Cancer Molecular Biology

Graeme K. Ambler

June 14, 2006

1 Introduction

This is an attempt to synthesise the various parts of the Lent term MIMS lectures into a coherent whole. Information from the prologue lectures has thus as much as possible been put in the context of the other lectures rather than having a separate section for it. Before we can really discuss cancer we must first introduce the biochemical basis of inheritance, and thus we begin by discussing some of the content of Steve Bell's lectures. This is not an exhaustive treatment of our lecture course. Topics not covered include the experimental stuff: PCR (SB), DNA sequencing methodology (SB), footprinting (NS13) and chromatin immunoprecipitation (NS13).

2 Organisation and structure of DNA

2.1 Basics

Although DNA was discovered in 1868 by Friedrich Mieschner, it was not until 1944 that it was proved to be the bearer of genetic inheritance. The now famous experiment was carried out by Avery, MacLeod and McCarty. They extracted and purified DNA from heat-killed virulent bacteria (*Streptococcus pneumoniae*) and added it to a sample of non-virulent bacteria. The non-virulent bacteria were transformed by this procedure and capable of inducing infection in mice.

In the late 1940s the primary structure of DNA was solved, with the discovery that it was a polymer consisting of a sugar-phosphate backbone with nitrogenous bases protruding from the backbone, of which there were four varieties: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The sugar in DNA is deoxyribose. There is a slightly confusing nomenclature for the monomers of DNA (or RNA).

- The bases themselves are grouped into two subclasses. Adenine and Guanine are called **purines**, while Thymine and Cytosine are called **pyrimidines**. Purines have a double ring structure, while pyrimidines have only a single ring.
- A nitrogenous base together with a pentose (5 carbon) sugar is called a **nucleoside**. The four nucleosides present in DNA are called Adenosine (A), Thymidine (T), Guanosine (G) and Cytidine (C).

- A nitrogenous base together with a pentose sugar, to which is attached one or more phosphate groups is called a **nucleotide**. Thus adenosine triphosphate (ATP) could be called either a *nucleoside triphosphate* or simply a *nucleotide*.

While we are on the subject of nucleic acid, we may as well talk about RNA. There are two key differences between DNA and RNA.

1. The sugar in RNA is ribose not deoxyribose.

2. Thymine is not present in RNA. The pyrimidine uracil is used in its place.

The sugar in ATP and GTP is ribose. The nucleotides which polymerise to form DNA are referred to as dATP, dGTP, etc.

Returning now to history, Chargaff discovered in the late 1940s that the number of Thymidine residues in DNA equalled the number of Adenosine residues and the number of Guanosine residues equalled the number of Cytidine residues. This, together with Rosalind Franklin's X-ray diffraction work showing that DNA formed a helix led Watson and Crick to propose their now famous model for the three-dimensional structure of DNA. Here is a picture of a flattened-out section of the double stranded helix they proposed.

$$
\begin{array}{rr}
| & | \\
P & P \\
| & | \\
dRb - A\cdots T - dRb \\
| & | \\
P & P \\
| & | \\
dRb - C\cdots G - dRb \\
| & | \\
P & P \\
| & | \\
dRb - G\cdots C - dRb \\
| & | \\
P & P \\
| & |
\end{array}
$$

The Ps in the picture represent phosphate groups; dRb represents deoxyribose; A, G, C and T represent bases; solid lines represent covalent bonds and dotted lines represent hydrogen bonds. Note that A only ever pairs with T and G with C in this structure, which explains Chargaff's findings. There are actually three hydrogen bonds between G and C and only two between A and T, so that stretches of DNA rich in G and C bind much more tightly than stretches which are rich in A and T. The ribbon shown in the picture on the left is coiled into a right-handed helix 20Å wide with a periodicity of 36Å corresponding to 10.5 base pairs. The strands actually run antiparallel to each other, and I thought of putting the letters on one of the two strands upside down to convey this idea, but couldn't be bothered to figure out how to do it.

The helix has a large major and small minor groove. Although the bases are on the inside of the helix, some of their atoms are exposed, thus allowing sequence-specific interactions with other molecules (in particular proteins). The phosphate groups in the backbone are negatively charged at physiological pH, thus repelling each other.

There are several forms of DNA found in nature, though by far the most prevalent is DNA-B, described above. Others include DNA-A, another right-handed helix, and DNA-Z, a left-handed helix. Certain DNA sequences can also adopt more exotic structures such as hairpins or triple or quadruple helices.

2.2 Supercoiling

The 23 chromosomes of the human genome have a combined length of nearly 1 metre. Since we are a diploid species this means that we have nearly 2 metres of DNA inside the

nucleus of each of the cells of our body (with some exceptions like erythrocytes which have no nucleus and gametes which are haploid). Since the nucleus of the average human cell is only about 10 microns across this is a fantastic degree of compaction. There are several ways in which this is achieved.

1. Negative supercoiling, where underwinding of the double helix causes DNA to wrap around itself rather like the tangled cable on an old-fashioned telephone.

2. Wrapping around proteins such as IHF in bacteria, which introduces 180° bends in the space of just 20 base pairs. In eukaryotes DNA wraps round **histone** octomers made of two copies each of histones H2A, H2B, H3 and H4. 150 base pairs of DNA wraps around one of these particles 1.65 times forming a **nucleosome core particle**. The remaining histone, H1, binds core particles together to introduce further compaction. This generates 30nm helices with about 6 nucleosome core particles per turn. Histone proteins are positively charged at physiological pH (they are basic) and thus help to neutralise the negatively charged phosphate backbone enabling tighter packing.

3. 30nm fibres are attached in loops to a proteinacious scaffold in the nucleus to provide the final level of compaction.

The combination of DNA and histone proteins in eukaryotes is called **chromatin**.

3 DNA replication

DNA replication proceeds from an origin of replication (unique in bacteria, multiple on each chromosome in eukaryotes) by using each strand as a template, resulting in semiconservative duplication (each new double helix is formed from one original strand and one new strand). Replication also proceeds in both directions from the origin of replication, so that there is effectively four lots of copying initiated by a single origin of replication, cutting the time required to replicate a given strand by half. Replication only takes place in the 5' to 3' direction (on a 3' to 5' template), meaning that one of the strands replicated is done in short stretches in the direction opposite to the direction which replication moves along the polymer. These short stretches are called Okazaki fragments. In bacteria they are usually around 1000 bases long, but are only about 100 bases long in eukaryotes. DNA polymerase cannot initiate replication without a primer sequence, which is made of RNA by a primase. Most of the remainder of this section is summarised in Table 1, which lists the proteins responsible for each of the processes involved in DNA replication. Questions from this table seem to come up every year, so it's probably worth just committing the damn thing to memory by rote.

Three additional things to note about replication which are not covered in the table are as follows.

1. DNA polymerase has built-in error correcting. When it detects an error, its inherent 3' to 5' exonuclease activity backs it up, removing the incorrect nucleotide as it goes.

	Bacteria (*E. coli*)	**Eukaryotes**
Origin recognition Protein	Dna A	ORC (6 subunits)
Helicase loading	Dna C	Cdc6 and Cdt1
Replicative helicase (Melts and unwinds the DNA)	Dna B	MCM (Heterohexamer)
Single Stranded Binding Protein	SSB	RPA (3 subunits)
Primase (Makes RNA primers for replication)	Dna G	Primase complex PriL, PriS, B subunit, DNA polymerase α
Polymerase	DNA polymerase III (Family A polymerase)	DNA Polymerase α, δ, ε (Family B polymerases)
Sliding clamp (confers processivity to polymerase)	β Clamp (Dimer)	PCNA (Homotrimer)
Clamp loader (opens the ring of the sliding clamp)	γ Complex (5 different subunits)	RFC (1 large subunit and 4 small subunits)
Ligase	DNA ligase (NAD dependent)	DNA ligase (ATP dependent)
Primer Removal	DNA Polymerase I	FEN1/Rad 2

Table 1: Proteins involved in DNA replication.

2. Since DNA is coiled into a double helix, it must unwind in order to be replicated. The molecules which do this are called **topoisomerases** and come in two types, I and II. Type I cuts a single strand and either winds it around the other strand before resealing it (type IB) or passes the intact strand through the gap (type IA), while type II cuts both strands and passes another part of the DNA molecule through the gap created. Faulty topoisomerases have something to do with hereditary breast cancer and Bloom's Syndrome. **Ciprofloxacin** inhibits a type II bacterial topoisomerase called DNA girase, which unwinds positive supercoiling. Eukaryotic topoisomerases are unaffected, making ciprofloxacin an effective antibiotic. **Doxorubicin** and **Etopocide** inhibit eukaryotic topoisomerase II and are therefore used as chemotherapeutic agents in cancer treatment. **Camptothecin** inhibits eukaryotic type IB topoisomerases and is thus also used in cancer treatment.

3. Telomeres. They form unique structures at the end of eukaryotic chromosomes and

are replaced by telomerase which is a self-templating reverse transcriptase.

It may be of interest to note that the helicase loading protein cdc6 is degraded as DNA polymerase passes the origin of replication, thus preventing MCM from re-attaching. Presumably there is some cell-cycle checkpoint which prevents replication from beginning until all origins of replication are bound by a combination of ORC, cdc6 and MCM. This would then ensure that all cdc6 can be degraded at the same time and thus preventing any DNA from being copied more than once. See Section 8 for more details on cell cycle checkpoints.

4 DNA repair

It has been estimated that about 10^5 DNA lesions happen in each cell in our bodies each day. In order to repair all of these lesions, the DNA repair machinery must thus carry out about 10^{18} repairs every day since there are about 10^{13} cells in our bodies. In eukaryotes a complex set of signals are also sent to arrest the cycle while the damage is repaired. If the damage is too severe to repair, the cell will commit suicide through the process of apoptosis[1].

There are many things which cause mutations in DNA. Examples include the following.

- Ionising radiation, e.g. γ rays, X-rays, u.v. radiation, α and β radiation.
- Reactive oxygen species (ROS). Formed from ionising radiation or normal enzymatic activity. Examples are the hydroxyl radical, hydrogen peroxide and the superoxide radical.
- Chemical carcinogens:
  - Polycyclic aromatic hydrocarbons.
  - Aromatic amines.
  - Alkylating agents.
  - Electrophilic carcinogens, which react with hydrophilic sites in purine and pyrimidine rings to form covalent adducts with DNA, resulting in errors in repair and hence mutations.
  - Tobacco contains several potent carcinogens which form a wide range of DNA adducts.

Thus DNA repair enzymes are important tumour suppressor genes. We review the four main mechanisms of DNA repair below. Another method, called one-step repair, also exists. It utilises the enzyme alkyltransferase to reverse the effects of alkylating agents on guanine.

[1] Apoptosis is the subject of Andrew Wyllie's lectures.

4.1 Base excision repair

Sometimes bases themselves mutate. The commonest example of this is that cytosine can spontaneously deaminate, turning it into uracil. Since uracil base pairs with adenine instead of guanine this can result in a change from G-C to A-T when the cell divides, or if the base is within a gene coding sequence an A in the transcribed RNA instead of a G. This spontaneous source of mutation explains why DNA uses thymine instead of uracil, as if uracil were normally present in the genome there would be no way to recognise this kind of mutation. The enzyme uracil DNA glycosylase (UDG) recognises the aberrant uracil and removes it by breaking the glycosidic link with its sugar. AP endonuclease recognises the resulting abasic site and cleaves the backbone. DNA polymerase I then replaces the original cytidine in bacteria. In eukaryotes this is carried out by DNA polymerase β.

4.2 Nucleotide excision repair

A variety of environmental factors including ultraviolet radiation and cigarette smoke can cause distortions in the helical structure of DNA. Examples include pyrimidine dimers and benzoprene-guanine. Phosphodiester bonds several nucleotides away from the lesion are broken in the strand containing the lesion and the piece of DNA containing the lesion is removed. A DNA polymerase then fills in the missing section by base-pairing with the intact strand. In prokaryotes the uvrABC exiconuclease complex recognises the lesion and breaks the phosphodiester bonds and uvrD excises it. DNA polymerase I fills in the gap. The method is similar in eukaryotes, though it does not use any of the same enzymes. The gap is filled in by DNA polymerase ε.

Inherited mutation in the nucleotide excision repair machinery leads to hypersensitivity to sunlight and is associated with the syndrome **Xeroderma Pigmentosa**, which is characterised by a predisposition to skin cancers.

4.3 Mismatch repair

Although the DNA polymerases are very good at producing accurate copies in replication, errors do occur. The mechanism for repairing these errors has not been fully worked out for many organisms (including humans), so the remainder of this section refers to *E. coli*. In *E. coli* this problem is dealt with by tagging DNA with methyl groups at GATC sequences. Because methylation occurs only some time (between a few seconds and a few minutes) after replication, the newly synthesised strand will be hypomethylated with respect to the template strand. A complex of enzymes called MutL and MutS bind to mismatched DNA and then track along the DNA in both directions by spinning out a loop until a hemimethylated (only one strand methylated) GATC sequence is reached. An enzyme called MutH then cleaves the phosphodiester bond on the 5' side of the G on the unmethylated strand. The unmethylated strand is then degraded from there past the mismatch and DNA polymerase III replaces it with (hopefully) the correct sequence by base pairing with the template strand once more. In this way, Mismatches have been shown to be successfully repaired within about 1000 base pairs of a hemimethylated GATC sequence.

Proteins analogous to MutS and MutL have been found in all eukaryotic cells thus far studies, though nothing analogous to MutH has been found. In humans, mutations in these genes is associated with a high incidence of colorectal cancer. Inheritance of faulty mismatch repair genes causes **Hereditary Nonpolyposis Colorectal Cancer**.

4.4 Recombination and double stranded break repair

4.4.1 Homologous recombination

It may appear strange that we discuss recombination, a process which is critical in generating diversity during meiosis, within a section on DNA repair. However there is considerable evidence that the most important function of recombination is to repair double-stranded breaks.

Broken ends of DNA molecules are prone to degradation. Thus if a cell were to simply stick two broken ends together, not only could part of one chromosome be transposed onto another chromosome if both were broken at the time of repair, but part of the sequence at the break may be lost. Fortunately, most multi-cellular organisms are diploid, that is to say that they normally contain two copies of each chromosome. While these are not identical (one having been inherited from the mother and one from the father) they are at least nearly identical, and thus the intact chromosome can act as a template for the broken chromosome.

The exact process is best described as complicated in prokaryotes and ridiculously complicated in eukaryotes. In brief (and believe it or not this really is a brief summary), the broken ends of the DNA molecule are first trimmed back in such a way as to leave a single-stranded piece with a 3' end sticking out the end of each broken end. This single-stranded end is recognised by highly conserved enzymes. In prokaryotes the enzyme is called RecA. In eukaryotes there are several steps involving several enzymes including RPA (the single-stranded binding protein involved in DNA replication) and an enzyme called Rad51, which displaces RPA and also binds to the corresponding section of the intact chromosome. RecA and Rad51 then direct a process called **strand invasion**, which replaces one of the strands of the intact chromosome with the single-stranded end from the broken chromosome. The strands bind by base pairing in the normal way. The same happens with the other broken end slightly further along. In a process called **branch migration**, the branch point moves along (catalyzed by a molecular motor protein complex called RuvAB in prokaryotes) and further strand invasion completes the crossing over between the two homologous chromosomes. The gaps caused by the break and subsequent trimming can then be resolved by copying the template provided by the intact chromosome (using DNA polymerase I or III in prokaryotes). The points at which DNA strands switch from one double-strand to another are called **Holliday junctions** after Robin Holliday, who proposed the method in 1964 to explain the phenomenon of crossing over. In order to get back to two independent pieces of double stranded DNA, one strand of each double-stranded piece between the two Holliday junctions is cut and reattached to the opposite strand. This can be done in two different ways, each of which appears to be equally likely. Either the two strands which do not cross over at the Holliday junctions are cut and rejoined, resulting in one end from the broken chromosome and one end from the

previously intact chromosome, or vice versa.

The proteins BRCA1 and BRCA2 both interact with Rad51 in the homologous recombination pathway, though their precise function is not clear. An inherited defect in either of these genes results in an 80% risk of developing breast cancer across a woman's lifetime. Homologous recombination can also go wrong. The most famous example of this is the so-called Philadelphia chromosome, where parts of chromosomes 9 and 22 are exchanged, resulting in a greatly shortened chromosome 22 with a small piece of chromosome 9 on the end of it and a lengthened chromosome 9 which has lost a small piece off the end, which has been replaced by a large piece of chromosome 22. The former is called the Philadelphia chromosome after the city in which it was first recorded. This abnormality is present in nearly all cases of chronic myelogenous leukemia.

4.4.2 Non-homologous end joining

Non homologous end joining is another method of repairing double stranded DNA breaks. As the name suggests, this is simple joining of the broken ends. This can result in the loss of some genetic material, as the ends often need to be tidied up before they can be joined. One use for non-homologous end joining other than DNA repair is in generation of diverse antibodies. While this is really interesting, it does not appear that we need to know the details so I'll leave the description of it to the textbook authors (Lehninger is particularly good). One interesting thing to note is that the enzymes which mediate this process of antibody diversity generation are closely related to enzymes which move mobile genetic elements called transposons around the genome (and even between genomes). It is possible therefore that they originated as transposons which conferred great evolutionary advantage to the host. Forty-five percent of the human genome appears to be composed of transposons or their fossils, so it is certainly clear that many transposons have been taken up and retained over the course of our evolutionary history.

5 Genes and Transcription

5.1 What's in a gene?

In prokaryotes, a gene is simply a stretch of DNA which is **transcribed** into a functional RNA molecule. Most of the genome of prokaryotes consists of genes. In eukaryotes the situation is a little more complex due to the presence of **non-coding** stretches of DNA. About 75% of the human genome contains no genes. In addition, most eukaryotic genes are split up into stretches which are spliced out of the transcribed RNA molecule called **introns** and stretches called **exons** which are actually translated into proteins (*intr*ons are *inter*vening sequences, *ex*ons are *ex*pressed sequences). Counting the introns along with other non-coding sequences, about 98.5% of the human genome is non-coding. Even RNA with the introns removed contains non-coding sections called untranslated regions (UTRs) at both ends. Genes also contain regulatory sequences which ensure that the right genes are expressed at the right time.

5.2 Transcription in prokaryotes

Genes are encoded in both of the two DNA strands, but for any one section of double stranded DNA the notes say that only one strand is copied (called the template strand). This does not appear to be true for all organisms, however, as there is a picture in Lehninger of the adenovirus genome, which appears to contain several stretches where both strands code for protein in an overlapping way. The non-template strand is called the coding strand, as its base sequence is the same as that of the transcribed RNA (except that uracil replaces thymine).

RNA is synthesised from NTPs by a single RNA polymerase enzyme by repetitive addition to the 3' end with release of PP_i. Thus the RNA chain grows in the 5' to 3' direction, the same as for DNA synthesis. The initial nucleotide is not bonded to anything other than phosphate at its 5' end and thus retains its three phosphate groups. RNA polymerase has 5 core subunits ($\alpha\alpha\beta\beta'\omega$) and a σ subunit, of which there are many forms. For some reason the notes don't mention the ω subunit. This must just be a typo, as all of the books are quite clear about its existence. It is the σ subunit which is responsible for recognising promoter sequences upstream of genes and thus guiding RNA polymerase to transcribe specific sequences rather than at random. When the core RNA polymerase enzyme binds a σ subunit it is said to be in its holoenzyme form. The σ subunit breaks off after RNA polymerase has transcribed about 6–8 nucleotides, and can then bind to another core enzyme to direct further transcription. The most common σ is called σ^{70}, as its molecular weight is around 70kDa. The σ subunit involved in directing transcription of heat shock genes is called σ^{32}, as it weighs about 32kDa. Aren't biochemists imaginative when it comes to names? There are two especially highly conserved promoter sequences about 10 and 35 nucleotides upstream of genes. The -10 region is sometimes called the Pribnow box and has the consensus sequence TATAAT. The -35 region has the sequence TTGACA and is slightly less highly conserved. By convention when gene sequences (here consensus sequences) are discussed, it is the sequences on the coding strand (not the template strand) which are described. After RNA polymerase binds to dsDNA it unwinds about 10–15 base pairs of DNA so that the template strand can be transcribed and then initiates transcription. Unlike DNA polymerase, RNA polymerase has no built-in error correcting capabilities, as it has no 3' to 5' exonuclease activity. Thus errors are much more frequent than in DNA replication, occurring about once in every 10^4 bases. However the higher turnover of RNA (average $t_{\frac{1}{2}} \approx 2$ minutes in prokaryotes), and the synonyms that exist in the genetic code mean that this matters a lot less than errors in DNA replication would, which are effectively un-repairable once methylation has taken place.

The remaining question about prokaryotic transcription is how RNA polymerase knows when to stop. Termination usually happens in one of two ways.

1. A self-complimentary GC rich sequence centred 15–20 nucleotides upstream of the termination point form a hairpin. This causes RNA polymerase to stall. In addition, several (4–10) consecutive downstream uracil bases in the RNA molecule reduce binding affinity, promoting dissociation.

2. A protein called ρ which shares many of the features of F_1-ATPase binds the RNA molecule in its central channel and uses hydrolysis of ATP (or other nucleoside

triphosphates) to drive itself up the RNA in the 3' direction. The ρ protein recognises a GU-rich sequence called a *rut* (rho utilisation) in the transcript. The exact mechanism by which it causes termination of transcription is unclear. Presumably, when it reaches RNA polymerase it interacts with it, causing the RNA to dissociate. The ρ protein may then simply proceed off the end of the transcript, or the interaction with RNA polymerase may cause it to dissociate as well. In its open position, ρ is helical rather than circular like F_1-ATPase. It is not clear whether this helicity is preserved when it binds RNA and adopts its closed position.

There are several drugs which inhibit transcription in prokaryotes and are therefore useful antibiotics. As mentioned in Section 3, ciprofloxacin inhibits DNA girase, a type II bacterial topoisomerase. By unwinding DNA in order to transcribe it, RNA polymerase causes positive supercoiling downstream and negative supercoiling upstream. Inappropriate superhelicity causes transcription to halt and thus ciprofloxacin inhibits transcription as well as DNA replication. Two other related antibiotics, **rifamycin B** and **rifampicin** also inhibit transcription[2]. These antibiotics inhibit transcription by binding to the β subunit of RNA polymerase. Rifampicin is useful in the treatment of tuberculosis, among other things.

5.3 Transcription in Eukaryotes

The basic principles of eukaryotic transcription are the same as those of prokaryotic transcription but the details are (as you might expect) far more complicated. Instead of a single RNA polymerase, there are three nuclear RNA polymerases[3]. RNA polymerase I transcribes only a single RNA, which is the precursor of three out of the four eukaryotic ribosomal RNAs (rRNAs). It is located in nucleoli. RNA polymerase II transcribes the precursors of mRNA, many small nuclear RNAs (snRNAs) including U1–U5 snRNAs, and many micro RNAs (miRNAs). In addition, the introns of some genes contain small nucleolar RNAs (snoRNAs), which are involved in splicing pre-rRNAs, so RNA polymerase II obviously transcribes these too. RNA polymerase III transcribes many small RNAs including the transfer RNAs (tRNAs), the remaining rRNA (5S rRNA) and U6 snRNA. The eukaryotic RNA polymerases are also typically large, containing 10–14 subunit, several of which are homologous to the five basal subunits of prokaryotic RNA polymerase. Five of the smaller subunits are identical between the three polymerases, and the the two largest subunits are the same in RNA polymerases I and III. Since RNA polymerase II transcribes the genes which are destined to be translated to protein, and the central dogma of molecular biology gives great weight to those genes, as it was once thought that every gene represented the code for a protein, we will focus on RNA polymerase II, with only occasional references to the other two.

The role of the σ subunit of prokaryotic RNA polymerase is played by a rather complex structure known as the **pre-initiation complex** (PIC), which guides RNA polymerase II

[2]Rifamycin B is derived from *Streptomyces mediterranei*. Rifampicin is a semi-synthetic derivative of rifamycin B.

[3]There are also RNA polymerases inside mitochondria and chloroplasts which resemble prokaryotic RNA polymerase

to the start of genes, ensuring accurate initiation of transcription. The basic structure of this complex is built up as follows.

1. TBP binds at the promoter.
2. TAFs bind TBP to make the TFIID complex.
3. TFIIB binds the TFIID complex.
4. A complex of TFIIF and RNA polymerase binds TFIID, the DNA downstream of TFIID, and TFIIB.
5. TFIIE and TFIIH bind the downstream side of RNA polymerase II. TFIIH also binds the DNA upstream of RNA polymerase II.

TBP stands for TATA binding protein, which binds to the **TATA box**, a highly conserved sequence centred about 25–30 nucleotides upstream of the start of the gene which resembles the Pribnow box in consensus sequence (TATAAA). TAF stands for TBP associated factor. TFII is an abbreviation for "transcription factor for RNA polymerase II". In addition to this basic sequence of steps, some other transcription factors may also bind, for example TFIIA often binds upstream of TFIID at the same time as TFIIB.

In addition to the TATA box, which assumes the role of the Pribnow box, there exist other well conserved promoter sequences, for example the CCAAT box, located between 70 and 90 nucleotides upstream, and the so-called **upstream promoter element**, which is a GC-rich region between 107 and 187 nucleotides upstream of the gene start position.

TFIIH has two important roles. Firstly, it acts as a helicase, unwinding the DNA so that transcription can begin. Secondly, it phosphorylates RNA polymerase II on several of the residues of the carboxyl-terminal domain (CTD) of one of its subunits, causing it to string out behind the polymerase. Cyclin-dependent kinase 9 (CDK9) also phosphorylates the CTD of RNA polymerase II at this stage. In addition to helping to release some of the initiation factors, this phosphorylation plays an important role in RNA processing, the importance of which will become apparent in the next section. After initiation, TFIIE dissociates and TFIID and B (and TFIIA if present) are left behind, ready to initiate the binding of another TFIIF/RNA polymerase II complex. Some books say that TFIIH dissociates shortly after initiation, leaving only TFIIF to remain bound throughout the elongation phase of transcription, while others say that both remain bound throughout elongation. I'll try to chase this down at a later date. TFIIH is also associated with the nucleotide excision repair pathway, and can recruit the enzymes of that pathway if an atypical base is encountered during transcription. I'm not sure exactly how it would do this if it has dissociated from the transcription complex, so maybe it doesn't dissociate, or can re-associate if RNA polymerase stalls. During elongation, other proteins called elongation factors typically bind to RNA polymerase II, increasing the speed of transcription and coordinating transcription with post-transcriptional processing. Like DNA polymerase, and unlike prokaryotic RNA polymerase, RNA polymerase II has error correcting ability due to built-in 3' to 5' exonuclease activity. This makes sense when you consider the fact that $t_{\frac{1}{2}}$ can be several days for some RNA transcripts.

Several chemicals block transcription in eukaryotes. Actinomycin D and Acridine inhibit both prokaryotic and eukaryotic transcription by intercalating between neighbouring GC base pairs in DNA. They are too toxic for general use, though actinomycin D is sometimes used in cancer therapy. The poisonous mushroom *Amanita phalloides* contains several toxins including α-amanitin, which binds tightly to RNA polymerase II. It is unclear exactly how α-amanitin inhibits transcription, but the current theory is that it interferes with the step which feeds the growing RNA chain through the polymerase, preventing the 3' nucleotide from vacating the active site.

Unlike prokaryotic transcription, which has two well-defined pathways for terminating transcription, eukaryotic transcription seems to just peter out. This has been shown experimentally, as pre-mRNA has a variable 3' end. As we will see in the next section, post-processing of the RNA ensures that this does not result in proteins with a variable 'tail'.

5.4 RNA processing

By and large, RNAs are only processed prior to translation in eukaryotes. While there are a few exceptions to this rule, it is is easy to see why this is. In eukaryotes the processes of transcription and translation are separated, as the former takes place in the nucleus while the latter takes place in the cytoplasm. By contrast, in prokaryotes the processes of transcription and translation are coupled so that on electron micrographs it is possible to see chains of ribosomes attached to an RNA as it is being transcribed from a DNA template. Thus there is little opportunity for RNA processing. In addition to this, RNA processing makes it possible to tag certain RNAs for export to the cytoplasm, while retaining other RNAs in the nucleus, enabling further separation of nuclear processes catalysed by some of the retained RNAs.

One of the most important things to realise about RNA processing is that it is not a separate process from transcription, but actually happens while transcription is going on. There are three key steps in RNA processing: addition of a cap to the 5' end (the end which is transcribed first), splicing to remove introns, and addition of a long stretch of adenosines to the 3' end (the so-called poly-A tail). It is traditional to discuss 5' capping and 3' polyadenylation first, but I feel that this makes it easy to forget that these processes take place during transcription, so that polyadenylation must, by necessity, occur last, and 5' capping first. So I'll break with tradition and discuss events in the order in which they occur *in vivo*.

Shortly after initiation, when the transcript is only around 30 nucleotides long, the 5' end is capped by guanosine. There are three steps in this process. First of all, one of the three phosphate groups is removed from the 5' nucleotide of the RNA[4]. Secondly, GTP is added to the trailing 5' phosphate with the loss of PP_i, making a triphosphate bridge between the 5' group of the first nucleotide of the RNA and the 5' group of the guanosine. This 5' to 5' triphosphate link is very unusual and I couldn't find any reference to it occurring in any other circumstance. Finally, the N_7 group of the capping guanosine

[4]Recall that unlike all other nucleotides in RNA, the initial nucleotide in a transcribed RNA molecule retains all three phosphates.

is methylated by the transfer of a methyl group from SAM to guanine-7-methyltransferase and then to the N_7 group of the guanine base. See the metabolism notes for a discussion of SAM. The 2' hydroxyl groups of the first two nucleotides may also be methylated by another process which uses SAM.

The enzymes which catalyse this process all appear to be bound to the CTD of RNA polymerase II, which reinforces the point that transcription and processing are all part of the same process. The cap is essential for export to the cytoplasm and for efficient initiation of translation. After capping, the cap remains attached to the **cap binding complex** (CBP), which is also attached to the CTD of RNA polymerase II and helps to protect the growing RNA from the action of exonucleases, as well as facilitating the splicing of introns by keeping it close to the CTD.

There are four classes of introns. Classes I and II are **self-splicing** and were the first catalytic RNA molecules to be discovered[5]. Unfortunately, fascinating though these are they are not discussed in the notes and so there's not much point learning about them! The third class of introns require the action of a huge complex of proteins and snRNAs called a **spliceosome**. Individual protein-RNA complexes within the spliceosome are called snRNPs (pronounced *snurps*): small nuclear ribonucleoproteins. Introns in this class begin and end with the sequences GU and AG respectively and have an A somewhere within 20–50 nucleotides of the 3' end. The sequence of events involved is as follows.

1. U1 RNP binds to a well-conserved sequence at the 5' end of the intron by complementary base pairing of the U1 RNA.

2. U2 RNP binds to a recognition sequence including the aforementioned mid-intron adenine residue by complementary base pairing of the U2 RNA.

3. U4, U5 and U6 RNPs bind to both U1 and U2, bringing the two domains together.

4. The RNA is cleaved at the 5' end and rejoined to the 2' of the adenosine, forming a loop.

5. The 3' end of the intron is cleaved and resulting 5' end of the next exon is bonded to the now free end of previous exon. All of the bound RNA is then released, leaving a lasso-shaped intron called a **lariat**, and the intron-less RNA transcript.

As has already been hinted at, the spliceosome appears to form on the CTD of RNA polymerase II, and thus splicing also occurs during transcription. Both U1 and U2 appear to remain fixed to the CTD until splicing is complete. Although the formation of the spliceosome requires energy in the form of ATP, the actual splicing does not. As the process involved is very similar to that of type II introns (not discussed!), it has been suggested that the spliceosome has evolved from self-splicing precursors, possibly to improve efficiency or accuracy.

Alternative splicing is involved in several key processes including the generation of variable regions in the heavy chain of antibodies. A cancer-related example is the Wilm's tumour gene WT1, which is a tumour suppressor gene which acts as a transcription factor.

[5]RNA molecules with catalytic properties are known as **ribozymes**.

Although alternative splicing occurs normally in this gene, altered patterns of alternative splicing caused by mutation at splice sites can lead to loss of some of the DNA binding domain or other critical regions.

As mentioned in the previous section, transcription in eukaryotes is not terminated at a specific site, but rather terminates at an arbitrary point some distance downstream of the end of the transcript. This does not affect transcription, however, as the RNA is cleaved to get rid of this superfluous region and a sequence of around 240 adenosine nucleotides is added in its place. The steps of polyadenylation are as follows.

1. The cleavage and polyadenylation specificity factor (CPSF) binds the well-conserved recognition sequence AAUAAA.

2. CPSF recruits a multi-enzyme complex including poly(A) polymerase (PAP) and an endonuclease which cleaves the RNA at a specific site around 15 to 25 nucleotides beyond the aforementioned recognition sequence.

3. PAP then catalyses the addition of adenosine nucleotides to the resulting 3' end.

The multi-enzyme complex of which PAP is a part yet again binds to the CTD of RNA polymerase II and thus is possibly also involved in termination of transcription.

The poly-A tail appears to lengthen the cytosolic lifespan of mRNA, although it is not clear whether this is due to the tail itself or the fact that a protein called poly-A binding protein (PABP) binds to it, as either could conceivably protect the important bit from the action of exonucleases.

Although not really part of the normal process of RNA processing, editing of the RNA transcript also occurs in a few cases. The most frequent editing that occurs is of tRNAs, which are just full of unusual bases such as inosine, which is formed from deamination of uridine. Deamination of cytosine also occurs, creating uridine. Spurious RNA editing has also been implicated in neurofibromatosis 1, where the codon CGA becomes UGA when cytosine is deaminated, replacing an arginine residue in the middle of the protein with a stop codon, truncating the resulting protein. We discuss the genetic code in the following section.

6 Translation

6.1 The genetic code and tRNAs

The genetic code is pretty simple really. It is non-overlapping and uses three nucleotides to represent each amino acid. Each three nucleotide sequence is called a **codon**. In addition, there is also a "start" codon (AUG) and three "stop" codons (UGA, UAG and UAA). There is a certain amount of redundancy in the system, as there are $4^3 = 64$ possible three nucleotide sequences but only 20 amino acids (plus start and stop). As we have already seen, there are three different "stop" codons. While some amino acids are represented by only a single codon (eg. tryptophan is UGG), others are represented by several (eg. leucine is represented by UUA, UUG, CUU, CUC, CUA and CUG). The start and stop codons are sometimes referred to as nonsense codons, though I'm not sure why the start codon

is referred to in this way as it perfectly legitimately represents methionine. Perhaps it is because a mutation which generates an extra AUG in the 5' UTR may lead to nonsense.

The three nucleotide codons representing amino acids are recognised by base pairing by small RNA molecules called transfer RNAs (tRNAs). Unlike mRNA, tRNAs contain a large number of unusual bases including pseudouridine y, dihydrouridine and inosine. Transfer RNAs fold up into a shape that looks like a cross in two dimensions, though in three dimensions it actually looks more like an "L". The anticodon, which base pairs with the codon on the mRNA is at the opposite end of the molecule from the long arm of the cross, which terminates in the 5' and 3' ends, with a little piece of the 3' end sticking out. It is this unpaired 3' end which binds the amino acid, making an aminoacyl-tRNA. Amino acids are joined to tRNAs by molecules called aminoacyl tRNA synthetases (RSs). There are 20 different RSs, one for each of the amino acids. The two steps involved include first of all addition of AMP to the amino acid and a tight non-covalent binding by the enzyme. The AMP is donated by ATP and involves release of pyrophosphate (PP_i). The second step involves donation of the bound amino acid to its tRNA with release of the AMP bound in the first step. RSs require magnesium as a coenzyme. The reaction can be summarised as follows.

$$\begin{array}{ccccc} \text{ATP} & & & & \text{tRNA} \\ & \searrow & & & \searrow \\ \alpha\alpha \rightarrow & \text{Enzyme} \cdot \alpha\alpha\text{AMP} & & \rightarrow \alpha\alpha\text{tRNA} \\ & \nearrow & + & & + \\ \text{Enzyme} & & PP_i & & \text{AMP} \end{array}$$

Here $\cdot$ represents a tight but non-covalent bond, and $\alpha\alpha$ is shorthand for an amino acid. A more standard chemical equation which recognises the role of magnesium as a cofactor is:

$$\text{amino acid} + \text{tRNA} + \text{ATP} \underset{\text{aminoacyl-tRNA synthetase}}{\overset{\text{Mg}^{2+}}{\rightleftharpoons}} \text{aminoacyl-tRNA} + \text{AMP} + PP_i.$$

Although there are 64 codons, there are between 45 and 100 different tRNAs present in different types of cells. The reason that some cells can get away with only 45 tRNAs to recognise 64 different codons is down to the redundancy of the code together with something called **wobble**. When several different codons represent the same amino acid, the difference is usually found at the third base. This is because the third codon base is bound only loosely by the first base of the anticodon[6]. For example if the first anticodon base is U it can recognise either A or G, or if it is inosine (I), it can recognise A, U or C. Thus the anticodon IAU can recognise any of the sequences AUA, AUU or AUC, which all encode isoleucine.

6.2 Ribosomes and initiation of translation

Ribosomes are basically big RNA enzymes with a bit of protein for additional support. They consist of a large and small subunit which only come together when the small subunit

[6]Recall that nucleotide sequences are always read in the 5' top 3' direction and thus the first base of the codon is bound by the third base of the anticodon, the second by the second and the third by the first since base pairing is *antiparallel*.

has already bound the initiation sequence of an RNA molecule. Their size is measured in **Svedberg** units, which is a measure of sedimentation rate and is thus a very crude measure of size and weight. The bacterial subunit are 30S and 50S, making 70S when combined. Eukaryotic ribosomes are a little larger — 40S and 60S, making 80S when combined. Yes that's right folks, the usual rules of arithmetic do not apply to Svedberg units.

In prokaryotes, the sequence of translation initiation is as follows. Initiation factors IF-1 and IF-3 bind to a free 30S subunit. IF-3 prevents the small and large subunits coming together prematurely, while IF-1 prevents premature binding of tRNAs. The 30S subunit then binds to the mRNA and is guided to the correct initiation sequence by the **Shine-Dalgardo** consensus sequence, which is a purine-rich sequence 8–13 nucleotides upstream of the initiation sequence. Its consensus sequence is GGAGG and it is recognised by a sequence near the 3' end of the 16S rRNA, which is part of the 30S subunit. A special initiation aminoacyl-tRNA then binds, together with IF-2, which has a molecule of GTP bound. Finally, the 50S subunit binds, hydrolysing the bound GTP to GDP+P_i and releasing IF-1, IF-2 and IF-3.

In eukaryotes the sequence is slightly different and the initiation factors are called eIFx instead of IF-x, where x is a number. Firstly, there is no Shine-Dalgarno recognition sequence. Instead, the 40S subunit first binds the initiation aminoacyl-tRNA and then scans along the mRNA until it reaches the first AUG[7]. Thus it is still base pairing which enables recognition of the correct start site, but with eukaryotes it is between tRNA and mRNA, whereas with prokaryotes it was between rRNA and mRNA. Secondly, eukaryotes have at least nine initiation factors, including eIF4F, which binds to the 5' cap, and PAB, which binds the poly-A tail and also eIF4F, thus keeping the two ends of the mRNA together.

In prokaryotes, the initiation aminoacyl-tRNA is N-formylmethionine-tRNA (fMet-tRNA$^{\text{fMet}}$), while in eukaryotes it is Met-tRNA$_{\text{i}}^{\text{Met}}$. In both cases, although the start codon AUG specifies methionine, the tRNA is different from normal Met-tRNA$^{\text{Met}}$, which adds methionine in the middle of a polypeptide chain. In prokaryotes, **deformylase** cleaves the formyl group. In all organisms the initial methionine is sometimes cleaved by an aminopeptidase, depending on what the next amino acid residue is.

In rare cases, translation in eukaryotes is initiated in the middle of an mRNA. This is called **internal initiation**. Instead of scanning to the initiation site, an **internal ribosome entry site** directs binding of the 40S subunit. It was discovered in small RNA viruses called (funnily enough) picornaviruses. This class of viruses include polio, hepatitis C, foot and mouth disease and rhinoviruses (responsible for most cases of the common cold). Some of these viruses include a protease which cleaves one of the cap binding factors (part of eIF4F), thus preventing the host cell from translating its own proteins. Internal initiation is important because the IRES of these viruses provides a nice drug target. In addition, the human oncogene *c-myc* has an IRES. Mutation in this region which increases ribosomal affinity correlates with multiple myeloma.

[7]Scanning requires hydrolysis of ATP.

6.3 Elongation of the polypeptide chain

There are three binding sites on ribosomes for tRNAs: the aminoacyl site (A), the peptidyl site (P) and the exit site (E). Initiation factor IF-1 binds the A site, preventing binding prior to initiation. Normal tRNAs can only enter a ribosome at the A site, and initiation tRNAs can only enter a ribosome at the P site. Once the ribosome has come together, IF-1 is released, and aminoacyl-tRNAs can bind to the A site. This binding requires the presence of an elongation factor, which is a GTPase. Following hydrolysis of GTP to GDP+P_i and binding of the aminoacyl-tRNA in the A site, the elongation factor is released and transfer of the peptide bond from the tRNA in the P site to amino acid of the aminoacyl-tRNA in the A site is catalysed by the rRNA of the ribosome itself. This step requires no energy. Finally, the ribosome moves along the mRNA by three nucleotides, moving the tRNA which was in the P site to the E site and the aminoacyl-tRNA which was in the A site to the P site. This step again requires an elongation factor which is again a GTPase. The tRNA in the E site freely dissociates from the ribosome and the cycle begins again. Note that at this point the growing polypeptide chain is attached to the tRNA at the P site of the ribosome. In prokaryotes the first elongation factor is called EFTu, while the second is called EFG. In eukaryotes they are called eEF-1 and eEF-2.

6.4 Termination

The remaining question is what happens about stop codons. While it might be expected that stop codons are matched by tRNAs with no aminoacyl groups, in fact proteins called release factors (RFs) bind to the A site when a stop codon is present, move the tRNA in the P site along to the E site and facilitate the dissociation of the small and large subunits of the ribosome. The binding of the initial release factor to the A site modifies the peptidyl transferase activity of the rRNA so that the bond is simply hydrolysed (transferred to water). As it is no longer attached, the polypeptide chain dissociates from the ribosome.

6.5 Antibiotics and toxins which interfere with translation.

The following antibiotics interfere with translation (in prokaryotes).

Streptomycin binds to the 30S subunit and causes misreading of mRNA. In high concentrations it also inhibits initiation.

Tetracycline blocks the A site on the ribosome, thus inhibiting binding of aminoacyl-tRNA to the 30S subunit.

Chloramphenicol inhibits the peptidyl transferase activity of the 50S subunit. Used selectively, as it also inhibits mitochondrial protein synthesis and is thus relatively toxic.

Erythromycin causes "molecular constipation" by blocking the exit site for the growing polypeptide chain on the 50S subunit.

Oxazolidines are experimental drugs still in clinical trials which inhibit formation of the initiation complex.

In addition to antibiotics there are other chemicals which interfere with translation.

Cyclohexidine blocks the peptidyl transferase of eukaryotic ribosomes.

Diphtheria toxin modifies one of the residues of eEF-2, inactivating it.

Ricin inactivates the 60S subunit by depurinating a specific adenosine in 28S rRNA.

7 Regulation of gene expression

7.1 Regulation in prokaryotes

In prokaryotes, almost all regulation of gene expression occurs at the level of transcriptional activation. There are three main classes of proteins which regulate transcription. We have already met the first kind, called **specificity factors** in the form of the σ subunit of prokaryotic RNA polymerase. In addition to the existence of different σ subunits, the similarity of the -10 and -35 boxes to the consensus sequences which the σ subunit recognises can cause expression to vary over a 3000-fold range without even changing σ.

The other two classes of proteins, **repressors** and **activators**, fine-tune this rather gross control according to more subtle environmental conditions, such as the availability of sugars or amino acids. For historical reasons these proteins are usually presented within the settings of the *lac* and *trp* **operons**[8]. We will only discuss the *lac* operon. The *trp* operon is regulated in a quite different way, but since it is not discussed in the notes we will leave its discussion to the textbooks — see Lehninger Section 28.2 for further details.

Three genes, β-galactosidase, galactoside permease and thiogalactoside transacetylase are involved with the metabolism of lactose, a disaccharide. The first hydrolyses lactose to galactose and glucose while the second transports lactose and other galactosides into the cell. The function of the third has not been definitively established[9]. The operon is controlled by two regulatory proteins, the ***lac* repressor** and the **catabolite gene activator protein (CAP)**, also called **cAMP receptor protein (CRP)**. The former, as the name suggests, is a **negative regulatory protein**, or repressor, while the latter is a **positive regulatory protein**, or activator. Both of these regulatory proteins respond to small molecules which indicate the need for the activation of the lactose pathway by

[8] An operon is a collection of genes (and their regulatory regions) which are co-expressed on a single RNA and thus controlled by a single promoter.

[9] What is known is that thiogalactoside transacetylase, or galactoside transacetylase as it has come to be known, transfers acetyl groups to galactosides which cannot be degraded but are nevertheless transported into the cell by galactoside permease. It has been shown that a functional version of this gene confers a competitive advantage in the presence of indigestible galactosides *in vitro*. The modified galactosides are transported out of the cell, but cannot get back into it as galactoside permease will no longer transport them in their modified form. The reason for the doubt about its function is that physiologically available indigestible galactosides have not been identified, so it seems to be an enzyme without a purpose *in vivo*. A useful review can be found at Roderick (2005) *Comptes Rendus Biologies* 328(6) pp. 568–575.

changing shape between an active configuration, which binds DNA, and an inactive configuration which does not. When the *lac* repressor binds lactose, it changes to its inactive configuration and dissociates from its operators (of which there are three, two of which are bound by the repressor forming a loop of DNA in between). Thus the presence of lactose prevents the repressor from binding and thus allows the operon to be transcribed. It is only when CAP/CRP binds cAMP, however, that it is converted to its active configuration, allowing it to bind DNA. Hence the presence of cAMP, which is an indicator that glucose is in short supply and other pathways are needed to supplement it, allows the activator to bind and thus stimulate transcription. It is not always this way round. Sometimes binding of a metabolite activates a repressor, or inactivates an activator. In the presence of the repressor, transcription is almost completely turned off. In the absence of both the activator and the repressor, transcription is increased 20-fold. However, when the activator is present and the repressor is not, transcription increases to 1000-fold its basal level. Sections of DNA recognised and bound by repressors are called **operators**, while sections recognised and bound by activators don't seem to have a name in prokaryotes, which seems a little weird, but that's biochemistry for you.

7.2 Regulation in eukaryotes

Much is still unknown about the regulation of gene expression in eukaryotes. In many cases molecular biologists have been reduced to noting correlations between transcriptionally active or inactive genes and features of the genome at that point as causation has not been established. As with prokaryotes much regulation appears to happen at the level of transcription. The first thing to notice, however, is that whereas prokaryotic genes generally have a significant basal level of transcription, eukaryotic genes are simply not expressed without the presence of enhancers. This restriction of transcription may be the result of the closer association of eukaryotic DNA with protein in the form of chromatin. There are two types of chromatin found in interphase (non-mitotic) cells. Highly condensed regions of chromatin associated with structures like centromeres is called **heterochromatin**. Transcription seems to be totally suppressed in these regions. The remaining chromatin, called **euchromatin**, is less compact and allows DNA binding proteins access. Even in euchromatin transcription does not always occur. Regions which *are* transcriptionally active generally lack the H1 histone, thus leaving loops of DNA free between adjacent nucleosome core particles. In addition, the core histones are extensively modified by methylation of *Lys* residues, phosphorylation of *Ser* or *Thr* residues, acetylation of *Lys* residues and ubiquitination. Although the way that these modifications alters gene expression is still not clear, some patterns have arisen. For example, increased acetylation of the histones by histone acetyltransferases[10] (HATs) in the region of a gene seems to upregulate a gene, possibly by reducing the positive charge of the histone, making it bind DNA less tightly. Increased methylation, on the other hand, seems to cause downregulation though again the mechanism is unclear. Finally, as we discussed in Section 4.3, DNA itself is subject to methylation. The DNA of expressed genes seems to be hypomethylated with respect to

[10]The notes call these simply histone acetylases, but that does not seem to be the accepted name for these proteins.

unexpressed regions.

Molecules such as HATs are referred to as **co-activators**, and together with the basal transcription factors discussed in Section 5.3 make up two out of the three classes of proteins which affect transcriptional control in eukaryotes. The third class, consisting of activators and repressors, bind to enhancers and silencers respectively, which are sequences of DNA in the region of the gene. Unlike promoters, these do not have a fixed location with respect to the start point of the gene and can actually be several thousand nucleotides upstream, or even within the gene itself (usually in an intron). These effect transcription in a similar way to CAP/CRP and the *lac* repressor discussed in the previous section. These activators and repressors effect transcription initiation by looping the DNA between them and the start of the gene so as to bind their recognition sequence and also effect initiation which, if the DNA were linear, would be happening some distance away.

Having discussed the general pattern of eukaryotic transcriptional regulation, we now introduce several important examples from the literature. We have already met the cyclic AMP response element binding protein (cREB) when we discussed β-adrenergic receptors in the cell signalling part of the course. There is also a protein in this pathway called CBP (cREB binding protein), which is a HAT. It binds cREB only when cREB is phosphorylated and increases transcription by acetylating histones in the nucleosome core particles the gene is wrapped around.

Another example is that of steroid hormones[11] and vitamin D_3[12]. These are hydrophobic molecules which must be carried round the blood stream on carrier proteins since they will not dissolve. Their hydrophobicity becomes advantageous when they reach cells, as they can easily diffuse across the hydrophobic core of the plasma membrane. Once inside the cell, they bind to receptors[13]. Receptor-hormone complexes then form **dimers** and diffuse into the nucleus if they are not there already, where they bind to stretches of DNA known as **hormone response elements** (HRE). In addition to their ligand (hormone) binding domain, each receptor has a consensus sequence (which binds to the HRE) and a variable region, which binds to a sequence of DNA unique to the hormone (otherwise all steroid hormones would regulate the same genes). While steroid receptors form dimers, other nuclear receptors (for vitamin D_3, etc.) form either homodimers like steroid receptors, heterodimers, or sometimes even simply act as monomers.

Other examples in the notes are MyoD, which controls expression of muscle-specific genes; Oct-2, which regulates expression of immunoglogulins; c-Fos, an oncogene; and p53, a tumour suppressor gene. We'll say plenty about these last two later when we discuss cell cycle regulation.

The final example of transcription factors in the notes is that of homeobox genes, which are expressed in embriogenesis and appear to be well conserved across widely separated higher eukaryotes such as fruit flies, mice and humans. Their expression coordinates

[11]Steroid hormones include sex hormones, glucocorticoids such as cortisol and mineralocorticoids such as aldosterone. All three are secreted by cells in the adrenal cortex (though not exclusively)

[12]Although not steroids, retinoids and thyroid hormones have the same mode of action so can also be considered as part of this class of transcriptional modifiers.

[13]When these receptors have not bound their ligands they form multi-protein complexes. Upon binding their ligand they dissociate from the complex. In the case of glucocorticoid receptors, a protein called HSP90 (a heat shock protein with a molecular weight of about 90kDa) forms part of the complex.

differences in the main head-to-tail axis of animals, and they appear to be transcription factors. Unfortunately this is such a complex area that not much seems to be known beyond this simple fact.

7.3 Control of translation

Although transcriptional control seems to be the dominant form, translational control does occur, and falls into two categories, global and selective. Global control regulates the rate of translation in general, while selective control modifies the rate of translation of specific mRNA transcripts. One example of global control is the phosphorylation of the eukaryotic initiation factor eIF-2. Recall that in prokaryotes the initiation factor IF-2 aids binding of initiator aminoacyl-tRNA (fMet-tRNA$^{\text{fMet}}$ in the case of prokaryotes). There is a eukaryotic initiation factor which performs the same role (though at a slightly different point in the sequence, see Section 6.2 for details) which is called eIF-2[14]. This protein is regulated by phosphorylation, which inhibits its function, preventing it from being recycled from its GDP bound state to its active GTP bound state[15]. An important example of this is the hemin controlled repressor, which is an eIF-2 kinase.

An example of selective control of translation is given by the iron response element (IRE), which regulates a number of genes involved in iron homeostasis of which we describe two: ferritin and transferrin. The iron response element is a sequence of around 30 nucleotides which form a stem loop structure and is present in the 5' UTR of ferritin mRNA and the 3' UTR of transferrin. It is bound by a protein called aconitase[16], also known as iron regulatory protein 2 (IRP2). Aconitase binds the IRE in its native configuration. When it binds iron, however, a conformational change is triggered which causes it to dissociate from the IRE. What is different about these two examples is that while binding of aconitase to the IRE in ferritin inhibits translation, binding of aconitase to the IRE in transferrin *activates* translation. The reason for this is simple. Iron is highly toxic in its unbound form but essential to cellular function, as many cellular enzymes have heme groups or iron-sulphur cores (see metabolism notes). In order to solve this conundrum the protein ferritin binds free iron, thus providing a safe store of iron. The cell must also have a way of getting additional iron when it is short on supplies. The enzyme transferrin is a membrane-bound iron transporter. Thus when the cell is short of iron, not much aconitase will have iron bound to it and it will thus bind the IRE and stimulate translation of transferrin while inhibiting translation of ferritin, enabling more iron to be brought into the cell and preventing it from being mopped up by additional ferritin. When there is an excess of free iron, however, some of it will bind aconitase and cause it to dissociate from the IRE, thus activating translation of more ferritin to mop up the superfluous iron while inhibiting translation of transferrin so that more iron does not get brought into the cell. The mode of regulation is also different in these two examples. While ferritin translation is

[14]Well who'd have thought it — names for proteins which are broadly similar between prokaryotes and eukaryotes!

[15]Isn't it funny how all of the translation proteins seem to be GTPases?

[16]The cytosolic aconitase discussed here should not be confused with mitochondrial aconitase, the citric acid cycle enzyme discussed in the metabolism notes. Although both catalyse the conversion of citrate to isocitrate and contain 4Fe-4S centres, only cytosolic aconitase binds IREs (*PNAS* **88** pp. 10109–10113).

inhibited by preventing binding of eIF4F in the initiation complex, transferrin translation is activated by promoting stability of the mRNA. When aconitase is not bound to the IRE in the 3' UTR of the transferrin mRNA, endonucleases quickly degrade the mRNA. However when aconitase is bound the endonucleases cannot and thus the mRNA is not degraded, resulting in greater translation.

7.4 RNA interference

RNA interference (RNAi) is a special method of highly selective translational control which silences genes by interacting with mRNA. It may have evolved as a protection against RNA viruses, as it allows the cell to prevent them from translating their genomes into working proteins. This is done via small double-stranded RNA. An enzyme called DICER snips short interfering RNAs (siRNA) from longer double stranded RNAs made by

1. self-copying gene sequences,
2. replicating viruses, or
3. regulatory RNA sequences known as microRNAs.

All of these can suppress gene expression. The short double stranded siRNAs unwind into single stranded RNAs, which then combine with several proteins to form an RNA-induced silencing complex (RISC). The RISC then binds to native mRNA by complementarity. This prevents the mRNA from being translated in one of two ways. If the complementarity is perfect, or near-perfect, the native mRNA is cleaved into little bits and degraded. If the pairing is imperfect, the RISC-mRNA complex simply prevents ribosomes from moving along the mRNA. Imperfectly binding interfering RNAs are called micro RNAs, or miRNAs. Perfectly binding interfering RNAs are called small interfering RNAs, or siRNAs.

Although this is the current theory on the mechanisms of RNAi, no-one really understands how many of these there are, how exactly they work and where they came from (evolutionarily). However it is clear that some act as tumour suppressors by silencing oncogenes, and at least one acts as an oncogene by silencing a tumour suppressor.

The example given in the notes is the 13q31 locus, which is amplified in some lymphomas and other tumours. The *c13orf25* gene is contained within this locus, and can encode seven miRNAs. Expression of this group together with *myc* accelerates tumour growth. Bizarrely, *myc* itself seems to activate transcription of two of these miRNAs by binding to the gene. These two miRNAs inhibit expression of the transcription factor $E2F_1$, which promotes cell cycle progression, while *myc* activates transcription of $E2F_1$. Thus *myc* activates transcription of $E2F_1$ but inhibits its translation through RNAi!

7.5 Protein degradation

In addition to degradation in lysosomes (chaperone-mediated autophagy of proteins containing the KFERQ motif in starvation, or macroautophagy) the main method of degradation of proteins in eukaryotes is the ubiquitin pathway. Ubiquitin, a 76 amino acid protein,

is added to the ε NH_2 group of lysine side chains by a process involving three types of enzyme called E1, E2 and E3. E1 uses energy from ATP hydrolysis to add ubiquitin to one of its SH groups. E2 then swaps it onto one of its, and E3 swaps it onto the ε NH_2 of a lysine on the target protein. More ubiquitin molecules may then be added to the first ubiquitin by the same process, forming a chain. Presumably the longer the chain the higher the chances of it being recognised by the degradation machinery and thus the quicker the protein is degraded. Ubiquitin peptidases also exist which remove ubiquitins and thus switch off degradation.

The degradation machinery alluded to above is called the proteasome and is a 26S structure composed of many different proteins. It looks kind of like a dustbin with a lid at both ends. The "lids" are called the 19S proteasome, while the "bin" part is called the 20S proteasome. Degradation proceeds by the four steps of **recognition** of ubiquitin; **dissociation** of the ubiquitins and ATP-dependent unravelling of the protein; **translocation** of the protein into the "bin" part; and **destruction** of the protein into 8 amino acid peptides.

There are also specific sequences of residues which aid destruction, and the N-terminal amino acid plays a large part in determining speed of degradation, so addition of an amino acid to the N-terminal end of a protein can vastly change its degradation rate.

8 The cell cycle

Now that we have discussed the basic mechanisms by which cells function, we move on to the topic of cell proliferation. There are five stages in the cell cycle.

M-phase is mitosis phase, where cell division occurs. This stage is further subdivided into prophase, metaphase, anaphase, telophase and cytokinesis.

G_0-phase is not really part of the cell cycle. It is the phase which terminally differentiated cells inhabit.

G_1-phase is the first gap phase, between mitosis and S-phase.

S-phase is synthesis phase, where DNA duplication occurs.

G_2-phase is the second gap phase, between S-phase and mitosis.

8.1 Cyclins and their dependent kinases

The proteins whose activity defines the cell cycle are called **cyclins**. Each of the ten or so cyclins present in mammalian cells is active at a different point in the cell cycle. We will only mention cyclins A, B, D (1, 2 and 3) and E. Cyclins are regulatory subunits of protein kinase complexes. The kinase subunits of these complexes are called rather obviously **cyclin-dependent kinases** or cdks.

The most important cyclin/cdk complex in metaphase is the cyclin B/cdk1 complex[17]. It phosphorylates just about everything that has anything to do with mitosis:

- Nuclear lamins — causing depolymerisation and thus breakdown of the nuclear envelope.
- Condensin — a multiprotein complex which causes chromosome condensation.
- Histones H1 and H3 — not sure why this promotes mitosis as I would have expected it to cause the DNA to unwind from round the nuclear core particles (see Section 2.2 for details).
- microtubule-binding proteins such as kinesin-related motor proteins — promoting spindle formation.
- Golgi matrix components — causing fragmentation of the Golgi apparatus.
- APC/C — the anaphase promoting complex, or cyclosome, which we discuss in some detail below.

While cyclin B/cdk1 is important in M phase, the combination of cyclin D1 with cdk4 is important in regulating entry to S phase, though for obvious reasons it is actually active in G_1 phase rather than S phase itself. Its importance relates to the retinoblastoma protein Rb. When Rb is not phosphorylated it binds a transcription factor called E2F-1. This prevents E2F-1 from binding DNA and thus activating its genes for transcription. E2F-1 controls transcription of over 200 genes, most of which are crucial for DNA synthesis (e.g. orc1, mcm2–7 — see Table 1). Other important complexes are cyclin D2/cdk4 and cyclin D3/cdk6 both again in G_1; cyclin E/cdk2 in the transition from G_1 phase to S phase; and cyclin A/cdk2, which is active from part way through S phase, peaks in prophase and is then rapidly degraded so that its activity is negligible in metaphase and beyond. Cyclin E/cdk2 also phosphorylates Rb, thus sustaining transcription of E2F-1 genes through the transition into S phase.

8.2 The anaphase promoting complex

It is interesting to note that the cyclin B/cdk1 complex effectively triggers its own destruction, as APC/C is responsible for ubiquitinating proteins to target them for destruction in the proteasome as discussed in Section 7.5 above. It is cyclin B which is ubiquitinated and then degraded, and this degradation causes the inactivation of cdk1. There must be some underlying phosphatase activity going on which is overwhelmed by cdk1 when active, as upon deactivation the nuclear lamins immediately start polymerising again (causing the nuclear envelope to re-form) and the chromosomes decondense. Timing is critical here, as APC/C has some other roles to play. A multiprotein complex called cohesin (which is similar to condensin but has a different role) holds sister chromatids together at the metaphase plate. This must be broken down by a protease called separase in order to

[17]While some cyclin/cdk complexes are activated simply by the subunit coming together, cyclin B/cdk1 requires further de-phosphorylation by cdc25 to make the ATP binding site accessible.

ensure equal division of chromatids to daughter cells. However separase is bound by an inhibitor called securin, preventing it from carrying out its role. This is where APC/C comes in again — it ubiquitinates securin, which is then degraded, freeing separase to cleave cohesin and allowing chromatids to be pulled to opposite poles of the cell. It is not entirely clear how these processes are controlled so that APC/C only begins ubiquitinating when all sister chromatids are securely attached at the metaphase plate, but it may have something to do with kinetochore proteins like MAD2, Bub1 and BubR1. What is clear is that a critical cofactor called cdc20 must be released, which binds to APC/C, creating the fully active subunit. Phosphorylation by cyclin B/cdk1 seems to allow this binding to take place, creating an additional checkpoint before the cell cycle can progress to anaphase.

8.3 Cell cycle checkpoints

There are many cell cycle checkpoints so we will discuss only a few of them. Firstly, INK4 proteins inhibit cdk4 and cdk6 and thus halt the cycle in G_1. Secondly, p21, p27 and p57 inhibit cdks 2, 4 and 6 and can thus halt the cycle in phases G_1, S or G_2.

The main protein which controls the activity of these around the entry to S phase is p53. The concentration of p53 in the nucleus is mainly controlled by its binding of the protein MDM2, whose transcription is activated by p53 itself. Binding of MDM2 facilitates degradation of p53, thus lowering its concentration. The affinity that p53 has for MDM2 (or perhaps that should be "the affinity that MDM2 has for p53") is regulated by several phosphorylation sites on p53, which are phosphorylated in response to events such as double-stranded DNA breaks (see Section 4.4 for details of how these are repaired). The protein ATM (Ataxia Telangiectasia Mutation) is responsible for this phosphorylation. Phosphorylation *decreases* affinity in this case. Another protein responsible for decreasing the affinity of p53 for MDM2 is p14 ARF (transcribed from an alternative reading frame of the INK4 gene). One method by which p53 affects a halt in the cell cycle is by increasing the transcription of p21. P21 binds the cyclin E/cdk2 complex, preventing it from phosphorylating Rb and thus inhibiting transcription of the E2F-1 genes responsible for DNA replication.

While p53 controls entry to S phase, kinases chk1 and chk2 regulate the entry to M phase by inhibiting cdc25. These are also activated by ATM, and also by another protein called ATR.

Finally, as mentioned above, entry into anaphase is delayed until all the sister chromatids have lined up on the metaphase plate by delaying activation of APC/C.

8.4 Growth factors

As mentioned above, most cells spend their time in the G0 phase of the cell cycle and are thus not actively dividing. Re-entry into phase G_1 is controlled by extracellular growth factors. Most of these are soluble, although some are embedded in the cell membrane and thus mediate their actions through direct cell-to-cell contact. Examples of growth factors are EGF, FGF, HGF, PDGF and CSF-1. The receptors for these growth factors are usually receptor tyrosine kinases. One notable exception is the receptors for the transforming growth factor β (TGFβ) family of cytokines, which are serine/threonine receptor kinases.

There are two important receptor tyrosine kinase pathways to consider. The first we have already met in the cell signalling notes — the PI3K pathway. The second is the ras pathway. A growth factor binds to the receptor tyrosine kinase, which recruits grb2 via its SH2 domain. Grb2 in turn recruits sos, which activates a GTPase called ras. In the course of activating raf, ras activates its GTPase activity, converting GTP to GDP. Raf in turn activates MAPKK, also called MEK, which activates MAPK. Finally, MAPK activates a dimeric protein called AP-1, which is made up of the two peptides fos and jun. AP-1 is a transcription factor which activates lots of genes which are active in the early phase of response to a growth factor. The most important of these is myc. The PI3K and ras pathways are shown in Figure 1.

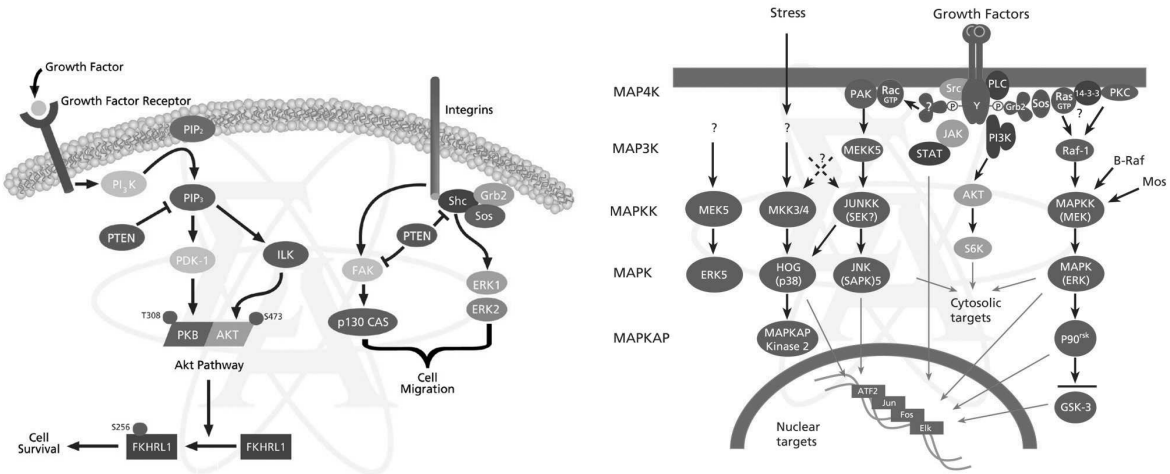

Figure 1: The PI3K and MAPK pathways. The important bits are on the left hand side of the left hand figure and the right hand side of the right hand figure.

A third important pathway is the wnt pathway. Wnt is a paracrine (locally mediated) growth hormone which is active in development, as well as in some forms of cancer. This is not a receptor tyrosine kinase pathway. Although the exact mechanism of how the receptor works is as yet unclear, a seven-membrane-spanning receptor called frizzled which is *not* a G-protein coupled receptor binds wnt and recruits a second receptor called LRP (LDL receptor related protein). Together, by an unknown mechanism, they activate an intracellular signalling protein called dishevelled, which mediates the effects of wnt. When there is no wnt signal (and thus no active dishevelled), GSKβ phosphorylates a protein called β-catina, which marks it for degradation. In order to perform this phosphorylation, GSKβ must be bound to a protein called APC (named because it is mutated in adenomatous polyposis coli). This is achieved by binding of the two together and to a scaffold protein called axin. It seems that APC helps the complex to bind β-catina, but it is GSKβ which catalyses the phosphorylation. In the presence of active dishevelled, however, GSKβ is inhibited, so β-catina is not marked for degradation. This will cause the concentration of β-catina to rise, causing it to migrate to the nucleus (it is usually bound to E-cadherins on the inside of the cell membrane) and bind a transcription factor called LEF-1/TCF. This

displaces an inhibitory transcription factor called Groucho and activates transcription of a number of genes, the most important of which is myc.

The C-terminal domain of APC contains binding sites for the plus-end of microtubules, causing APC to localise to kinetochores, where it probably interacts with proteins such as Bub1. This seems to be important, as cells with mutated APC genes tend end up segragating their chromosomes unequally to daughter cells (so-called chromosome non-disjunction).

8.5 Apoptosis

9 Tumour development and angiogenesis

One of the most important things limiting tumour growth is the availability of nutrients. Most people seem to have a lot of microtumours which grow to about 10^5 cells big and then stop as they have now grown to the limit of what the current vascular system can support. Thus tumours must vascularise to grow. Another purpose of tumour vascularisation is to allow metastatic cells to travel to other parts of the body in the blood and lymphatic systems. Tumour-induced angiogenic factors include basic fibroblast growth factor (bFGF), vascular endothelial growth factor (VEGF) and matrix metalloproteases (MMPs). These factors promote breakdown of the basement membrane and growth of new blood vessels towards the source of the growth factor (the tumour).

Competing with the rate of tumour growth is not only avascular necrosis but also apoptosis. Labelling of the broken ends of DNA found in apoptosis reveals that apoptotic cells are present in most tumours. Thus it is only if there are sufficient mutations in the apoptotic and angiogenic pathways that a tumour may grow. As the tumour grows it will recruit macrophages and leukocytes associated with the inflammatory response. These may be stimulated to produce growth factors and thus accelerate tumour growth. As the tumour increases in size it will often become polyclonal as further mutations are added in subsets of the cells. Those which favour the most rapid growth will then contribute most to the growth of the tumour.

As mentioned above, the progress of the tumour to the metastatic stage is characterised by the expression of proteases which degrade the basement membrane and allow tumour cells to enter the blood or lymphatics and spread to distant parts of the body where they encounter areas that allow them to exit the vasculature and invade a foreign tissue. Some types of cancer seem to spread preferentially to certain parts of the body, though the reasons for this are not clear. For example, breast tumours tend to spread first to lymph nodes, then bone, liver and lung.

As mentioned above, tumours must stimulate angiogenesis to proliferate. It turns out that in addition to the angiogenic factors mentioned above, there are several angiostatic factors. This was discovered after the observation that removal of a primary tumour often prompted secondary metastases to grow and proliferate quickly. After investigation it was found that the primary was expressing high levels of a protein called angiostatin (probably as a homeostatic mechanism in response to the heightened levels angiogenic factors). Angiostatin is a fragment of plasminogen. Since then other angiostatic factors

have been discovered such as endostatin, which is derived from collagen XVIII. The gene for collagen XVIII is on chromosome 21. The trisomy 21 which causes Down's syndrome also results in elevated levels of endostatin. Astonishingly people with Down's syndrome almost never acquire cancer, implying that endostatin is enormously protective.

In summary, the balance between amounts of angiogenic factors such as VEGF, TNF-α, PDGF-BB, bFGF and MMPs is balanced by angiostatic factors such as angiostatin, endostatin, thrombospondin-1 and interferon-α. Excess production of angiogenic factors seems to be partially compensated for by increased production of angiostatic factors. It is only when this compensatory mechanism is overwhelmed that tumours can grow and metastasise.

10 Cancer causing mutations

Cancer genes fall into three distinct classes.

1. **Proto-oncogenes**, the normal products of which are components of signalling pathways which regulate proliferation and which, in their mutated form, become **dominant** oncogenes (i.e. mutation of only one copy is enough to stimulate tumour growth).

2. **Tumour suppressor genes**, which are generally check-points along the road of cell cycle progression, cellular adhesion, etc. and thus loss of both copies of which will result in a lack of control over the cell cycle. In other words, tumour suppressor genes generally exhibit **recessive** effects.

3. **DNA repair enzymes**, which when mutated will reduce the stability of the genome. These are discussed in detail in Section 4.

In addition, epigenetic factors also influence cancer development. Tumour cells' DNA is hypomethylated compared to normal cells' DNA, though gene silencing by promotor hypermethylation is at least as common as mutational inhibition of tumour suppressor gene function.

10.1 Oncogenes

Oncogenes were first identified in retroviruses. So far no retrovirus has yet been shown to be directly oncogenic in humans, though it is probable that subversion of normal cellular control mechanisms by the transcription factors present in the genomes of human T-lymphotropic viruses (HLTV-1 and HLTV-2) or HIV may cause latent development of cancer. In addition, certain transforming herpes viruses encode a cyclin D homolog which complexes with cdk6 but does not respond to p16INK4 or p21, causing the G_1–S checkpoint to fail to function.

There are a variety of mechanisms by which proto-oncogenes may become true oncogenes. In virally-induced cancers mutation may occur during acquisition of the protooncogene by the virus, turning it into an oncogene. Alternatively, the insertion of viral promot-

ers or enhancers into the cellular genome may cause the transformation of proto-oncogenes already present to become oncogenic.

In human cancers there are a number of additional ways that proto-oncogenes may become oncogenic.

10.1.1 Mutations

These take several forms: point mutations of one or a few bases; deletions; or insertions. They occur from the action of the mutagens discussed in Section 4. One example of this is *Ras*, which can be transformed into an oncogene by substitution of a single base, commonly replacing glycine with valine at residue 12 or lysine with glutamine at residue 61. This modifies the GTP-binding domain, making it permanantly active. Another example is that of *Kras* missense mutations, which are present in 80% of pancreatic tumours. Finally, truncation mutations (changing a normal amino acid codon to a stop codon) which remove the extracellular (ligand-binding) domain of the EGF receptor produce a molecule which signals continuously.

10.1.2 Aneuploidy

Aneuploidy is the term given to loss or gain of a whole chomosome. An example is loss of chromosome 10 in glioblastomas. In addition, the genes *Rb1*, *p53* and *Apc* are frequently lost in a range of human cancers, though these are tumour suppressor genes not oncogenes, so I'm not sure why this is in this section... Who said these notes were logical?

10.1.3 Translocation

Sometimes a piece of a chromosome is translocated to a location on another chromosome. This can cause a whole host of weird and wonderful changes, from fusion proteins to over- or under-expressed proteins due to the novel location of promoters and enhancers in these patchwork chromosomes. Usually the occurance of chromosomal translocations is purely random, but there are consistent rearrangements in a group of leukemias and lymphomas which have been identified.

10.1.4 Gene amplification

For some reason, the amplification of specific sequences of DNA (sometimes several megabases) occurs frequently in tumour cells, though it has not been found in normal mammalian cell development. Presumably this involves failures to the mechanisms mentioned at the end of Section 3 that ensure that DNA is only copied once. This duplication will obviously result in an increase in the copy number of genes encoded in the amplified segment. This amplification is sometimes enough for proto-oncogenes to become tumour-promoting. Examples where this has been found to happen include *EGFR*, *myc*, *ras* and the 11q13 locus.

10.2 Tumour suppressor genes

It is estimated that 10% of all cancers result from inheritance of a mutant form of a tumour suppressor gene. The best characterised examples of tumour suppressor genes are *Rb1* and *p53*. Both of them exert a controlling influence on the G_1–S transition.

As we saw in Section 8.1, the effects of *Rb1* are mediated through its inhibition of transcription factors that are required for the expression of genes involved in DNA replication (E2F proteins), but it also mediates transcriptional activation (by RNA polymerase I, II and III). *Rbl1* and *Rbl2* are related genes with overlapping functions.

The *p53* gene was the first tumour suppressing gene to be identified and is known as the "guardian of the genome" because it prevents entry to S phase if there is damage to the DNA. Indeed it may even divert the cell to apoptosis. Its protein product binds to DNA as a tetramer, so the gene does not exhibit totally recessive behaviour. Mutations of *p53* occur with high frequency in lung cancer, 60% of breast cancers and around 40% of brain tumours. In addition, the region deleted in most colorectal neoplasms (whatever that is!) includes *p53*. Familial inheritance of a mutated copy of p53 leads to a condition called **Li-Fraumeni Syndrome**

11 Some cancer statistics

- A typical human cancer cell will contain around 10,000 mutations.
- Around 200 of these will be in protein-coding sequences ($< 1\%$ of the genome).
- It is estimated on p. 8 of the prologue notes that the number of mutated oncogenes or tumour suppressor genes needed for a tumour to form is between 2 and 20. On p. 11 however, the range is given as 5 to 15. Confusing, huh?!!
- About 200 oncogenes and 50 tumour suppressor genes have been identified.
- Early stage tumours are typically monoclonal, having derived from a single cell which acquired the necessary mutations.