Mutation

CAUSES OF MUTATION

Mutations are random changes to the base sequence of a gene. A mutation that replaces one base in a gene with a different base is a base substitution. Mutations are important as they are a source of the genetic variation that is necessary for evolution to occur, but very few mutations prove to be beneficial and some cause genetic diseases or cancer. The mutation rate is increased by two types of mutagen:

- **high energy radiation** including X-rays, short or medium wave UV, gamma rays and alpha particles from radioactive isotopes
- **mutagenic chemicals** such as nitrosamines in tobacco, mustard gas that was used as a chemical weapon and the solvent benzene.

Because mutagens increase the mutation rate they are a cause of both genetic diseases and cancer.

The effects of radiation can be studied using two incidents, the nuclear accident at Chernobyl and the nuclear bombing of Hiroshima. The common feature of these incidents is that radioactive isotopes were released into the environment and as a result people were exposed to potentially dangerous levels of radiation. Chernobyl released far more radioactive material but will probably have caused fewer deaths than Hiroshima, because the isotopes released were spread over a wider area and have longer half-lives so the doses of radiation have been spread over a longer period.

NUCLEAR BOMBING OF HIROSHIMA

The atomic bomb that was detonated over Hiroshima in 1945 killed 90,000–166,000; people either died directly or within a few months. The city was devastated with few buildings remaining.

The health of a large group of survivors of both the Hiroshima and Nagasaki nuclear bombs has been followed since then by the Radiation Effects Research Foundation in Japan. There have been long-term effects from the radiation with increased deaths due to cancer. The larger the dose of radiation received by a survivor, the higher the risk of both leukemia and other cancers.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Leukemia</td>
</tr>
<tr>
<td>&lt;0.005</td>
<td>0.25</td>
</tr>
<tr>
<td>0.005–0.1</td>
<td>0.23</td>
</tr>
<tr>
<td>0.1–1.0</td>
<td>0.44</td>
</tr>
<tr>
<td>≥1.0</td>
<td>2.36</td>
</tr>
</tbody>
</table>

Most of the deaths due to leukemia occurred in the first ten years after the bombing, but deaths due to other cancers have continued.

Apart from cancer the other main effect of the radiation that was predicted was mutations, leading to stillbirths, malformations or death. The health of 10,000 children that were fetuses when the atomic bombs were detonated has been monitored. No evidence has been found of mutations caused by the radiation. There are likely to have been some mutations, but the number is too small for it to be statistically significant even with the large numbers of children in the study. Despite this lack of evidence of mutations due to the atomic bombs, survivors have sometimes felt that they were stigmatized. Some found that potential wives or husbands were reluctant to marry them for fear that their children might have genetic diseases.

NUCLEAR ACCIDENT AT CHERNOBYL

The accident at Chernobyl, Ukraine, in 1986 caused explosions and a fire in the core of a nuclear reactor. Radioactive iodine-131, caesium-134 and caesium-137 were released and spread over large parts of Europe. About six tonnes of uranium and other radioactive metals in fuel from the reactor were broken up into small particles by the explosions and escaped.

- 28 workers at the nuclear power plant died from the effects of radiation within three months. There have also been increased rates of leukemia in other workers exposed to high radiation doses.
- Concentrations of radioactive iodine in the environment rose and resulted in drinking water and milk with unacceptably high levels. Iodine is absorbed by the thyroid gland. More than 6,000 cases of thyroid cancer can be attributed to the radioactive iodine released. Horses and cattle near the plant died from damage to their thyroid glands.
- Bioaccumulation caused high levels of radioactive caesium in fish as far away as Scandinavia, Germany and Wales. Consumption of lamb contaminated with radioactive caesium was banned for many years in some areas due to the long half-life of caesium-137.
- There will almost certainly have been a small increase in the risk of cancer and genetic disease for large numbers of people in Europe due to radiation from Chernobyl, but it is hard to prove this.
- 4 km² of pine forest downwind of the reactor turned ginger brown and died due to high doses of radiation, but in the absence of humans some wildlife such as lynx and wild boar have thrived.

Hiroshima Peace Memorial
**GENES AND CHROMOSOMES**

*Genetics* is the study of variation and inheritance. The basic unit of inheritance is the gene.

A gene is a heritable factor that consists of a length of DNA and influences a specific characteristic.

Every gene occupies a specific position on a chromosome. For example, in humans the gene for making the beta polypeptide of hemoglobin is located near the end of the short arm of chromosome 11.

- **locus of gene for β polypeptide of hemoglobin**
- **centromere**
- **short arm of chromosome 11**
- **long arm of chromosome 11**

**NUMBERS OF GENES**

A typical animal or plant cell nucleus contains thousands of genes.

The total number of genes is not yet known precisely for humans or other species but these are current estimates:

- *Homo sapiens* (humans) 23,000
- *E. coli* (a gut bacterium) 3,200
- *Drosophila melanogaster* (fruit fly) 14,000
- *Takifugu gambiae* (puffer fish) 25,500
- *Oryza sativa* (rice) 41,000

These estimates illustrate some trends:

- bacteria have fewer genes than eukaryotes
- some other animals have fewer genes than humans but some have more
- plants may seem less complex than humans but some have more genes.

**ALLELES**

A gene consists of a sequence of bases on a piece of DNA. There are different versions of some genes that have almost the same base sequence but differ in just one or a very small number of bases. These variant forms of a gene are called alleles.

Alleles are different forms of the same gene because they influence the same characteristic, occupy the same position on a type of chromosome and have base sequences that differ from each other by one or only a few bases.

For example, in the human gene for the beta polypeptide of hemoglobin, the second base in the sixth codon of the gene is adenine [A] in the commonest allele of the gene. There is a less common allele in which this base is thymine [T]. This allele causes the genetic disease sickle cell anemia.

**SICKLE-CELL ANEMIA**

This is a genetic disease that demonstrates how a single base substitution mutation can have very significant consequences. The mutation occurred in HBβ, the gene for the beta polypeptide of hemoglobin. This polypeptide consists of 146 amino acids.

- **Part of HbA**
- **Base substitution**
- **Part of HbS**

<table>
<thead>
<tr>
<th>Effect on the phenotype</th>
<th>Effect on the phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>In some conditions red blood cells containing the altered hemoglobin become sickle shaped</td>
<td>Sickle cells may carry oxygen less efficiently but can give resistance to malaria</td>
</tr>
<tr>
<td>Normal red blood cells that carry oxygen efficiently but are affected by malaria</td>
<td></td>
</tr>
</tbody>
</table>

The allele HbS that causes sickle-cell anemia has become quite common in some parts of the world affected by malaria. In these regions the malaria resistance that it causes is an advantage.
Gene sequencing

USING DATABASES
Databases have been developed since the 1960s to help store complex information and let researchers get access to it. They are ideally suited to storing the vast amount of base sequence data that is currently being generated by genome research. Exponential increases in the amount of data needing to be stored have been matched by increases in database capacity.

There have also been great improvements in the ease with which users can search databases and extract data. The development of the internet opened up access to databases from anywhere in the world, so data can be shared far more easily. Two uses of gene databases are explained below on this page.

GENE LOCI AND PROTEIN PRODUCTS
The locus of a gene is its particular position on homologous chromosomes. The loci of human genes can be found using the OMIM website (Online Mendelian Inheritance in Man),

- Search for OMIM and open the home page.
- Choose OMIM Gene Map
- Enter the name of a gene into the Search gene map box.

This should bring up a table with information about the gene, including its locus and the protein product of the gene.

Example: entering HBB brings up the information that this gene codes for the beta polypeptide of hemoglobin and the location is 11p15.4. The first number indicates that the gene is on chromosome 11, the letter p that it is on the short arm of this chromosome [q indicates the long arm] and 15.4 which region of the short arm the gene is in.

Comparing base sequences of genes
Changes in the base sequence of genes occur over time. If a species splits to form two separate species, differences between the base sequences of the genes of those two species will gradually accumulate. The number of differences can give an indication of how long ago species diverged from the common ancestral species. It is therefore useful to be able to compare the base sequences of genes and find out how many differences there are. This can be done using base sequence data from the GenBank database and downloadable software.

- Search for GenBank and open the home page.
- Choose 'Gene' from the search menu.
- Enter the name of a gene plus the organism, such as HBB human.
- Move your mouse over the section 'Genomic regions, transcripts and products' until Nucleotide links appears.
- Choose FASTA and the entire base sequence of the gene should appear.

You can then search for similar sequences in the GenBank database using BLAST software. There may be other alleles in the same species or genes with similar sequences in other species.

Alternatively you can compare the sequence with other selected genes (for example HBB in chimpanzees) using ClustalX software which can either be downloaded or used online:

- Copy the whole sequence in the FASTA format (including the symbols starting with >) and paste it into a text file or notepad file.
- Repeat with a number of different species that you want to compare. Copy them to the same file, separating by pressing the return button on your keyboard and saving the file each time.
- Open ClustalX and enter the text file containing the sequences that you wish to compare. A base by base comparison of the sequences should appear with any differences highlighted.

ClustalX (1.83)

Multiple Alignment Mode ▼ Font Size: 10 ▼
DNA technology

PCR – THE POLYMERASE CHAIN REACTION
PCR consists of a cycle of stages carried out again and again to produce many copies of a DNA molecule:

Millions of copies of the DNA can be produced in a few hours. This is very useful when very small quantities of DNA are found in a sample and larger amounts are needed for analysis. DNA from very small samples of semen, blood or other tissue or even from long-dead specimens can be amplified using PCR. [Reasons for the use of Taq DNA polymerase in PCR are described in Topic 2.]

DNA is heated to 95 °C to separate the two strands.
The temperature is increased to 73 °C, which encourages Taq DNA polymerase to replicate both strands, starting at the primer, producing two double-stranded copies of the original DNA.
The temperature is reduced to 53 °C, which allows primers to bind to both strands of the DNA next to the sequence that is to be copied.

GEL ELECTROPHORESIS
Gel electrophoresis is a method of separating mixtures of proteins or fragments of DNA, which are charged. The mixture is placed on a thin sheet of gel, which acts like a molecular sieve. An electric field is applied to the gel by attaching electrodes to both ends. Depending on whether the particles are positively or negatively charged, they move towards one of the electrodes or the other. The rate of movement depends on the size of the molecules – small molecules move faster than larger ones.

GENE SEQUENCING TECHNOLOGY
Many developments in scientific research follow improvements in technology. In some cases research projects stimulate improvements in technology.

Methods for finding the base sequence of genes were developed in the 1970s and the technology has been improved repeatedly since then.

The idea of sequencing the entire human genome seemed impossibly difficult at one time but improvements in sequencing technology towards the end of the 20th century made it possible, though still very ambitious. The Human Genome Project began in 1990 and was expected to take 15 years but improvements in technology continued once the project was underway and draft sequences were completed much sooner than expected in 2000.

Further advances are allowing the genomes of other species to be sequenced at an ever increasing rate and lower cost.

By 2008 the genomes of over a thousand different humans from all parts of the world had been sequenced, to study genetic variation, and by 2012 the cost of sequencing a human genome had dropped below $10,000. By 2014 the genomes of hundreds of prokaryotes had been sequenced and over a hundred eukaryotes. The 1,000 Plant Genomes Project was well on its way towards the planned sequencing of the genomes of a thousand different plant species.
DNA profiling

In the DNA of humans and other organisms there are loci in the chromosomes where instead of a gene consisting of a long sequence of bases there are much shorter sequences of three, four or five bases that are repeated many times. The repeated sequences are called short tandem repeats (STR). At these STR loci there are many different possible alleles that vary in the number of repeats. STR alleles are used in DNA profiling (also called DNA fingerprinting).

1. A sample of DNA is obtained from a person. It must not be contaminated with DNA from anyone else or another organism.
2. DNA from a selection of STR loci is copied by PCR. The DNA from between 11 and 13 loci is copied in commonly used DNA profiling methods. It is very unlikely for two individuals to have the same number of repeats at each of these loci.
3. The copies of STR alleles made by PCR from one person's DNA sample are separated by gel electrophoresis. The result is a pattern of bands. Two individuals are extremely unlikely to have the same pattern of bands unless they are identical twins.

DNA profiling is used in forensic investigations (obtaining evidence to use in court cases) and investigating paternity (who the father of a child is).

Forensic investigations

The first DNA profiles to be used in a forensic investigation (the Enderby double murder case) are shown below. Key:
- a = hair roots from the first victim
- b = mixed semen and vaginal fluids from the first victim
- c = blood of second victim
- d = vaginal swab from second victim
- e = semen stain on second victim
- s = blood of prime suspect

Two bands in track b indicated by arrows must be from DNA in the culprit's semen but are not present in DNA from the prime suspect, who was not guilty despite having confessed to the murders.

Paternity investigations

The DNA profiles of a family of dunnocks (Prunella modularis) are shown above. Dunnocks are small birds found in Europe. The tracks from left to right are: the mother, two resident males that might have been the father of the offspring and four offspring. There are bands in the profiles of three offspring (D, E and F) that are found in the profile of the ( male, but not the ( male or the mother, showing that the ( male fathered them despite being less dominant than the ( male.
GENE TRANSFER USING PLASMIDS

Genetic modification is the transfer of genes from one species to another. Organisms that have had genes transferred to them are called genetically modified organisms (GMO) or transgenic organisms. The transfer of the gene for human insulin to bacteria was outlined in Topic 2. The methods used for gene transfer to bacteria are explained in the flow chart below. Genes are transferred between species using a vector. In this case the vector is a small loop of DNA called a plasmid. Two enzymes are used to insert genes into plasmids: restriction endonucleases cut DNA molecules at specific base sequences and DNA ligase makes sugar–phosphate bonds to link nucleotides together and form continuous strands of nucleotides. A plasmid with a gene from another species inserted is called a recombinant plasmid. The cell that receives the gene is a host cell. In the example below E. coli is the host cell.

1. Messenger RNA coding for insulin is extracted from human pancreas cells that make insulin.
2. DNA copies of the messenger RNA are made using the enzyme reverse transcriptase.
3. Sticky ends are made by adding extra G nucleotides to the ends of the gene.
4. Plasmids are cut open using restriction enzymes.
5. Sticky ends are made by adding extra C nucleotides to the ends of the cut plasmid.
6. The insulin gene and the plasmid are mixed. They link by complementary base pairing [C – G], between the sticky ends.
7. DNA ligase seals up the nicks in the DNA by making sugar-phosphate bonds.
8. The recombinant plasmids are mixed with the host cells (E. coli). The host cells absorb them.
9. The genetically modified E. coli are cultured in a fermenter.
10. The E. coli bacteria start to make human insulin, which is extracted, purified and used by diabetics.

BENEFITS AND RISKS OF GENETIC MODIFICATION OF CROPS

The production of human insulin using bacteria has enormous benefits and no obvious harmful effects. Genetic modification of crop plants is more controversial. An example of this is corn or maize (Zea mays). A gene from a bacterium (Bacillus thuringiensis) has been transferred to some varieties. The gene codes for a bacterial protein called Bt toxin, which kills insect pests feeding on the crop, especially corn borers that can cause serious damage.

Potential benefits of Bt maize
1. Higher crop yields and thus more food for humans, due to less pest damage.
2. Less land needed for crop production, so some could become areas for wildlife conservation.
3. Less use of insecticide sprays, which are expensive and can be harmful to farm workers and to wildlife.

(Possible harmful effects of Bt maize
1. Insects that are not pests could be killed. Maize pollen containing the toxin is blown onto wild plants growing near the maize. Insects feeding on the wild plants, including caterpillars of the Monarch butterfly (Danaus plexippus) are therefore affected even if they do not feed on the maize. Leaves and stems from the crop after harvest still contain the toxin which could harm insect detrivores in the soil and in streams.
2. The transferred gene might spread to populations of wild plants by cross-pollination, making them also toxic to insects feeding on them.
3. The insects pests of corn may develop resistance to the Bt toxin.)
Cloning

**CLONES AND CLONING**
A group of genetically identical organisms derived from a single original parent cell is a clone and production of an organism that is genetically identical to another organism is cloning. Asexual reproduction is a natural form of cloning. Many plant species and some animal species can do this, using mitosis to produce the genetically identical cells required. For example, plants clone themselves by growing extra bulbs, tubers, runners or other structures. Female aphids [greenfly] can give birth to young formed asexually from their own cells.

There are also methods of artificial cloning for both plants and animals. Cloning is very useful if more organisms with a desirable combination of characteristics are wanted.

**ARTIFICIAL CLONING OF ANIMALS**
The simplest method of cloning an animal is to break up an embryo into more than one group of cells at an early stage when it consists entirely of embryonic stem cells. Each group of cells develops into a separate genetically identical individual. The drawback is that at the embryo stage the characteristics of an animal are mostly unknown.

It is much more difficult to clone an adult animal with known characteristics, but methods have been developed. One method is somatic-cell nuclear transfer, in which the nucleus is removed from an egg cell and replaced by a nucleus from a differentiated somatic [body] cell. This method [see below] was used to produce Dolly, the first mammal to be cloned from an adult somatic cell.

**CLONING ADULT ANIMALS USING DIFFERENTIATED CELLS**

**INVESTIGATING FACTORS AFFECTING ROOTING IN STEM CUTTINGS**

Stem cuttings are short lengths of stem that are used to clone plants artificially. If roots develop from the stem, the cutting can become an independent new plant. Some plant species root when the base of the cutting is placed in water but others root better when it is inserted into a solid medium. The diagram [right] shows a basic method for rooting a cutting.

Many factors affect whether the cutting will form roots or not. One of these could be investigated – it is the independent variable. For example, the independent variable could be how many leaves are left on the cutting, whether a hormone rooting powder is used, how warm the cuttings are kept and whether a plastic bag is placed over the cuttings. The dependent variable could simply be whether any roots are formed or not, or to make the investigation quantitative the number of roots could be counted. All other factors that could affect rooting are control variables and must be kept the same. For example, cuttings from the same species of plant should be used for the whole investigation. There should be repeats to make the investigation reliable and avoid anomalous results leading to false conclusions.
DNA AS THE GENETIC MATERIAL

In the early 1950s it was still unclear whether genes were made of DNA or protein. Hershey and Chase used a virus that infects cells of the bacterium *E. coli* to investigate this. Viral proteins start being made in the cytoplasm of *E. coli* soon after the virus comes into contact with it, showing that the viral genes have entered the bacterium. The virus was T2. Viruses such as T2 consist only of DNA inside a protein coat. DNA contains phosphorus but not sulphur, and protein contains sulphur but not phosphorus. Hershey and Chase used this difference to prepare two strains of T2, one having its DNA radioactively labelled with $^{32}P$ and the other having its protein labelled with $^{35}S$.

These two strains of labelled T2 were each mixed with *E. coli*. After leaving enough time for the bacteria to be infected, the mixture was agitated in a high-speed mixer and then centrifuged at 10,000 rpm to separate into a solid pellet containing the bacteria and a liquid supernatant. A Geiger counter was used to locate the radioactivity. The results are shown in the diagram.

Analysis of results

T2 binds to the surface of *E. coli* and injects its DNA into the bacterium. This explains the high proportion of radioactivity with the bacteria in the pellet when $^{32}P$ was used. Agitation shakes many of the protein coats of the viruses off the outside of the bacteria and these coats remain in the supernatant. This explains the very high proportion of radioactivity in the supernatant when $^{35}S$ was used.

The small proportion of radioactivity in the pellet can be explained by the protein coats that remain attached to the bacteria and also the presence of some fluid containing protein coats in the pellet. This and other experiments carried out by Hershey and Chase give strong evidence for genes being composed of DNA rather than protein.

THE HELICAL STRUCTURE OF DNA

If a beam of X-rays is directed at a material, most of it passes through but some is scattered by the particles in the material. This scattering is called diffraction. The wavelength of X-rays makes them particularly sensitive to diffraction by the particles in biological molecules including DNA.

In a crystal the particles are arranged in a regular repeating pattern, so the diffraction occurs in a regular way. An X-ray detector is placed close to the sample to collect the scattered rays. The sample can be rotated in three different dimensions to investigate the pattern of scattering. Diffraction patterns can be recorded using X-ray film.

DNA cannot be crystallized, but in 1950 Maurice Wilkins developed a method of producing arrays of DNA molecules that were orderly enough for a diffraction pattern to be obtained, rather than random scattering.

Rosalind Franklin came to work in the same research department as Wilkins. She developed a high resolution detector that produced very clear images of diffraction patterns from DNA. The figure below shows the most famous of the diffraction patterns that she obtained.

Analysis of results

From this diffraction pattern Franklin was able to make a series of deductions about the structure of DNA:

- The cross in the centre of the pattern indicated that the molecule was helical in shape.
- The angle of the cross shape showed the pitch (steepness of angle) of the helix.
- The distance between the horizontal bars showed turns of the helix to be 3.4 nm apart.

Rosalind Franklin's research is an excellent example of the importance of making careful observations in science. She was painstaking in her methods of obtaining X-ray diffraction images of DNA and in her analysis of the patterns in them. Her observations were critically important in the discovery of the double helix structure of DNA by Crick and Watson.
**DNA replication**

**LEADING AND LAGGING STRANDS**
The two ends of a strand of nucleotides in DNA or RNA are different. They are known as the 3' and 5' ends (3 prime and 5 prime). The 3' end in DNA has a deoxyribose to which the phosphate of another nucleotide could be linked. The phosphate would bond with the -OH group on the C3 of the deoxyribose. The 5' end in DNA has a phosphate that is attached to C5 of deoxyribose.

Nucleotides are linked to the end of a DNA strand during replication by one of a group of enzymes called DNA polymerases. These enzymes always add the phosphate of a free nucleotide to the deoxyribose at the 3' end of the strand. The direction of replication is therefore 5' to 3'.

The two strands in a DNA molecule are antiparallel because they run in opposite directions. Each end of a DNA double helix therefore has one strand with a 3' end and one with a 5' end.

Because of the antiparallel structure of DNA, the two strands have to be replicated in different ways.

- On one strand DNA polymerases can move in the same direction as the replication fork so replication is continuous. This is the leading strand.
- On the other strand DNA polymerases have to move in the opposite direction to the replication fork, so replication is discontinuous. This is the lagging strand.

**ROLES OF ENZYMES IN PROKARYOTIC DNA REPLICATION**
Semi-conservative replication is carried out by a complex system of enzymes. There are differences between prokaryotes and eukaryotes in the mechanism of replication, though the basic principles are the same. The system used in prokaryotes is shown below.

1. **DNA gyrase** moves in advance of helicase and relieves strains in the DNA molecule that are created when the double helix is uncoiled. Without this action the separated strands would form tight supercoils.

2. **Helicase** uncoils the DNA double helix and splits it into two template strands. Single-stranded binding proteins keep the strands apart long enough to allow the template strand to be copied.

3. **DNA polymerase III** adds nucleotides in a 5' to 3' direction. On the leading strand it moves in the same direction as the replication fork, close to helicase.

4. **DNA primase** adds a short length of RNA attached by base pairing to the template strand of DNA. This acts as a primer, allowing DNA polymerase to bind and begin replication.

5. Short lengths of DNA are formed between RNA primers on the lagging strand, called Okazaki fragments.

6. **DNA ligase** seals up the nick by making another sugar-phosphate bond.

7. **DNA polymerase I** removes the RNA primer and replaces it with DNA. A nick is left in the sugar-phosphate backbone of the molecule where two nucleotides are still unconnected.

8. **DNA polymerase III** starts replication next to the RNA primer and adds nucleotides in a 5' to 3' direction. It therefore moves away from the replication fork on the lagging strand.
SANGER SEQUENCING

Frederick Sanger developed a method of base sequencing that was used very widely for 25 years. It is based on nucleotides of dideoxyribonucleic acid (ddNA). These contain dideoxyribose instead of deoxyribose, so have no -OH group on carbon atom 3.

If a dideoxyribose is at the end of a strand of DNA, there is no site to which another nucleotide can be added by a 5’ to 3’ linkage. In the sequencing machine single-stranded copies of the DNA being sequenced are mixed with DNA polymerase and normal DNA nucleotides, plus smaller numbers of ddNA nucleotides. The replication is repeated four times, once with dideoxynucleotides with each base, A, C, G and T.

The fragments of replicated DNA that are produced vary in length depending on how far replication got before it was terminated because a ddNA nucleotide was added to the end of the chain. The fragments are separated according to length by gel electrophoresis with four tracks, one for each base in the ddDNA nucleotide that terminated replication. Each band in the gel represents one length of DNA fragment produced by replication. All the fragments of the same length end in the same base, so there is only one band in one of the four tracks for each length of fragment. This allows the base sequence of the DNA to be deduced quite easily from the gel.

A typical section of gel is shown [right]. Part of the base sequence is indicated.

The whole base sequence can easily be deduced. This was initially done by hand but fluorescent markers were introduced that allowed the base sequence to be read by a machine.

FUNCTIONS OF DNA BASE SEQUENCES

There are thousands of sequences of bases that code for proteins in the DNA of a species. These coding sequences are transcribed and translated when a cell requires the protein that they code for.

There are also non-coding sequences. Some non-coding sequences have important functions.

- **Regulating gene expression** – some base sequences are sites where proteins can bind that either promote or repress the transcription of an adjacent gene.
- **Introns** – in many eukaryote genes the coding sequence is interrupted by one or more non-coding sequences. These introns are removed from mRNA before it is translated. Introns have numerous functions associated with mRNA processing.
- **Telomeres** – these are repetitive base sequences at the ends of chromosomes. When the DNA of a eukaryote chromosome is replicated, the end of the molecule cannot be replicated, so a small section of the base sequence is lost. The presence of the telomere prevents parts of important genes at the ends of the chromosomes from being lost each time DNA is replicated.
- **Genes for tRNA and rRNA** – transcription of these genes produces the transfer RNA used during translation and also the ribosomal RNA that forms much of the structure of the ribosome.
**Bioinformatics**

Computers now allow huge amounts of data to be stored and analysed, allowing the branch of biology called bioinformatics to develop. Base sequences are the main type of data stored and analysed in bioinformatics. Sequencing was at first only possible with short lengths of DNA such as single genes, but now whole genomes can be sequenced and the amount of data generated is growing exponentially.

![Graph: Growth of EMBL-Bank (1982–2012)]

One of the main types of analysis in bioinformatics is locating genes that code for polypeptides within genomes. This is done using computers to search for ORFs (open reading frames). The details of this procedure are described in Option B. Another type of analysis is to search for conserved sequences in the genomes of different organisms. These are base sequences similar enough for them to have been most likely inherited from a common ancestral gene. The conserved sequences are analysed to find differences in the base sequences [see Topic 3]. Classification of living organisms has been revolutionized by these techniques [Topic 5].

**Tandem Repeats**

Within the genomes of humans and other species there are regions where adjacent sections of DNA have the same base sequence. These are called **tandem repeats**.

The length of the repeated sequence can be anything from two bases to 60 or more.

Examples:

- **ACACACAC**
  - two nucleotide repeat (dimeric)

- **GATAGATAGATAGATAGATA**
  - four base repeat (tetrameric)

The number of repeats varies between different individuals with some tandem repeats. These are therefore known as **variable number tandem repeats**.

DNA profiling (fingerprinting) is based on variable number tandem repeats. The methods used are described in Topic 3.

**Nucleosomes**

DNA in eukaryotes is associated with proteins to form nucleosomes. These are globular structures that have a core of eight histone proteins with DNA wrapped around. Another histone protein called H1 binds the DNA to the core. A short section of linker DNA connects one nucleosome to the next.

The eight histones in the core have **N-terminal tails** that extend outwards from the nucleosome. During the condensation of chromosomes in the early stages of mitosis and meiosis the tails of histones in adjacent nucleosomes link up and pull the nucleosomes together. This is part of the process of **supercoiling**.

During interphase, changes to the nucleosomes allow chromosomes to decondense [uncoil]. The N-terminal tails are reversibly modified by adding acetyl or methyl groups. This prevents adjacent nucleosomes from packing together. The H1 histone protein is removed so the binding of DNA to the nucleosome core is loosened. The DNA then resembles a string of beads. Where these changes occur they allow access to the DNA by polymerase enzymes that carry out replication and transcription. Some sections of chromosomes remain condensed during interphase and genes in these sections are therefore not transcribed. Nucleosomes thus help to regulate transcription in eukaryotes, by controlling which sections of the chromosomes are condensed or decondensed during interphase.

**Jmol molecular visualization software** can be used to analyse the association between protein and DNA within nucleosomes. Go to ‘Molecule of the Month’ on the Protein Data Base (PDB). Select ‘Nucleosome’ and then ‘DNA in a nucleosome’ in the list of discussed structures. The Jmol image of a nucleosome can be rotated and coloured in different ways to show components.
**STAGES IN GENE EXPRESSION**

Gene expression is the production of mRNA by transcription of a gene and then the production of polypeptides by translation of the mRNA. In prokaryotes translation can occur immediately after transcription, because there is no nuclear membrane. Translation can even begin before an mRNA molecule has been fully transcribed. In eukaryotes the mRNA is produced by transcription in the nucleus. It is modified while still in the nucleus, then passes out to the cytoplasm via nuclear pores and is translated in the cytoplasm.

**PROMOTERS AND TRANSCRIPTION**

Gene expression can be controlled at the transcription stage – at any time in the life of a cell some genes in the nucleus are being transcribed and others are not. Control of gene expression involves a promoter. This is a base sequence close to the start of a gene. Every gene has a promoter, but the base sequences vary, allowing particular genes to be transcribed and not others. The promoter is not itself transcribed and does not code for an amino acid sequence, so it is an example of non-coding DNA with a function.

RNA polymerase (RNAP) binds directly to the promoter in prokaryotes and then starts transcribing. Repressor proteins can bind to the promoter and prevent transcription. The control of gene expression is more complicated in eukaryotes. Proteins called transcription factors bind to the promoter, which allows RNAP to bind and then initiate transcription. Several transcription factors are required, some of which may need to be activated by the binding of a hormone or other chemical signal. Repressor proteins can prevent transcription.

After transcription has been initiated RNAP moves along the gene, assembling an RNA molecule one nucleotide at a time. RNAP adds the 5' end of the free RNA nucleotide to the 3' end of the growing mRNA molecule, so transcription occurs in a 5' to 3' direction. The elongation of RNA by transcription was described in Topic 2. Transcription is terminated at the end of the gene and the DNA, RNA and RNAP separate.

**IDENTIFYING POLYSOMES**

The figure below is an electron micrograph showing groups of ribosomes called polysomes (or polyribosomes). A polysome is a group of ribosomes moving along the same mRNA, as they simultaneously translate it.

In the micrograph below the arrow indicates where transcription of a prokaryote gene is being initiated. Along the DNA of the gene are nine polysomes. Only in prokaryotes can translation begin before transcription is finished.
**Epigenetics**

**INHERITANCE OF ACQUIRED CHARACTERISTICS AND EPIGENETICS**

A fundamental theory of modern biology is that characteristics acquired during an individual's lifetime cannot be inherited by their offspring. The alternative theory that acquired characteristics can be inherited was propounded by the French biologist Lamarck, so is referred to as **Lamarckism**.

Evidence has sometimes been presented for inheritance of acquired characteristics, but has been falsified and Lamarckism was dismissed as tantamount to heresy. The discovery that DNA is the genetic material added to the evidence against Lamarckism – the environment of an individual during its lifetime cannot cause specific changes to the base sequences of their genes.

Nevertheless there is mounting evidence that the environment can indeed trigger heritable changes. One explanation involves small chemical markers that are attached to DNA in the nucleus of a cell to fix the pattern of gene expression. These markers are usually passed to daughter cells formed by mitosis, and help to establish tissues with common patterns of differentiation, but they are mostly erased during the gamete formation. However a small percentage of markers persists and is inherited by offspring. The pattern of chemical markers established in the DNA of a cell is the **epigenome** and research into it is **epigenetics**.

Example of epigenetic inheritance: Methylation is one type of chemical marker. Variations in the pattern of methylation that affect height and flowering time in the model organism *Arabidopsis thaliana* (left) have been shown to be inherited over at least eight generations.

**METHYLATION AND EPIGENETICS**

Cytosine in DNA can be converted to methylcytosine by the addition of a methyl group (–CH₃). This change is catalysed by an enzyme and only happens where there is guanine on the 3' side of the cytosine in the base sequence. In some eukaryotes there is widespread **methylation** in parts of the genome.

Methylation inhibits transcription, so it is a means of switching off expression of certain genes. The cells in a tissue can be expected to have the same pattern of methylation and this pattern can be inherited in daughter cells produced by mitosis. Environmental factors can influence the pattern of methylation and gene expression.

Fluorescent markers can be used to detect patterns of methylation in the chromosomes. Analysis of the patterns has revealed some trends:

1. Patterns of methylation are established during embryonic development and the percentage of C-G sites that are methylated reaches a maximum at birth in humans but then decreases during the rest of an individual’s life.

   ![Graph showing percentage of methylation level](image)

   - Newborn: 82%
   - 26 years old: 79%
   - 103 years old: 72%

2. At birth identical twins have a very similar pattern of methylation, but differences accumulate during their lifetimes, presumably due to environmental differences. This is reflected in the decreasing similarity between identical twins as they grow older.

**POST-TRANSCRIPTIONAL MODIFICATION**

Eukaryotic cells modify mRNA after transcription. This happens before the mRNA exits the nucleus. In many eukaryote genes the coding sequence is interrupted by one or more non-coding sequences. These **introns** are removed from mRNA before it is translated. The remaining parts of the mRNA are **exons**. They are spliced together to form mature mRNA.

Some genes have many exons and different combinations of them can be spliced together to produce different proteins. This increases the total number of proteins an organism can produce from its genes.
 Ribosomes and transfer RNA

**TRANSFER RNA**

All transfer RNA molecules have:
- double-stranded sections with base pairing
- a triplet of bases called the **anticodon**, in a loop of seven bases, plus two other loops
- the base sequence CCA at the 3’ terminal, which forms a site for attaching an amino acid.

These features allow all tRNA molecules to bind to three sites on the ribosome – the A, P and E sites.

The base sequence of tRNA molecules varies and this causes some variable features in its structure. These give each type of tRNA a distinctive three-dimensional shape and distinctive chemical properties. This allows the correct amino acid to be attached to the 3’ terminal by an enzyme called a tRNA activating enzyme. There are twenty different tRNA activating enzymes – one for each of the twenty different amino acids. Each of these enzymes attaches one particular amino acid to all of the tRNA molecules that have an anticodon corresponding to that amino acid. The tRNA activating enzymes recognize these tRNA molecules by their shape and chemical properties. This is an excellent example of enzyme-substrate specificity.

Energy from ATP is needed for the attachment of amino acids to tRNA. ATP and the appropriate amino acid and tRNA bind to the active site of the activating enzyme. A pair of phosphates is released from ATP and the remaining AMP bonds to the amino acid, raising its energy level. This energy allows the amino acid to bond to the tRNA. The energy from ATP later allows the amino acid to be linked to the growing polypeptide chain during translation. Images of tRNA molecules made using molecular visualization software can be obtained from the Protein Data Bank and viewed with Jmol.

The image (above) shows a space-filling model of a tRNA for the amino acid phenylalanine. The position of the amino acid is indicated by the arrow and the anticodon by the three letters near the base.

**RIBOSOMES**

Ribosomes have a complex structure, with these features.
- Proteins and ribosomal RNA molecules (rRNA) both form part of the structure.
- There are two sub-units, one large and one small.
- There is a binding site for mRNA on the small sub-unit.
- There are three binding sites for tRNA on the large sub-unit:
  - A site for tRNA bringing in an amino acid
  - P site for the tRNA carrying the growing polypeptide
  - E site for the tRNA about to exit the ribosome.

The structure of a ribosome is shown in outline (below).

Much more accurate images of ribosome structure can be made using molecular visualization software. The image below is from the Protein Data Bank. This site allows a variety of three-dimensional coloured images of ribosomes to be produced and viewed from any angle.

In the cytoplasm there are free ribosomes that synthesize proteins primarily for use within the cell. There are also ribosomes attached to membranes of the endoplasmic reticulum. They are called bound ribosomes and synthesize proteins for secretion from the cell or for use in lysosomes.
**TRANSLATION**

**INITIATION OF TRANSLATION**
A sequence of events occurs once, to start the process of translation:
1. The small sub-unit of the ribosome binds to mRNA with the start codon in a specific position on the mRNA binding site of the small sub-unit.
2. A tRNA with an anticodon complementary to the start codon binds. The start codon is usually AUG, so a tRNA with the anticodon UAC binds. This tRNA carries the amino acid methionine.
3. The large sub-unit of the ribosome binds to the small unit. The mRNA is positioned so that the initiator tRNA carrying methionine is in the P site. The E and A sites are vacant.
4. A tRNA with an anticodon complementary to the codon adjacent to the start codon binds to the A site.
5. A peptide bond forms between the amino acids held by the tRNAs in the P and A sites.

**ELONGATION**
The elongation of polypeptides involves a repeated cycle of events.
1. The ribosome moves three bases on along the mRNA towards the 3' end. This moves the tRNA in the P site to the E site and the tRNA carrying the growing polypeptide from the A to the P site, so the A site becomes vacant.
2. The tRNA in the E site detaches and moves away so this site is also vacant.
3. A tRNA with an anticodon complementary to the next codon on the mRNA binds to the A site.
4. The growing polypeptide that is attached to the tRNA in the P site is linked to the amino acid on the tRNA in the A site by the formation of a peptide bond.

**TERMINATION OF TRANSLATION**
1. The ribosome moves along the mRNA in a 5' to 3' direction, translating each codon into an amino acid on the elongating polypeptide, until it reaches a stop codon.
2. A tRNA molecule has the complementary anticodon and instead release factors bind to the A site, causing the release of the polypeptide from the tRNA in the P site.
3. The tRNA detaches from the P site, the mRNA detaches from the small sub-unit, and the large and small sub-units of the ribosome separate.