts with the variable residues through hydrogen bonds to main-chain atoms, but in addition cavities exist within the binding surface that are not in contact with CD4 (Figure 12.18). Within these cavities are many conserved residues and their mutation (as a result of error-prone RT activity) perturbs the interaction between CD4 and gp120 despite an absence of direct interactions in the crystal structures. The binding of CD4 to gp120 also results in interactions with chemokine co-receptors, and this site of interaction is distinct from that involved in CD4 binding. Current evidence suggests the co-receptor binding site is based around the bridging sheet between inner and outer domains, but on the opposite side of the molecule.

Initial attempts to construct a vaccine against HIV focused on raising specific antibodies against the gp120 component of the virus since this represents the first step of molecular recognition. Unfortunately immune recognition of the CD4 binding site is poor because of the mix of conserved and variable regions forming the binding site, the V1 and V2 regions mask the CD4 binding site prior to binding, and the CD4-bound state may represent a transient conformation of gp120. Vaccine development based on gp120 has not completed clinical trials.

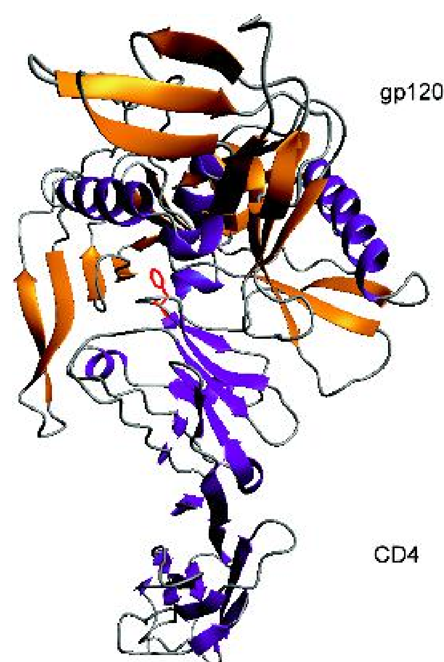

Figure 12.18 Interactions between the core of gp120 and the N-terminal immunoglobulin domain of CD4. The side chain of Phe43 of CD4 is shown capping a large hydrophobic cavity on gp120

The structure and function of gp41

The other major envelope protein is gp41, a transmembrane protein interacting with gp120, which consists of an N-terminal ectodomain, a transmembrane region and a C terminal intraviral segment within a single polypeptide chain of 345 residues. gp41 interacts with the matrix protein and its principal role is to mediate fusion between the viral and cellular membrane following gp120-based receptor binding.

At the N-terminal region a hydrophobic, glycine-rich, 'fusogenic' peptide is found and all structures presented to date have been of the ectodomain lacking this region. The ectodomain is a symmetrical trimer with each monomer consisting of two antiparallel α helices connected by an extended loop (Figure 12.19).

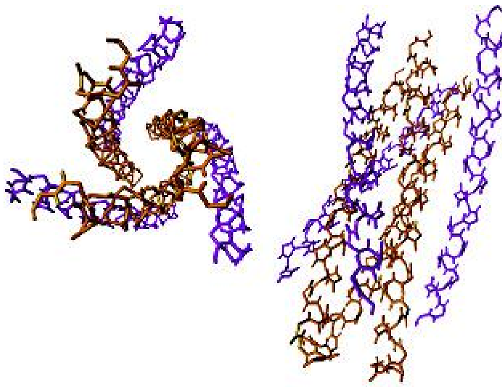

Figure 12.19 Part of the ectodomain region of gp41 is a coiled coil structure consisting of three pairs of helical segments

The combined action of gp120 and gp41 is the first contact point between virus and host cell. Unsurprisingly, if this stage can be inhibited then infection by HIV will be limited. Extensive proteolysis of a recombinant ectodomain revealed the existence of stable subdomains and a trimeric helical assembly composed of two discontinuous peptides termed N36 and C34 that are resistant to denaturation. The N- and C-peptides originate from the N-terminal and C-terminal regions of the gp41 ectodomain. The N-terminal derived peptides form a central trimeric 'coiled-coil' with the helical C-peptides surrounding the coiled coil in an antiparallel arrangement. This structural model was confirmed by X-ray crystallography for both HIV-1 and SIV peptides and is described as a trimer of hairpins (Figure 12.20).

In the trimer the C-peptides bind to the outside of the coiled-coil core interacting with a conserved hydrophobic groove formed by two helical N-terminal peptides. Although the structure of the intervening sequence of gp41 is unclear it would be required to loop around from the base of the coiled coil to allow the C-peptide to fold back to the same end of the molecule.

The mechanism of viral fusion to cells is slowly being unraveled and it may be very similar to fusion

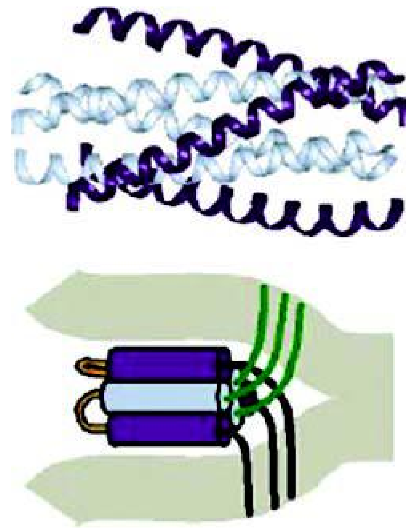

Figure 12.20 The trimer of hairpins exhibited by gp41 core domains (Reproduced by with permission from Eckert, D.E. & Kim, P.S. *Annu. Rev. Biochem.* 2001, **70**, 771–810. Annual Reviews Inc)

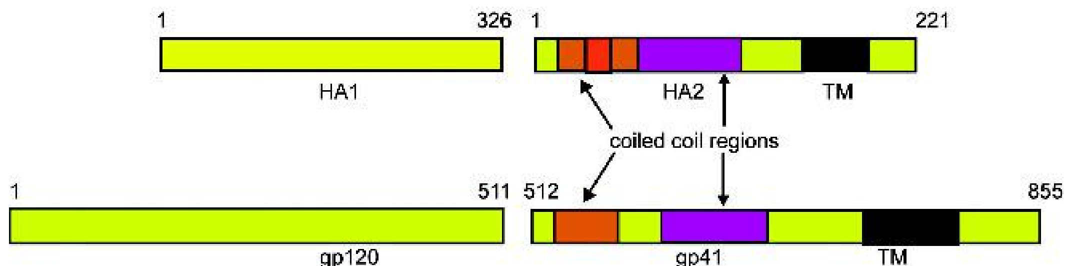

Figure 12.21 The organization of haemaglutinnin and HIV gp41/gp120 based around interacting helical regions or coiled coils

processes exhibited by the influenza virus during host cell infection. Some similarities exist between the sequences of gp41 and haemagglutinin (Figure 12.21) and may indicate potentially similar mechanisms exist *in vivo*.

However, this analogy cannot be taken too far, because HIV, unlike influenza virus fuses directly with cell membranes in a pH-independent manner. The trimer of hairpins is most probably the fusogenic state and is supported by the observation that gp41-derived peptides or synthetic peptides inhibit HIV-1 infection. Peptides derived from the C-terminal region are much more effective than those derived from N-peptides with inhibition noted at nM concentrations. These peptides act by binding close to the N-peptide region and preventing the hairpin structure. The observed inhibition of HIV by these peptides offers another route towards the development of antiviral drugs. Drugs that target additional steps in the viral life cycle, such as viral entry, are valuable in providing alternative therapies that may prove less toxic and less susceptible to viral resistance than current regimes.

The influenza virus

Influenza (or 'flu) is often considered a harmless disease causing symptoms slightly worse than a cold. Common symptoms are pyrexia, myalgia, headache and pharyngitis and the infection ranges in severity from mild, asymptomatic forms to debilitating states requiring several days of bed rest. The virus spreads by contact specifically through small particle aerosols that infiltrate the respiratory tract and with a short incubation period of 18 to 72 hours virus particles appear in body secretions soon after symptoms start causing further infection. In the respiratory tract the virus infects epithelial cells causing their destruction and susceptibility to infections such as pneumonia.

Modern medicines alleviate the worse symptoms of flu but this disease has killed millions in recent history and continues to present a major health problem. Periodically virulent outbreaks of influenza occur and one of the worst events occurring around 1918 and known as Spanish flu is conservatively estimated to have killed 20 million people. This flu pandemic resulted in more fatalities than all of the battles of

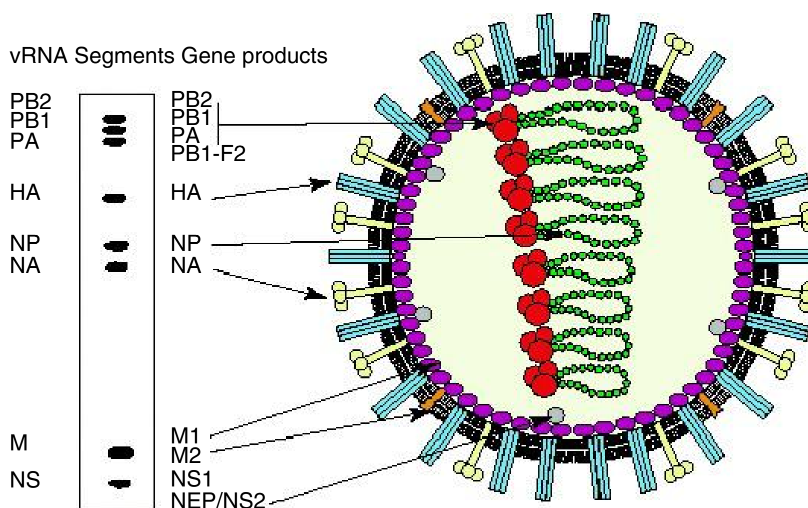

Figure 12.22 The influenza A virus. Nucleoprotein (green) associates with each RNA segment to form RNP complexes that contain PB1, PB2 and PA (red). The viral glycoproteins, haemagglutinin (blue) and neuraminidase (yellow), are exposed as trimers and tetramers. Matrix protein (purple) forms the inner layer of the virion along with the transmembrane M2 protein (orange). Non-structural proteins are also found (grey). (Reproduced with permission from Ludwig, S. *et al Trends Mol. Medicine* **9**, 46–52, 2003. Elsevier)

Table 12.3 The coding of different viral proteins and their functions by the segmented RNA genome

Segment	Size	Name	Function
1	2341	PB2	Transcriptase; Cap binding
2	2341	PB1	Transcriptase; Elongation
3	2233	PA	Transcriptase; Protease activity?
4	1788	HA	Haemagglutinin; membrane fusion
5	1565	NP	Nucleoprotein RNA binding; part of transcriptase complex; nuclear/cytoplasmic transport of vRNA
6	1413	NA	Neuraminidase: release of virus
7	1027	M1	Matrix protein: major component of virion
		M2	Integral membrane protein – ion channel
8	890	NS1	Non-structural: nucleus; role in on cellular RNA transport, splicing, translation. Anti-interferon protein.
		NS2	NS2 Non-structural: function unknown

World War I combined and emphasizes the serious nature of influenza outbreaks. Despite drugs and the use of vaccination influenza propagates around the world at regular intervals. There remains genuine fear of further major and lethal outbreaks striking again with a severity comparable to the events of 1918. When another pandemic occurs there is little doubt that we are in a much better position to deal with influenza than previous generations. This position of strength arises because the influenza virus has been extensively studied especially in relation to structural organization and interaction with cells.

The virus is a member of the orthomyxovirus family based around a (-) strand of RNA composed of eight segments in a total genome size of ~14 kb flanked by short 5′ and 3′ terminal repeat sequences (Figure 12.22). The segments bind to a nucleoprotein (NP) and associate with a polymerase complex of three proteins (PA, PB1 and PB2) that constitutes a ribonucleoprotein particle (RNP) and is located inside a shell. The shell is formed by the matrix (M1) protein and is itself surrounded by a lipid membrane that is derived from the host cell but contains two important glycoproteins – haemagglutinin (HA) and neuraminidase (NA) – together with a transmembrane-channel protein (M2). Two non-structural components (NS1 and NS2) are also encoded by the viral genome to make a total of 10 proteins[1] (Table 12.3).

Viral classification is based around the M and NP antigens and is used to determine strain type – A, B or C. Most major outbreaks of influenza are associated with virus types A or B with type B strain causing milder forms of the disease. The external antigens haemagglutinin and neuraminidase show greater sequence variation.

Viral infection and entry into the cell

The virus binds to sialic-acid-containing cell surface receptors via interactions with haemagglutinin entering the cell by endocytosis where fusion of the viral and endosomal membranes allows the release of viral RNA. After transfer to the nucleus the viral genome is transcribed and amplified before translation of mRNA in the cytoplasm. Late in the infection cycle the viral genome is transported out of the nucleus in RNP complexes and packaged with the structural proteins. Subsequent release of new virions completes a cycle

[1]For many years this was thought to be the complete complement of viral proteins but in 2002 further examination of the genome revealed an alternate reading frame in PB1 that allowed production of an eleventh protein (F2).

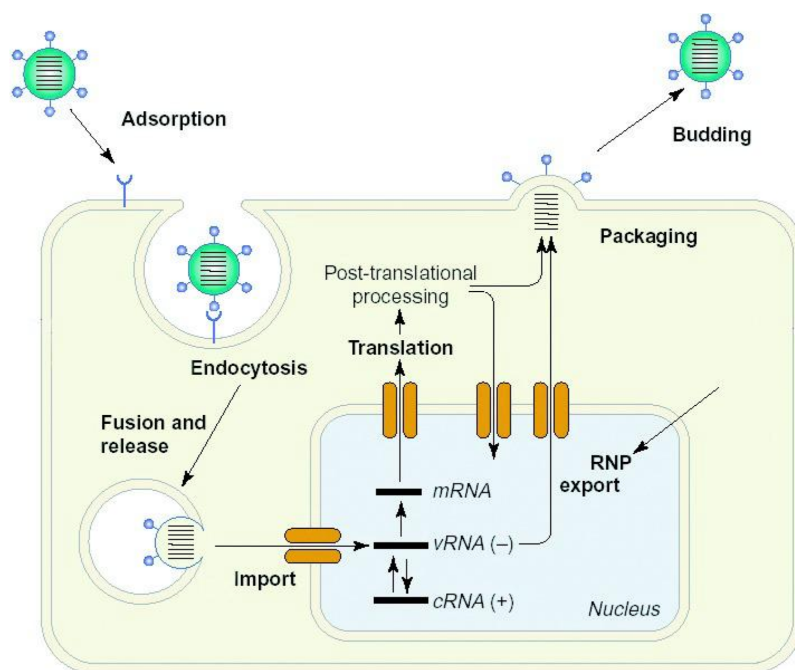

Figure 12.23 The lifecycle of the influenza A virus and the pathway of viral entry and exit from host cells. (Reproduced with permission from Ludwig, S. *et al Trends Mol. Medicine* **9**, 46–52, 2003 Elsevier)

that occurs over a time span between 6 to 16 h (Figure 12.23). Of critical importance to viral entry are the characteristic "spikes" of haemagglutinin and together with the second major envelope protein, neuraminidase, these proteins have been the targets for drug treatments to fight "flu".

Haemagglutinin structure and function

Haemagglutinin was identified from the ability of flu viruses to agglutinate red blood cells via the interaction between HA and sialic acid groups present on cell surface receptors. Haemagglutinin is anchored to the membrane by a short tail but limited proteolysis with enzymes such as bromelain or trypsin yield a soluble domain (Figure 12.24). The soluble domain was amenable to crystallographic analysis and in 1981 a structure of haemagglutinin determined by John Skehel, Donald Wiley and Ian Wilson provided the first detailed picture of any viral membrane protein.

Haemagglutinin is a cylindrically shaped homotrimer approximately 135 Å in length. It is synthesized as a precursor polypeptide (HA0) that associates to form a trimer but is first cleaved by proteases to give two disulfide linked subunits -HA1 and HA2. HA1 and HA2 are necessary for membrane fusion and virus infectivity and generally the polypeptide chains are separated by just one residue, a single arginine residue, removed from the C terminus of HA1 by carboxypeptidase activity. It is the HA1/HA2 assembly that is commonly, but slightly inaccurately, referred to as a monomer and haemagglutinin contains three HA1 and three HA2 chains. The membrane-anchoring region is located at the C terminal region of HA2 but structural studies were performed on tail less protein.

The crystallographic structure shows the free ends, at the cleavage site between HA1 and HA2, are separated by 20 Å and within the structure two distinct regions were readily identified; a globular head region responsible for binding sialic acid residues on host cell

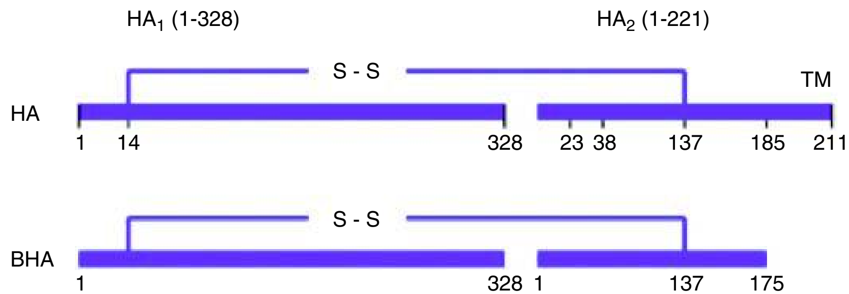

Figure 12.24 The membrane bound and protease treated form of haemagglutinin showing HA1 and HA2 regions. BHA refers to bromelain (protease) treated haemagglutinin

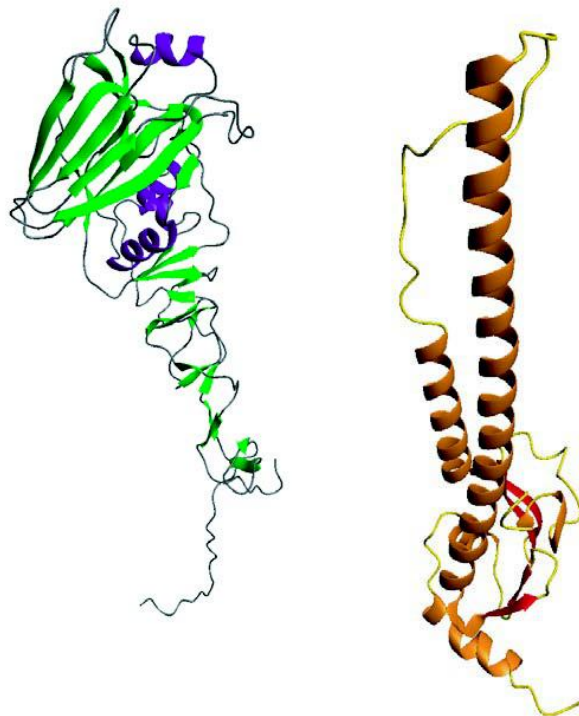

Figure 12.25 The individual HA1 and HA2 units of haemagglutinin

receptors and a longer extended stem-like region that contains a fusogenic sequence (Figure 12.25).

The three monomers assemble into a central α-helical coiled-coil region that forms the stem-like domain whilst at the top are three globular heads containing sialic acid-binding sites and responsible for cell adhesion or binding. The stem region consisting of a triple helix extends for ~76 Å from the membrane

surface and is formed predominantly by the HA2 polypeptide. The globular region is based mainly on the HA1 portion of the molecule and each globular domain consists of a swiss roll (jelly roll) motif of eight anti-parallel β-strands by residues 116–261. At the distal tips receptor (sialic acid) binding pockets were identified. The binding sites reside in small depressions on the surface and were detected from crystallographic studies in the presence of different sialic acid derivatives. Amongst the residues forming the active sites are Tyr98, Ser136 Trp153, His183, Glu190, Leu194, Tyr195, Gly225 and Leu226 with site directed mutagenesis suggesting that Tyr98, His183, and Leu194 are particularly important in forming critical hydrogen bonds with the substrate (Figure 12.26).

Sialic acid groups bind with one side of the pyranose ring facing the base of the site whilst the axial carboxylate, the acetamido nitrogen, and the 8- and 9-OH groups of the sugar form hydrogen bonds with polar groups of conserved residues. Histidine 183 and glutamate 190 form hydrogen bonds with the 9-OH group, tyrosine 98 forms hydrogen bonds with the 8-OH group whilst the 5-acetamido nitrogen forms a hydrogen bond with the backbone carbonyl of residue 135 and the methyl group makes van der Waals contact with the phenyl ring of tryptophan 153. The 4-OH group projects out of the site and does not appear to participate in binding interactions.

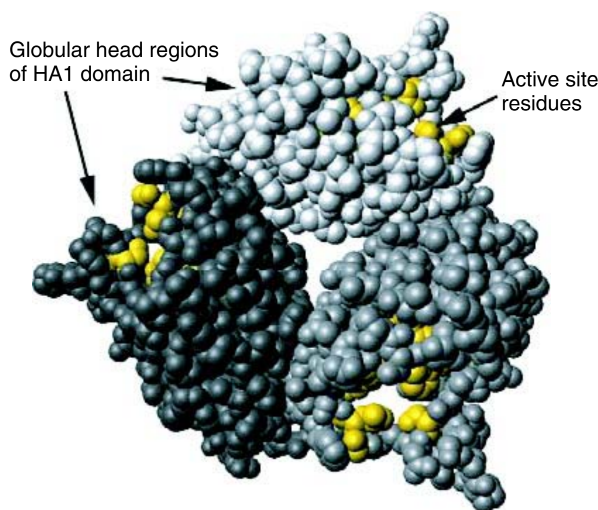

Figure 12.26 The active site residues of haemagglutinin are shown in the spacefilling representations of the globular HA1 domain. The view is from above the globular trimer

A key observation in understanding the structure and function of haemagglutinin was the demonstration that the molecule underwent a large conformational change when subjected to low pH. The HA proteolytic fragment does not bind to membranes and is functionally inactive but lowering the pH to a value comparable to that occurring in endosomes (~pH 5–6) during viral entry caused haemagglutinin to become a membrane protein. Electron microscopy showed protein aggregation into "rosettes" with an elongated and thinner structure. These observations suggested major conformational changes and performing crystallography on the 'low pH' form resolved the extent and location of structural perturbations (Figures 12.27 & 12.28).

At low pH the fusion region or peptide was located at the end of an elongated coiled-coil chain with domains near the C-terminal interacting with those at the N terminus. The low pH form is the thermodynamically stable form of HA2 and the initial state is best described as a metastable configuration. Exposure to low pH converts this metastable form into the thermodynamically stable form of HA2.

In total the structures of the ectodomain of haemagglutinin have been determined for the single-chain precursor (HA0), the metastable neutral-pH conformation and the fusion active, low pH-induced conformation. These structures provide a framework for interpreting results of experiments on receptor binding, the cyclical generation of new and re-emerging epidemics as a result of HA sequence variation as the process of membrane fusion during viral entry.

Neuraminidase structure and function

Neuraminidase is vital to viral proliferation. The enzyme is a glycosyl hydrolase cleaving glycosidic linkages between N-acetyl neuraminic acid and adjacent sugar residues by destroying oligosaccharides units present on host cell receptors. Viral neuraminidase

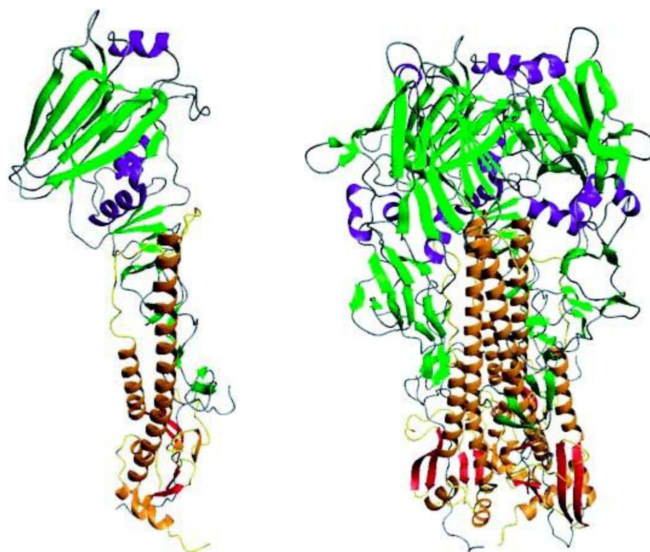

Figure 12.27 The structure of the HA1/HA2 'monomer' and a complete assembly of trimeric haemagglutinin (PDB:2HMG)

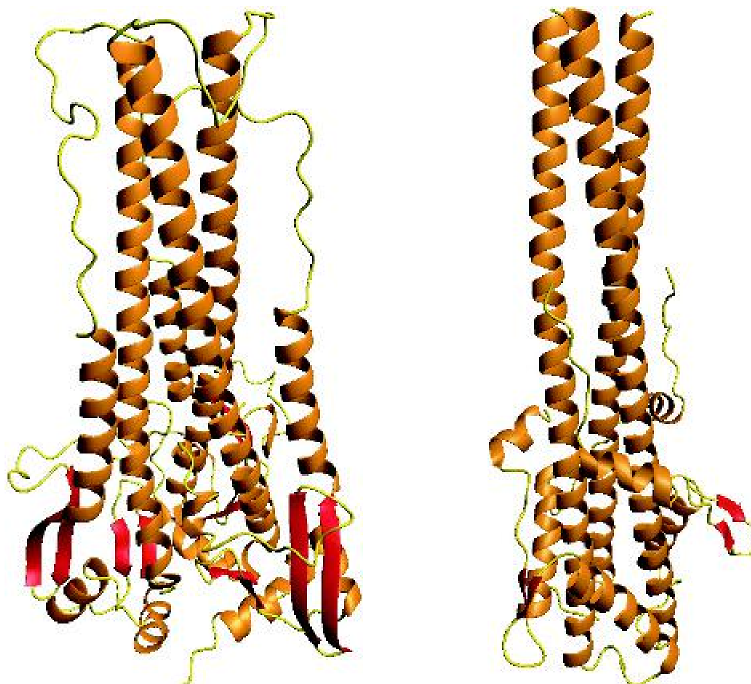

Figure 12.28 The changes in conformation associated with the coiled-coil regions of HA2 in the transition from neutral pH form to active haemagglutinin. To ascertain the change in conformation calculate the number of turns associated with the main helices

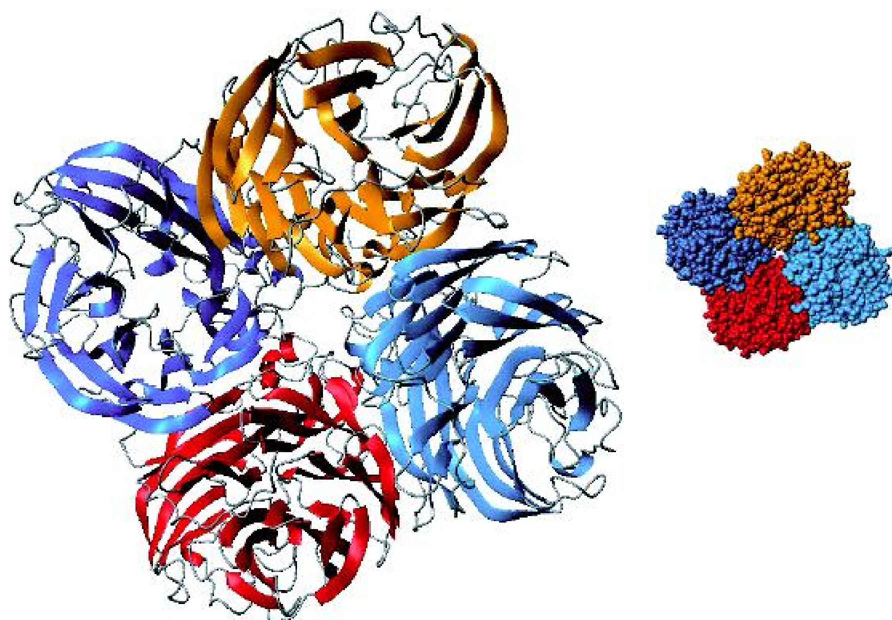

Figure 12.29 The box-like arrangement of the tetrameric state of neuraminidase shown alongside a spacefilling model of the same view

causes a loss of agglutination and allows subsequent infection of other host cells.

The amphipathic protein is a single polypeptide chain of ~470 residues with a hydrophobic membrane binding domain of ~35 residues connected to a larger polar domain of ~380 residues via a hypervariable stalk of ~50 residues. Limited proteolysis removes the N terminal membrane anchor and facilitated crystallization. The structure of the large fragment of neuraminidase revealed a homo-tetramer with monomers consisting of six topologically related units (Figure 12.29). Each unit contained a four-stranded sheet arrayed like the petals of a flower or more frequently described as one of the blades of a propeller (Figure 12.30).

Neuraminidase is an example of the β propeller motif. The four strands connect by reverse turns with the first strand of one sheet interacting across the top of the motif to the fourth strand of the adjacent motif. Disulfide bonds between adjacent strands contribute to a distortion of regular sheet conformation and introduce a distinct twist to each blade.

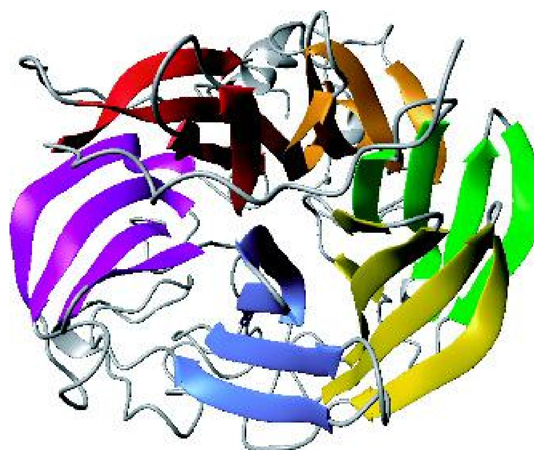

Figure 12.30 The neuraminidase monomer. Whilst the structure of the larger soluble fragment has been determined the structure of the membrane spanning region is unknown. The structure shows one monomer with six topologically related units (blades) composed of four β strands; each blade is shown in a different colour. (PDB:2BAT)

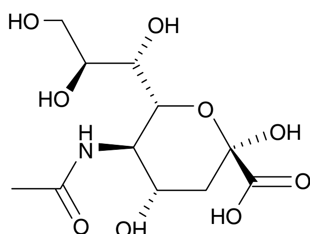

Figure 12.31 The structure of sialic acid (*N*-acetylneuraminic acid, Neu5Ac)

The monomeric units assemble into a box-shaped head with dimensions of ~100 Å × 100 Å × 60 Å attached to a slender stalk region. The tetrameric enzyme has circular 4-fold symmetry stabilized in part by metal ions bound on the symmetry axis with each subunit interacting via an interface dominated by hydrophobic interactions and hydrogen bonds.

Definition of the active site of neuraminidase proved important in designing inhibitors that interfere with neuraminidase activity and numerous crystal structures exist for enzyme-inhibitor complexes. Initial studies focused on a transition state analogue, 2-deoxy-2, 3-dehydro-*N*-acetylneuraminic acid (Neu5Ac2en), that inhibited the enzyme at micromolar concentrations but further refinement yielded inhibitors effective at nanomolar levels (Figures 12.31 & 12.32).

One of these compounds 4-guanidino-Neu5Ac2en (Zanamivir) inhibits viral proliferation as a result of significant numbers of hydrogen bonds formed between drug and *conserved* residues in the active site of neuraminidase (Figure 12.33).

Strategies to combat influenza pandemics

Influenza infection leads to a pronounced antibody response against haemagglutinin and neuraminidase based on the surface epitopes presented by the infecting virus. Ordinarily individuals have protective immunity but in the case of flu frequent changes in the sequences of these two antigens compromises the normal immune response. Vaccines against flu viruses are

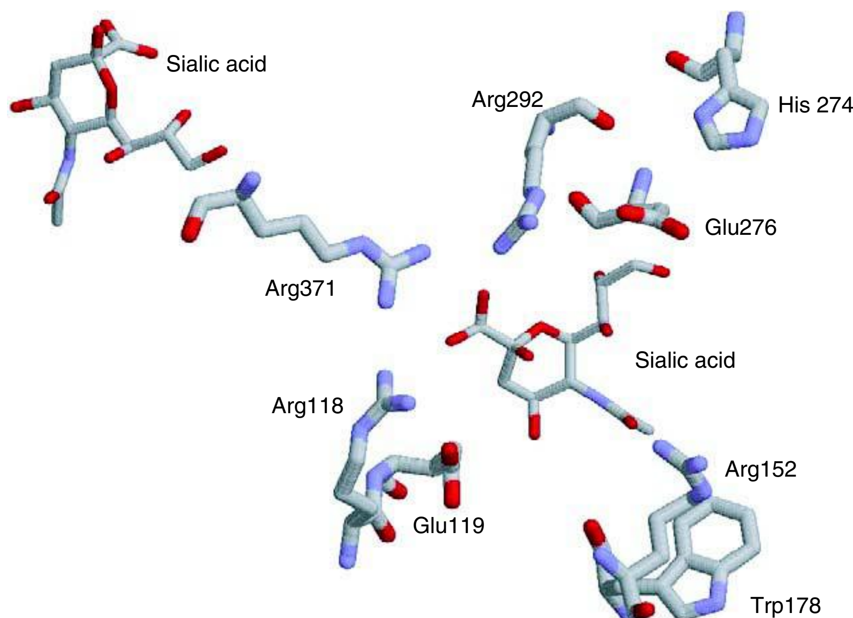

Figure 12.32 The active site of neuraminidase showing substrate bound to active site (PDB:1MWE). A second sialic acid binding site was determined on the surface of neuraminidase

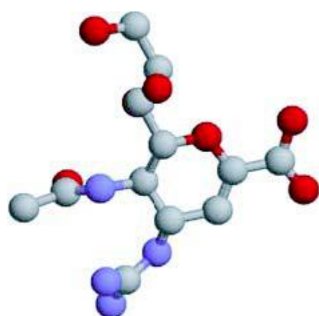

Figure 12.33 The molecular structure of Zanamivir. A potent inhibitor of neuraminidase that is now accepted as clinical treatment for flu in most countries

recommended by the WHO[2] to contain haemagglutinin components from two A strains and a single B strain and to be representative of viruses currently circulating the world. Unfortunately although vaccines are developed from purified HA the virus alters the sequence of residues around the epitopes without impairing protein function. This occurs *via* point mutations and leads to altered surface properties that eliminate cross-reaction with antibodies developed against previous forms of the virus. Two distinct mechanisms contribute to variation in surface properties of the HA. Point mutations cause antigenic drift whilst a second, potentially more serious, mechanism for variation is antigenic shift where the exchange of genes between influenza viruses creates new strains.

Haemagglutinin has at least fifteen different antigenic sub-types (H1–H15) whilst neuraminidase has nine antigenic subtypes (N1–N9). Most subtypes of type A virus infect avian species with sub-types H1, H2 and H3 infecting humans. The HA1 domains show greater variability than HA2 consistent with their role as binding sites for receptors and antibodies. With the recognition of these sub-types viral outbreaks are classified according to the virus class (usually A or B), the species from which the virus was isolated, its geographical region, the year of isolation and the H and N sub-group. The 20th century witnessed three influenza pandemics; Spanish influenza (defined as an H1N1 virus) in 1918, Asian influenza (defined as an H2N2 virus) in 1957, and Hong Kong influenza (H3N2 virus) in 1968. These global outbreaks caused severe illness with high mortality rates. DNA sequencing showed that the HA and NA genes of the H1N1 viruses emerged from an avian reservoir with subtle changes in primary sequence leading to altered antigenic properties. In contrast Asian and Hong Kong outbreaks were attributed to hybrid viruses arising from new combinations of avian and human viral genes. Aquatic birds are believed to be the source of influenza viruses affecting all other animal species. Recently in 1997, a highly pathogenic avian H5N1 influenza virus was transmitted directly from poultry to humans in Hong Kong. Of the 18 people infected, six died and this heightened worries about new pandemics in the future.

Influenza and HIV represent two viruses with an enormous impact on human health. They are not unique in their mechanism of infection but in both cases they have an unfortunate ability to mutate rapidly causing changes in antigenicity or loss of drug activity. This property creates pandemics and limits our ability to fight infection through vaccines and drugs. It is very unlikely that these viruses will remain the only threat to human health. The next decades are likely to see the emergence of new strains and new viruses, perhaps involving those crossing the species boundary, whose action will need to be countered through exploitation of knowledge about the structure and function of their constituent proteins.

Neurodegenerative disease

In comparison with influenza neurodegenerative diseases are less common, less easily transmitted and most of us will escape neurodegenerative disorders entirely. Despite a low incidence in all population groups there is evidence to suggest that these diseases are of greater importance to human health than previously thought. Many of these diseases have clear genetic links that enable insight into their molecular origin.

Prion based diseases are linked by the proliferative formation of insoluble aggregates of protein arising from mutations within the *PrP* gene. Creutzfeldt-Jakob disease, fatal familial insomnia, kuru and scrapie were linked to the PrP gene and lead to

[2]WHO = World Health Organization.

neurodegenerative conditions. All of these conditions lead to abnormal brain tissue and the presence of amyloid deposits. Collectively the term amyloidogenic disease stressed the link between protein misfolding and the development of neurodegenerative disorders.

BSE and new variant CJD

Scrapie was first reported in 1732 as a disease infecting sheep in England and has been present continuously since this time. It did not appear to effect bovine livestock. However, in the 1980s a new disease spread through cattle in the United Kingdom at an alarming rate producing symptoms normally associated with TSEs. The disease in cows led to abnormal gait and posture, poor body condition, nervous behaviour, decreased milk production and a progressive loss of mass. Diseased animals lost the ability to support themselves with death occurring 2–25 weeks after the appearance of symptoms. A feature of the disease was brain tissue of spongy appearance bearing large vacuoles. The disease was christened bovine spongiform encephalopathy (BSE) but its alarming rise in frequency yielded another sobriquet of 'mad cow disease' (Figure 12.34). The appearance of BSE raised a number of important questions. What was the origin of the disease, how was it transmitted and given the role of cattle in the human food chain was there a possibility of human infection?

BSE is a chronic degenerative disease and with the observation of scrapie like symptoms it was logical to look for a prion-like particle. At the time (mid 1980s) the concept of prions was controversial with the pioneering work of Prusiner slow to gain acceptance. The appearance of the disease in cattle was puzzling given the absence of simple routes of transmission but epidemiological data suggested BSE arose from animal feed containing contaminated protein sources. BSE had a substantial impact on the livestock industry in the United Kingdom with approximately 0.3 % of the total UK herd infected and continued to rise until 1992.

In 1988 the introduction of a total ban on the use of rendered ruminant based feed to cattle minimized transmission and was combined with the large scale slaughtering of infected cattle. In 1989 these preventative actions were supplemented by a ban on human

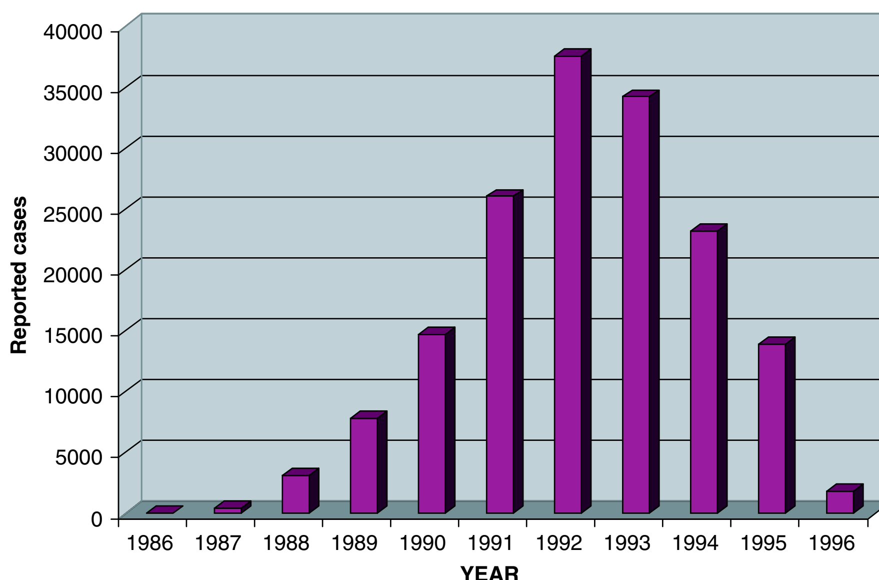

Figure 12.34 The yearly totals for clinical diagnosis of BSE in the UK herd (data derived from BSE Inquiry Report, Crown copyright 2000)

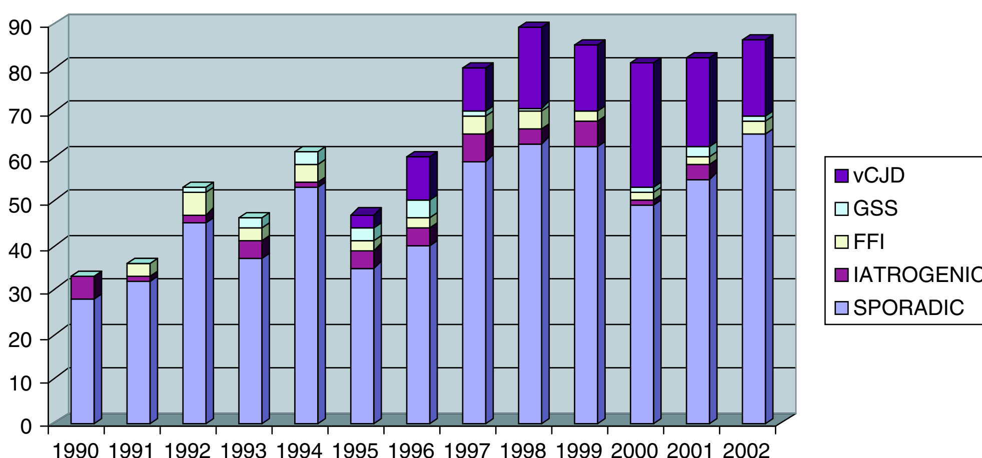

Figure 12.35 The incidence of all CJD cases in the United Kingdom (Data derived from CJD Surveillance unit, Edinburgh http://www.cjd.ed.ac.uk). From 1995 the number of cases of vCJD rise reaching a peak in 2000 before showing small decreases in the subsequent years

consumption of tissues shown to contain high levels of the infectious agent (when injected into mice brains) and therefore posing a potential danger to humans. Unfortunately this was not the end of the story because the long incubation period of BSE[3] ranging from 2 to 8 years meant that human consumption of infected cattle occurred throughout the 1980's.

In March 1996 the UK Spongiform Encephalopathy Advisory Committee announced the identification of 10 new cases of Creutzfeldt-Jakob disease (Figure 12.35). CJD is a rare disease and the observation of 10 cases was startling. Although direct evidence was lacking it seemed possible that exposure by ingestion to BSE agents was responsible for increase numbers of CJD cases. In 1994 initial speculation that BSE might be the causative agent of new cases of CJD was aired and despite a lack of definitive evidence the idea gained credibility and created widespread unease about the use of beef as a culinary product.

The 10 cases were unusual in terms of their high frequency and the clinical and pathological phenotypes associated with the condition. The new forms of CJD exhibited an onset of symptoms at a young age with teenagers or young adults most susceptible. The disease was associated with neurodegeneration and brain tissue showed lesions characteristic of spongiform encephalopathies. The new forms of CJD were called variant CJD (vCJD) and stemmed from exposure to BSE before the introduction of a specified bovine offal ban in 1989. Definitive proof of the link between BSE and vCJD required comparing both diseases at a molecular level to demonstrate that the etiological agents were prion proteins.

Comparing four different types of CJD (genetic, sporadic, iatrogenic, and "new variant") as well as BSE using SDS-PAGE showed that prion proteins exhibit distinct banding patterns when studied via Western blotting methods after treatment with proteinase K, a proteolytic enzyme that normally effectively degrades proteins into very small fragments but in the case of infectious prions produced protease resistant fragments. The pattern and intensity of classic forms of CJD (genetic, sporadic, or iatrogenic) were similar but variant CJD yielded a unique banding pattern suggesting a different genotype. The pattern was comparable to the profile seen in BSE infected organisms. Further analysis of the DNA sequence of individuals with vCJD showed that all were heterozygous at residue 129 (Valine/Methionine) suggesting that this transition conferred a genetic predisposition for contracting vCJD

[3]The time span between an animal becoming infected and then showing the first signs of disease.

whereas the homozygous (Valine/Valine) state was possibly resistant to infection.

Alzheimer's disease – a neurodegenerative disorder

Amyloidosis or protein aggregation is associated with Alzheimer's disease although like prion-based diseases little impact has been made towards effective treatment. An understanding of the molecular events involved in disease pathogenesis offers the possibility of future therapy.

Alzheimer's disease first recognized by Alois Alzheimer in the first decade of the 20th century is a slow progressive form of dementia normally seen in the later stages of life and resulting in intellectual impairment. Today a common diagnostic marker for physicians is memory impairment accompanied by a decline in language and decision-making ability, judgement, attention span, cognitive functions and personality. Diagnosis is usually confirmed by post mortem studies and the microscopic examination of brain tissue where a diagnostic feature is the presence of neurofibrillary tangles, aggregates of a single protein within brain tissue. The brain is seen to contain plaques and although plaques are detected in *all* brain tissue with age their number is drastically elevated in individuals with Alzheimer's disease.

Two types of Alzheimer's disease are recognized. Early onset disease results in symptoms before the age of 60 and has a genetic basis seen by its presence within families and an autosomal dominant form of inheritance. It accounts for less than 10 % of all forms of Alzheimer's. In contrast late onset forms are more common but the genetic basis is less well defined. Late onset disease develops in people above 65 with an incidence of 5–10 %.

The amyloid plaques found in Alzheimer's disease are composed of a 4.2 kD peptide called the Aβ protein. The plaques result from the transformation of soluble Aβ protein monomers into insoluble aggregates. The peptide called Aβ because of its amyloid properties and a partial β-stranded structure occurring within a 28-residue portion of the sequence is found in at least two forms; a 40 residue peptide and a 42 residue

```
                 *        20         *        40         *        60         *        80         *       100
APP : MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMNVQNGKWDSDPSGTKTCIDTKEGILQYCQEVYPELQITNVVEANQPVTIQNWCKR

                 *       120         *       140         *       160         *       180         *       200
APP : GRKQCKTHPHFVIPYRCLVGEFVSDALLVPDKCKFLHQERMDVCETHLHWHTVAKETCSEKSTNLHDYGMLLPCGIDKFRGVEFVCCPLAEESDNVDSAD

                 *       220         *       240         *       260         *       280         *       300
APP : AEEDDSDVWWGGADTDYADGSEDKVVEVAEEEEVAEVEEEEADDDEDDEDGDEVEEEAEEPYEEATERTTSIATTTTTTTESVEEVVREVCSEQAETGPC

                 *       320         *       340         *       360         *       380         *       400
APP : RAMISRWYFDVTEGKCAPFFYGGCGGNRNNFDTEEYCMAVCGSAMSQSLLKTTQEPLARDPVKLPTTAASTPDAVDKYLETPGDENEHAHFQKAKERLEA

                 *       420         *       440         *       460         *       480         *       500
APP : KHRERMSQVMREWEEAERQAKNLPKADKKAVIQHFQEKVESLEQEAANERQQLVETHMARVEAMLNDRRRLALENYITALQAVPPRPRHVFNMLKKYVRA

                 *       520         *       540         *       560         *       580         *       600
APP : EQKDRQHTLKHFEHVRMVDPKKAAQIRSQVMTHLRVIYERMNQSLSLLYNVPAVAEEIQDEVDELLQKEQNYSDDVLANMISEPRISYGNDALMPSLTET

                 *       620         *       640         *       660         *       680         *       700
APP : KTTVELLPVNGEFSLDDLQPWHSFGADSVPANTENEVEPVDARPAADRGLTTRPGSGLTNIKTEEISEVKMDAEFRHDSGYEVHHQKLVFFAEDVGSNKG

                 *       720         *       740         *       760         *
APP : AIIGLMVGGVVIATVIVITLVMLKKKQYTSIHHGVVEVDAAVTPEERHLSKMQQNGYENPTYKFFEQMQN
```

Figure 12.36 The primary sequence of the APP protein and its relationship to the Aβ peptide and other domains. The Aβ peptide extends from residues 672-711/713 (green) and overlaps with the single membrane spanning region from 700–723 (blue lettering and red block). The yellow block indicates a signal sequence (1–17) whilst the cyan block indicates the location of a domain showing homology to BPTI or Kunitz type inhibitors (291–341). The magenta blocks highlight poly-glutamate rich domains 230–260, threonine rich regions 265–282, heparin binding regions 391–424 and 491–522 and collagen binding domain 523–541

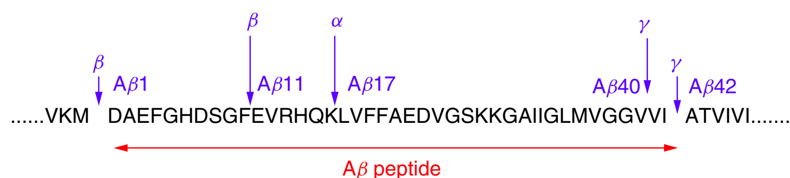

Figure 12.37 Schematic diagram of the Aβ peptide derived from the amyloid precursor protein. The amino acid sequence is shown with single letter codes and the α, β and γ secretase cleavage sites are shown along the polypeptide chain. The α secretase cleaves at the 17th residue whilst the β secretase cuts at the 1st and more rarely 11th residues and the γ secretase cuts at the 40th or 42nd residue

peptide designated as $A\beta_{1-40}$ and $A\beta_{1-42}$ respectively. Sequencing and database analysis showed that the peptides derived from a longer protein expressed in many tissues – the amyloid precursor protein (APP)- and DNA sequencing indicated a protein of ~770 residues with multiple domains (Figure 12.36).

APP undergoes extensive post-translational modification as a glycoprotein but despite domain recognition a clearly defined function for APP is absent although a role in cell adhesion is currently favoured. The APP gene was mapped to chromosome 21, contained 19 exons, with alternative splicing mechanisms giving rise to multiple transcripts (APP395, APP563, APP695, APP751, APP770). APP695 is the major isoform in neuronal tissue with APP751 dominating elsewhere.

The Aβ amyloid peptide (residues 672-711/713) is derived from a region of the protein that includes a small extracellular region together with its membrane spanning sequence (700–723). Understanding the origin of the Aβ peptide lies at the heart of discovering the molecular pathology of Alzheimer's disease. The amyloid peptide located *within* the full length protein arises from endoproteolysis of APP (Figure 12.37). The amyloid peptide was formed by cutting at residues 670 and 711/713 at sites associated with enzymes called β and γ secretase whilst the enzyme α secretase cuts APP at residue 687 within the amyloidogenic peptide. The properties and identity of these enzymes were initially obscure (hence their names) but slowly each system has been shown to occupy an important role in APP processing.

The demonstration that endoproteolysis of APP was associated with two different enzymes suggested that defining these 'activities' in more detail as well as the associated genes/proteins would assist understanding the molecular events contributing to disease. β secretase was identified as a new member of the aspartic protease family with the unusual property of being membrane tethered. The α secretase is a member of the ADAM family of proteins (standing for a disintegrin and metalloprotease) and finally γ secretase began to give up its secrets by suggesting similarity to a membrane protein called presenillin 1.

Although initially called β secretase the enzyme purified by different groups was given alternative names of BACE (β-site APP cleaving enzyme) and memapsin-2 (membrane-associated aspartic protease 2) and both are widely quoted in the literature. The enzyme, a single polypeptide chain of 501

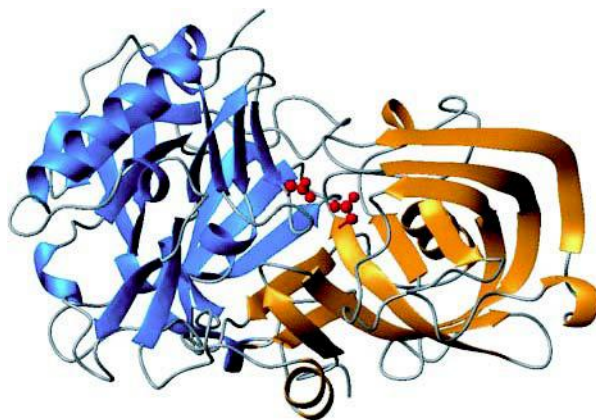

Figure 12.38 The organization of secondary structure into two lobes and the active site aspartate residues in β secretase (PDB:1M4H) shows similarities to previous aspartyl proteases

residues synthesized as a preproprotein, contains a single transmembrane region from residues 458 to 478 followed by a short cytoplasmic facing region of ~20 residues. The majority of residues are located in a large extracellular domain from residues 45–457 and this region of the enzyme has been studied to provide information on kinetics, substrate specificity and structure (Figure 12.38).

As an aspartyl protease the enzyme contains two conserved aspartate within a bilobal structure and exhibits an optimal pH for catalytic activity at pH 4.0 with substrates bound with K_m values between 1–10 mM and turnover rates ranging from 0.5–3 s^{-1}. The active site of β secretase is more open and less hydrophobic than corresponding sites in other human aspartic proteases but by defining the active site of this enzyme through substrate specificity and transition state analogues offers the hope of designing effective inhibitors. Inhibitors of β secretase are important pharmaceutical drug targets since they would prevent amyloid peptide formation.

The genetic basis of Alzheimer's disease has assisted understanding of disease pathology. Missense mutations are present in the APP gene in familial Alzheimer's disease and these mutations are frequently found in close proximity to the α-, β-, or γ secretase processing sites. In a Swedish pedigree a double mutation Lys → Asn and Met → Leu near the β-secretase cleavage site (671–672) leads to elevated production of the Aβ peptide (both Aβ40 and Aβ42) and therefore presumably accounts for the early onset of the disease. Defective genes are also responsible for early onset disease with mutations to presenilin 1 and presenilin 2 observed in families and leading to increased deposition of the Aβ peptide. Over 70 mutations are known most located in presenilin 1.

Despite a lack of a clearly defined function for APP the Aβ peptide aggregates into larger amyloid fibrils, destroying nerve cells and decrease levels of neurotransmitters. The correct balance of neurotransmitters is critical to the brain's activity and the levels of acetylcholine, serotonin, and noradrenaline are all altered in Alzheimers disease. With both structural defects and chemical imbalance in the brain Alzheimer's disease effectively disconnects areas of neuronal tissues that normally work cohesively together.

An interesting question is why have neurodegenerative diseases become more prominent in recent times? Undoubtedly part of the answer lies with longer life expectancy. In 1900 it is estimated that only 1 % of the world's population was above 65 years of age. By 1992 that figure had risen to nearly 7 % and in 2050 it is estimated that ~25 % of the population will be >65 years of age. The introduction of antibiotics, improved living conditions particularly in relation to water quality and food supply coupled with effective health care has meant that at least in the economically advantaged regions of the world average lifetimes for men and women are well above 80 years of age. One consequence of longer life expectancy is that diseases such as neurodegenerative disorders can develop over many years where previously an individual's short lifetime would preclude their appearance. Susceptibility to neurodegenerative diseases may be an inevitable consequence associated with longevity but as molecular details concerning these disease states are uncovered there is optimism about favourable treatment options in the future.

p53 and its role in cancer

Cancer involves the uncontrolled growth of cells that have lost their normal regulated cell cycle function. In many instances the origin of diseases lie in the mutation of genes controlling cell growth and division so that cancer is accurately described as a cell cycle based disease. Mutations alter the normal function of the cell cycle, with three classes of genes identified closely with cancer. Oncogenes ‘push’ the cell cycle forward, promoting or exacerbating the effect of gene mutation. In contrast, tumour suppressor genes function by normally applying the ‘brakes’ to cell growth and division. Mutation of tumour suppressor genes causes the cell to lose their ability to control coordinated growth. Finally, repair genes keep DNA intact by preserving their unique sequence. During the cell cycle mutagenic events result in DNA modification that in most circumstances are repaired unless the mutation resides in genes coding for repair enzymes.

One of the most important systems in the area of cancer and cell cycle control is the protein p53, and details of the action of this protein were largely the result of the research by Bert Vogelstein, David Lane

and Arnold Levine. The p53 gene is located on the short arm of chromosome 17 where the open reading frame codes for a protein of 393 residues via 11 exons. The protein identified before the gene was originally characterized through co-purification with the large T antigen in SV40 virus-transformed cells and had a mass of ∼53 kDa. Since the presence of the protein appeared to correlate with viral transformation of cells p53 was originally labelled as an oncogene. By the late 1980s it was clear that the gene product of cloned p53 was a mutant form of the protein and that normal (wild-type) p53 was the product of a tumour suppressor gene. Further modification of ideas concerning p53 function occurred with the demonstration that the protein caused G_1 cell cycle arrest but also activated genes by binding to specific DNA sequences – it was a transcription factor. The discovery that one transcriptional target of p53 was an inhibitor (p21) of cyclin dependent kinases (Cdk) provided a direct link to the cell cycle since p21 complex formation with cdk2 prevents cell division and represents one point of control. Mutant p53 does not bind effectively to DNA with the consequence that p21 is not produced and is unavailable to act as the 'stop signal' for cell division.

p53 has been called the 'Guardian of the Genome' and it is normally present in cells at low concentration in an inactive state or one of intrinsically low activity. When the DNA of cells is damaged p53 levels rise from their normal low levels and the protein is switched 'on' to play a role in cell cycle arrest, transcription and apoptosis. p53 regulates the cell cycle as a transcription factor so that when damage to DNA is detected cell cycle arrest occurs until the DNA can be repaired. Once repaired the normal cell cycle can occur but this process serves to ensure that damaged DNA is not replicated during mitosis and is not 'passed' on to daughter cells. If repair cannot take place apoptosis occurs (Figure 12.39).

With an important role within cells the structural organization of p53 was of considerable interest to many different fields ranging from cell biology to molecular medicine. The protein contains a single polypeptide chain that is divided into three discrete domains (Figure 12.40). These domains present a modular structure that has facilitated their individual study in the absence of the rest of the protein and are concerned with transcriptional activation, DNA binding and tetramerization. An N-terminal transactivation domain extends from residues 1–99 and is followed by the largest domain–a central core region from residues 100–300 that binds specific DNA sequences. The third and final region of p53 is a C-terminal domain (residues 301–393) that includes both a tetramerization domain (from residues 325–356) and a regulatory region (363–393) (Figure 12.41). As a result of the two functional activities in this region p53 is sometimes described as containing four as opposed to three composite domains.

The presence of discrete domains linked by flexible linkers creates a conformationally dynamic molecule and p53 has proved difficult to crystallize as the

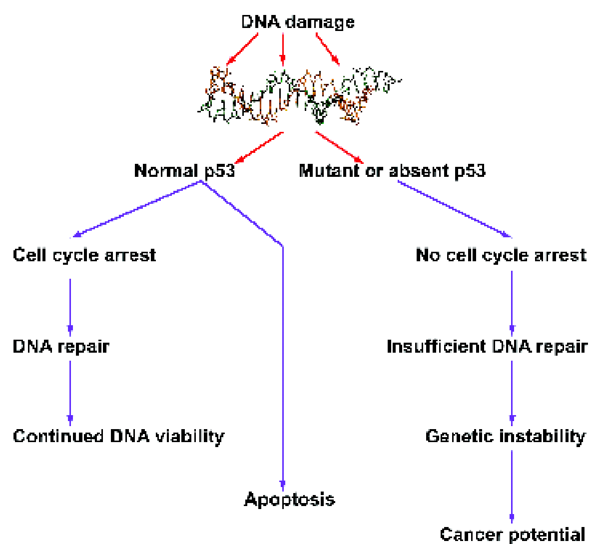

Figure 12.39 An outline of possible p53 actions with normal and damaged cells

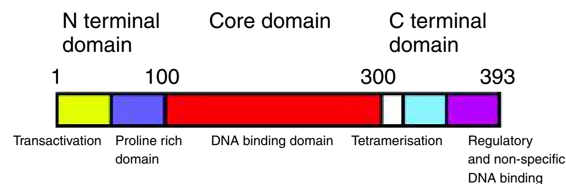

Figure 12.40 The organization of domains within the tumour suppressor p53

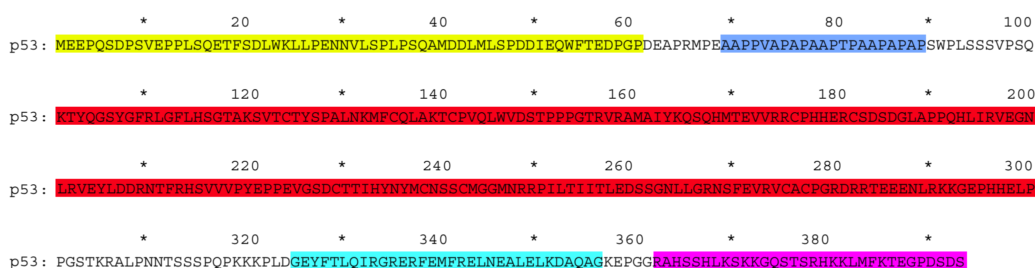

Figure 12.41 The primary sequence of p53 showing the partition of the primary sequence into discrete domains, using the colour codes of Figure 12.40

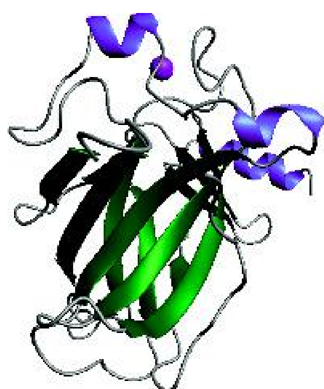

Figure 12.42 The structure of the isolated core domain of p53 showing β sandwich structure together with a bound Zn ion (purple)

full length molecule. This almost certainly results from molecular heterogeneity and structural insight into p53 followed a modular approach determining the structure of isolated domains. One incisive set of results were obtained by Nikolai Pavletich using proteolytic digestion of the full-length protein to yield a protease-resistant fragment of ∼190 residues (residues 102–292) that corresponded to the central portion of p53 involved in sequence-specific DNA-binding. DNA binding was inhibited by metal chelating agents and unsurprisingly within the core domain a Zn ion was identified (Figure 12.42). The central core extending from residues 100–300 contains the DNA binding residues and also those residues representing mutational hot spots (see below) within the p53 gene.

As a DNA binding protein the core domain binds to both the major and minor groove of the double helix through the presence of charged side chains arranged precisely in conserved loops that link elements of β strands (Figures 12.43 and 12.44). Numerous arginine and lysine side chains form specific interactions with DNA. Most DNA binding proteins are based on the helix-turn-helix (HTH), HLH or leuzine zipper motifs, but p53 is essentially a β sandwich formed by the interaction of antiparallel four- and five-stranded elements of β sheet.

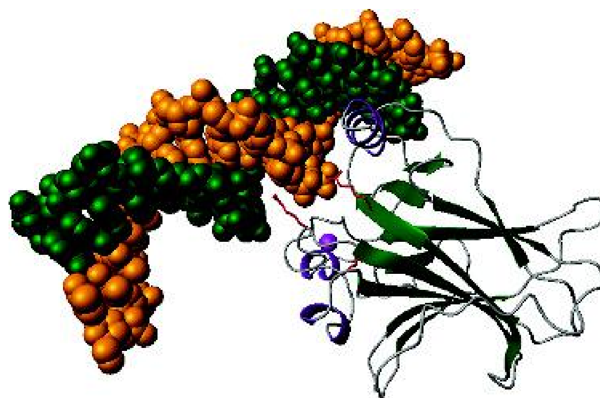

Figure 12.43 The structure of p53 core domain bound to DNA showing role of Arg175, Arg273 and Gly245 (red wireframe) in these interactions as well as the position of a bound Zn ion. The Zn ion is bound in a tetrahedral environment formed by the side chains of three cysteine residues together with a single histidine arranged as part of two loop regions and a single helix

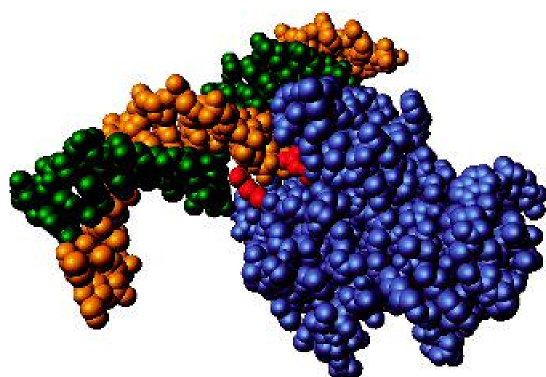

Figure 12.44 A spacefilling representation of p53-DNA complex showing precise interaction with major and minor grooves of helix (PDB:1TUP)

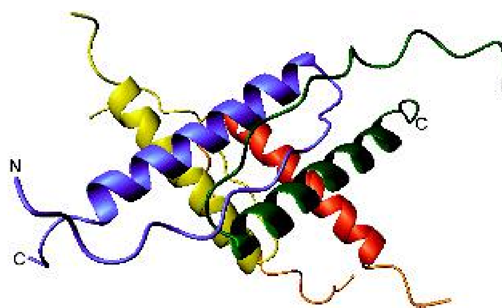

Figure 12.45 The structure of the isolated tetrameric region of p53. The blue and green helices pair together as one dimer with the yellow and orange helices forming the second dimer. The N- and C-terminals are indicated, and point in the case of the N-terminal domain towards the DNA binding core with the C-terminal region pointing towards the remaining parts of p53. The extended region before each helix resembles a β strand and two neighbouring strands on adjacent monomers also interact

The N-terminal domain consists of two contiguous transcriptional–activation regions (residues 1–42 and 43–63) together with an adjacent proline/alanine rich region (residues 62–91).

p53 binds to DNA with high affinity when assembled into a tetrameric complex. This reaction is directed by a tetramerization domain formed by the region between residues 320 and 360 in each polypeptide chain. It is joined to the previously described DNA binding domain via a long flexible linker. The structure of this domain determined in isolation as four 42 residue peptides consists of a dimer of dimers with a combined mass of 20 kDa. Each monomer consists of a regular helix (335–353) and an extended region resembling a poorly defined β strand conformation (residues 326–333) with the two elements linked via a tight turn based around Gly334. An antiparallel interaction of helices in each monomer promotes the formation of a pair of dimers. The interface between the pair of dimers is mediated solely by helix–helix contacts with the overall result being a symmetric, four-helix bundle (Figure 12.45). It has proved possible to mutate the tetramerization domain to favour dimer formation over that of the tetramer. Extensive mutagenesis combined with folding and stability studies suggest that two residues Leu340 and Leu348 promote tetramer formation via interactions of non-polar side chains.

A total of nine residues were identified as key residues in governing stability: Phe328, Ile332, Arg337, Phe338, Met340, Phe341, Leu344, Ala347 and Leu348. All of these residues are invariant in mammalian p53s with the exception of the minor transition Phe338Tyr, and point to important roles for these residues.

The realization that anomalous p53 activity is a common denominator in many human cancers was a major advance. In approximately 50 percent of tumours p53 is inactivated as a result of mutations within the p53 gene (Table 12.4). In other tumours the inactivation of p53 occurs indirectly through binding of viral proteins or through mutations in genes encoding proteins that interact with p53. These mutations alter the flow of information between p53 and the rest of the cellular apparatus.

Most p53 mutations occur in somatic cells and therefore only induce disease in the individuals carrying the defective gene. Mutations in germ line cells (egg and sperm) affect succeeding generations and for p53 an inherited defect known as Li–Fraumeni syndrome has been identified where the gene encoding p53 possesses a mutation. The mutation of p53 is associated with soft-tissue sarcomas and osteosarcomas (bone) where individuals with this inherited condition develop many different cancers from a young age. Since mutations in

Table 12.4 p53 malfunctions that may lead to human cancers

Mechanism of p53 inactivation	Typical tumours	Effect of inactivation
Mutation of residues in DNA binding domain	Colon, breast, lung, bladder, brain, pancreas, stomach, oesophagus and many others.	Prevents p53 binding to specific target DNA sequences and activating adjacent genes.
Deletion of C terminal domain	Occasional tumours at many different sites	Prevents the formation of p53 tetramers.
mdm2 gene multiplication	Sarcomas, brain	Enhances p53 degradation.
Viral infection	Cervix, liver, lymphomas	Products of viral oncogenes bind to p53 causing inactivation and enhanced degradation.
Deletion of $p14^{ARF}$ gene.	Breast, brain, lung and others especially when p53 is not mutated.	Failure to inhibit mdm2 with the result that p53 degradation is uncontrolled
Mislocalisation of p53 to cytoplasm.	Breast, neuroblastomas	p53 functions only in the nucleus.

Adapted from Vogelstein, B. *et al. Nature* 2000, **408**, 307–310.

p53 create genetic instability by the inability to repair damaged DNA or a failure to perform apoptosis defective DNA accumulates making individuals with the Li–Fraumeni syndrome very susceptible to cancer.

Of the mutations identified within p53 almost all are located in DNA binding and tetramerization domains and a vast array of different mutations have been identified. Databases exist where mutations of p53 found within tumours have been categorized; over 14 000 tumour-associated mutations of p53 have been described as of 2003. Within this vast panoply of mutations exist 'hot spots' along the gene. Three mutational hot spots are residues 175 (Arg), 248 (Arg) and 273 (Arg) of the protein. A distribution of all mutations identified to date suggest approximately 20 percent of mutations involve these three residues with residue 248 being the site of highest mutation (~7.5 percent). Along with a further three residues (245/Gly, 249/Arg and 282/Arg) these residues will account for 30 percent of all p53 mutations found in tumours.

If humans have an in-built tumour suppressor gene such as p53 why is cancer so common? Ignoring inherited conditions the answer lies in the influence of environmental conditions on p53. Mutagens alter the DNA sequence of p53 leads to base changes that are eventually reflected in an altered and functionally less active protein. One potent example of this environmental influence is the effect of cigarette smoking on p53. There is no doubt today about epidemiological data showing the overwhelmingly strong correlation between lung cancer and cigarette smoking, yet despite emphatic data a molecular basis for cancer development has taken longer to establish. One poisonous ingredient of cigarette smoke is a class of organic molecules known as benzopyrenes. Exposure of the p53 gene sequence to these mutagens caused alterations in base sequence most frequently the substitution of G by T. These mutagens covalently cross-link with DNA at the G residues found at codons 175, 248 and 273. Since p53 is a tetramer the mutation

of just one subunit perturbs the activity throughout the complex resulting in impaired DNA binding.

Another mechanism for cancer lies in the inactivation of p53 by the human adenovirus and papilloma virus. These viruses bind to p53 inactivating the protein function and preventing a successful repair of damaged DNA. At residue 72 of p53 two alleles exist with either arginine or proline found in the wild type protein. Initial studies suggested these variants were functionally equivalent but current evidence suggests the Arg72 variant is more susceptible to viral mediated proteolysis. For the human papillomavirus (HPV18) – a major factor in cervical carcinoma – this involves the preferential interaction of Arg72 p53 with the E6 protein of the virus *in vitro* and *in vivo*. From these studies it was suggested that cancer of the cervix is approximately seven times higher in individuals homozygous for the Arg72 allele.

Other cancers involve amplification of mdm2 – a protein that inactivates p53 by binding to the transcriptional domain. High levels of mdm2 cause a permanent inactivation of p53 and an increased likelihood that mutations will be passed on to daughter cells. In view of its role 'controlling' p53 mdm2 is widely believed to represent the next avenue for intensive cancer research.

The amount of information existing on all aspects of p53 function and mutant protein expression in human cancers is vast and reflects a key role in pathogenesis. p53 is just one component of a network of events that culminate in tumour formation. Although inactivation of p53 either through direct or indirect mechanisms is most probably responsible for the majority of human cancers it must be remembered that genes other than p53 are targets for carcinogens. The identification and subsequent characterization of these genes/proteins will increase the molecular approach to the treatment of cancer.

Emphysema and α_1-antitrypsin

Emphysema is a genetic, as well as an acquired disease, characterized by over-inflation of structures in the lungs known as alveoli or air sacs. This over-inflation results from a breakdown of the walls of the alveoli and this results in decreased respiratory function. As a result breathlessness is a common sign associated with emphysema, and together with chronic bronchitis, emphysema comprises a condition known as chronic obstructive pulmonary disease (COPD). Frequently COPD is associated with lung disease arising from smoking but in some instances emphysema has been identified as a genetically inherited condition.

Emphysema is defined by destruction of airways distal to the terminal bronchiole with destruction of alveolar septae and the pulmonary capillary region leading to decreased ability to oxygenate blood. The destruction of the alveolar sacs is progressive and becomes worse with time. The body compensates with lowered cardiac output and hyperventilation. The low cardiac output leads to the body suffering from tissue hypoxia and pulmonary cachexia. A common occurrence is for sufferers to develop muscle wasting and weight loss and they are often identified as 'pink puffers'. In the United States it is estimated that two-thirds of all men and one-quarter of all women have evidence of emphysema at death with total sufferers approaching $\sim$2 million. As part of COPD it is responsible for the fourth leading cause of death in the United States and represents a major demand on health service care.

As a genetic condition emphysema has been attributed to a deficiency in a protein called α_1-antitrypsin. Unlike the common form of emphysema seen in individuals who have smoked for many years this form of the disease is often labelled as α_1-antitrypsin deficiency form of emphysema. The disease can occur at an unusually young age (before 40) after minimal or no exposure to tobacco smoke and has been shown to arise in families. The latter observation pointed to a genetic link.

The biological role of α_1-antitrypsin is as an enzyme inhibitor of serine proteases and despite its name its main target for inhibition is the enzyme elastase. It is a member of a group of proteins known as serpins or serine protease inhibitors. Collectively these proteins inhibit enzymes by mimicking the natural substrate and binding to active sites. Serpins have been identified in a wide variety of viruses, plants, and animals with more than 70 different serpins identified. Members of the serpin family have very important roles in the inflammatory, coagulation, fibrinolytic and complement cascades. Consequently malfunctioning

serpins have been shown to lead to several forms of disease including cirrhosis, thrombosis, angio-oedema, emphysema as described here and dementia.

The site of action of α_1-antitrypsin is the serine protease elastase. Elastase is commonly found in neutrophils and is structurally homologous to trypsin or chymotrypsin (see Chapter 7). Neutrophils represent the primary defence mechanism against bacteria by phagocytosis of pathogens followed by degradation and in this latter reaction elastase is involved. At sites of inflammation hydrolytic enzymes of neutrophil origin are detected and this includes the enzyme elastase.

α_1-antitrypsin is the major serine protease inhibitor present in mammalian serum. It is synthesized in the liver and represents the major protein found in the α_1-globulin component of plasma. In the serum the concentration of α_1-antitrypsin is very high ($\sim$2 g l^{-1}) and this reflects its important role. A deficiency in certain individuals of α_1-antitrypsin was identified in the 1960s by detailed analysis of gel electrophoretic profiles but as with most diseases attention has recently been directed towards identifying the gene and its location within the human genome.

The gene is called the PI (protease inhibitor) gene and is located on the long arm of chromosome 14. Genetic screening of individuals presenting with early onset emphysema has allowed many mutations to be identified. The normal allele is the M allele and does not result in the disease. Genetic analysis of the M allele has shown that several different forms designated as M1, M2, M3, etc. can be distinguished with M1 representing approximately $\sim$70 percent of the population and distinguished by two variants; Val213 and Ala213. These two forms represent nearly all of the M1 phenotype and occur in a ratio of $\sim$2:1. Alternative alleles are attributed to an increased incidence of disease in individuals possessing one or more copies and the two most common are the S and Z forms. The Z allele is most frequently associated with codon 342 and results in the non-conservative substitution of glutamate by lysine as a result of a single base change from GAG to AAG. This mutation is estimated to be present in the north European population at $\sim$1 in 60. Individuals homozygous for the Z allele (often written as PI ZZ) are at greatly increased risk of emphysema *and* liver disease. The second form is the S allele and is associated with the substitution of glutamate by valine at codon 264. Individuals homozygous for the S allele (PI SS) do not display symptoms of the disease but individuals with one copy of the Z allele and one copy of the S allele (PI SZ) may develop emphysema. Over 70 different mutations have been identified and it appears that these mutations can be grouped into three major categories. Some mutations disrupt the serpin function of the protein, whilst rarer mutations are described as null or silent mutations and a third group appear to cause drastic reduction in the amount of α_1-antitrypsin present in the serum as opposed to its activity.

The single polypeptide chain of α_1-antitrypsin contains 394 amino acid residues after removal of a 24-residue signal sequence. The presence of three glycosylation sites leads to higher than expected molecular weight (M_r $\sim$52 000) when serum fractions are analysed by gel electrophoresis but the carbohydrate is not necessary for catalytic function. Within the polypeptide chain the sequence Pro357-Met358-Ser359-Ile360 represents a 'substrate' for target protease. This region of the polypeptide effectively forms a 'bait' for the cognate protease, in this case leucocyte elastase, and the reaction between serpin and protease leads to complete inactivation.

The structure of α_1-antitrypsin reveals three sets of β sheets designated A–C and nine helical elements. The basis of inhibition revolves around interaction between cognate serine protease and serpin by the presentation of a region of the serpin polypeptide chain that resembles natural substrates of the protease. Binding occurs at an exposed region of the polypeptide chain known as the reactive centre loop a region held at the apex of the protein and based around the critical Met358 residue (Figure 12.46). The mobile reactive centre loop presents a peptide sequence as the substrate for the target protease. Docking of the loop of the serpin with the protease is accompanied by cleavage at the P_1–P'_1 bond and the enzyme is locked into an irreversible and extremely stable complex. In the case of α_1-antitrypsin the P_1–P'_1 bond (i.e. the residues either side of the bond to be cleaved) lies between Met358 and Ser359 in the loop located at the apex of the protein.

Upon protease binding large conformational rearrangement occurs in which the two major structural transitions occur. The reactive centre loop is split between Met358 and Ser359 and the A_β sheet is able to

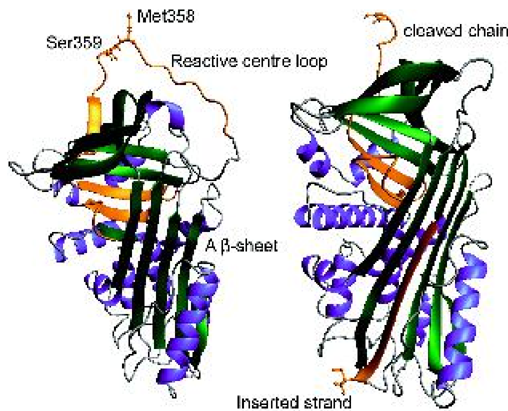

Figure 12.46 Two conformational states of α_1-antitrypsin showing conformational changes upon cleavage. Residues 350–394 are coloured orange to distinguish from the remainder of the molecule where the strands are shown in green and the helices in blue (PDB:2 PSI and 7API)

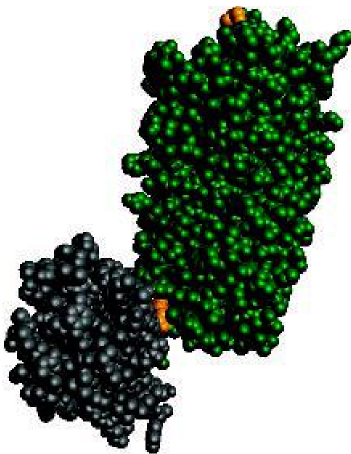

Figure 12.47 Space filling model of interaction between serpin and protease. Met358 and Ser359 can be seen at either end of the serpin (green)

open to accommodate much of the reactive centre loop region as a fifth β strand. This conformational change is vital for efficient inhibition and leads to Met358 being located at the foot of the molecule (Figure 12.47). In this conformation the α_1-antitrypsin and elastase are irreversibly bound in a stable complex that prevents further reactivity. Conformational mobility of these regions results in sensitivity to mutation and significantly the Z allele of α_1-antitrypsin is located within the reactive loop region (effectively as P'_4).

In many individuals with α_1-antitrypsin deficiency liver disease is detected as an associated condition. An explanation for this association has now been found. In the homozygous ZZ condition only about 15 percent of the normal level of α_1-antitrypsin is secreted into the plasma. The normal site of synthesis is the liver and it was observed that the remaining 85 percent accumulates in the ER of the hepatocyte. Although much is degraded the remainder aggregates forming insoluble intracellular inclusions. These aggregates are responsible for liver injury in approximately 12–15 percent of all individuals with this genotype. It is also observed that about 10 percent of newborn ZZ homozygotes develop liver disease leading to fatal childhood cirrhosis. Studies based on the structure of α_1-antitrypsin have demonstrated the molecular pathology underlying protein accumulation.

The principal Z mutation in antitrypsin results in the substitution of glutamate by lysine at residue 362 and allows a unique molecular interaction between the reactive centre loop of one molecule and the ‘gap’ in the A-sheet of another serpin. Effectively the reactive centre loop of one α_1-antitrypsin inserts into the A_β sheet region of a second molecule in a process known as loop to sheet polymerization. In mutant α_1-antitrypsin this occurs favourably at 37 °C and is very dependent on concentration and temperature. The polymerization is blocked by the insertion of a specific peptide into the A-sheet of α_1-antitrypsin.

α_1-antitrypsin is an acute phase protein and for mutant protein it will undergo a significant increase in association with temperature as might arise during bouts of inflammation. The result is harmful protein inclusions in the hepatocyte that may overwhelm the cell’s degradative mechanisms, especially in the newborn. The control of inflammation and pyrexia in ZZ homozygote infants is very important to overall management of the disease. The serpins show considerable homology and the mechanism of loop–sheet polymerization is the basis of deficiencies associated also with mutations of C1-inhibitor (the activated first component

of the complement system), antithrombin III, and α_1-antichymotrypsin implicated in other disease states. Collectively, these diseases could be called serpinopathies.

It has been shown that replacement of deficient protein by intravenous infusion of purified α_1-antitrypsin can limit disease progression. Unlike many of the previously described conditions this treatment represents an effective protocol that limits much of the necrosis of lung tissue associated with hereditary forms of emphysema. This form of emphysema is currently treated with protein isolated from plasma fractions, but with characterization of the cDNA encoding the protein α_1-antitrypsin has been expressed to allow high level production in the milk of transgenic sheep. Although this form of α_1-antitrypsin has yet to complete clinical trials it will offer a better alternative to conventionally purified protein. The underlying genetic condition will always remain but considerable improvement results from treatment with purified α_1-antitrypsin. The high concentration of α_1-antitrypsin in serum means that on average a patient may require over 200 g of highly purified proteins per year as the basis of a therapeutic regime.

For individuals with this genetic condition a lifetime avoidance of smoking is highly recommended since the effects of smoking as a primary cause of emphysema are well documented. In this context it has been shown that one effect of smoking can be to cause oxidation of the Met358 residue within α_1-antitrypsin. The oxidation of the methionine side chain to form a sulfoxide results in permanent irreversible damage to α_1-antitrypsin and eliminates its biological role as a serpin. A loss of inhibition allows unchecked elastase activity and promotes alveolar sac destruction that lead to emphysema. When combined with a genetic tendency towards emphysema it is clear that this situation must be avoided to prevent extreme lung damage. Genetically linked forms of emphysema represented one of the first molecular defects to be extensively characterized and the highly detailed structural and functional studies of serpins such as α_1-antitrypsin highlight the diagnostic, preventative and therapeutic powers of molecular medicine.

Summary

Molecular medicine represents a multidisciplinary approach to understanding human disease through the use of structural analysis, pharmacology, gene technology and even gene therapy. In 1955 sickle cell anaemia was identified as arising from a single amino acid change within haemoglobin. Many other diseases arise from similar genetic mutations.

Unprecedented progress has been made in understanding the molecular basis of many diseases including new, emerging, ones such as HIV and vCJD alongside familiar conditions such as influenza. Understanding at a molecular level offers the prospect of better disease diagnosis and in some cases improved therapeutic intervention.

Viruses represent a persistent human health problem, being responsible for many diseases. Vaccination using attenuated viruses or purified proteins effectively eradicated infections such as smallpox and polio but threats from rapidly mutating viruses such as influenza and HIV continue.

HIV continues to contribute to significant, worldwide, mortality in the 21st century whilst influenza returns periodically in the form of severe outbreaks or pandemics. In each virus the ability to mutate surface antigens rapidly compromises therapeutic drug action or antibody directed immune response.

The protein p53 has been called the 'Guardian of the genome'. Its major role is to prevent the perpetuation of damaged DNA by causing cell cycle arrest or apoptosis. The protein is a conformationally flexible molecule consisting of an N-terminal transcriptional activation domain, a central or core DNA binding domain and a third C terminal domain that contains multiple functions including the ability to promote tetramerization.

The core DNA binding domain is based on a β sheet scaffold that binds to major and minor grooves of the double helix through the presence of charged side chains arranged precisely in conserved loops that link elements of β strands.

Cancer and p53 are inextricably linked through the demonstration that many tumours have mutations in p53. Most, but not all, are located in the central DNA binding domain with mutational hot spots involving residues participating in DNA binding. The cell requires exquisite control pathways to modulate the activity of p53 and corruption of these pathways can often lead to the development of cancer.

α_1-antitrypsin is a serpin that acts to limit the unregulated activity of neutrophil elastase. The protein inhibits elastase by forming a tightly bound complex. It

is found at high concentrations in the globulin fraction of serum where its abundance helped to identify individuals deficient in serum α_1-antitrypsin.

Deficiency arises from a gene mutation and is an inherited condition leading to emphysema as a result of damage caused to the alveolar lining of the lungs. Mutant α_1-antitrypsin fails to bind elastase effectively.

Mutated forms of α_1-antitrypsin show a tendency to exhibit loop–sheet polymerization – a process that causes the irreversible association of serpins.

Problems

1. Why should individuals with sickle cell anemia avoid intense exercise or exposure to high altitudes?
2. Vaccines have been developed against smallpox, measles and many other viral based diseases. So what is the problem with developing a vaccine for flu and HIV?
3. The ionization of maleic acid (COOH–CH_2–CHOH–COOH) occurs with pK values of 1.9 and 6.2. Discuss the reasons for these values and what is the relevance to the activity of aspartic proteases.
4. Using the known properties of the active site of neuraminidase describe why Zanamivir is an efficient inhibitor of enzyme activity especially when compared with Neu5Acen. Can you suggest potentially other beneficial drug design.
5. Is vCJD a new disease or merely an old one firmly recognized and identified?
6. Identify the conformationally sensitive strands to inhibitor binding and active site aspartates in HIV protease.