6.5 Sanger sequencing and the polymerase chain reaction
Studying genomes, Sanger sequencing, electrophoresis and the polymerase chain reaction

Understanding DNA

DNA comprises some three billion base pairings (3 gigabases of DNA) in the human **genome**. A genome is the entire DNA sequence of an organism. It might be described as the entirety of an organism's heredity information. DNA is composed of smaller lengths called **genes** which code for the production of certain proteins and polypeptides, although only a small amount of DNA actually consists of **coding DNA** – a mere 1.5% of human DNA codes for polypeptides, the rest **junk DNA** (also called non-coding DNA).

Sanger sequencing

The most common method of sequencing used nowadays is **Sanger sequencing**, which comes from Frederick Sanger. It was introduced by the two-time Nobel laureate in 1975, and is also known as the *chain termination* method. To decode a sequence of DNA, the template DNA is copied repeatedly, with the copies terminating at different points. The copies can then be arranged into order of size to determine the sequence.
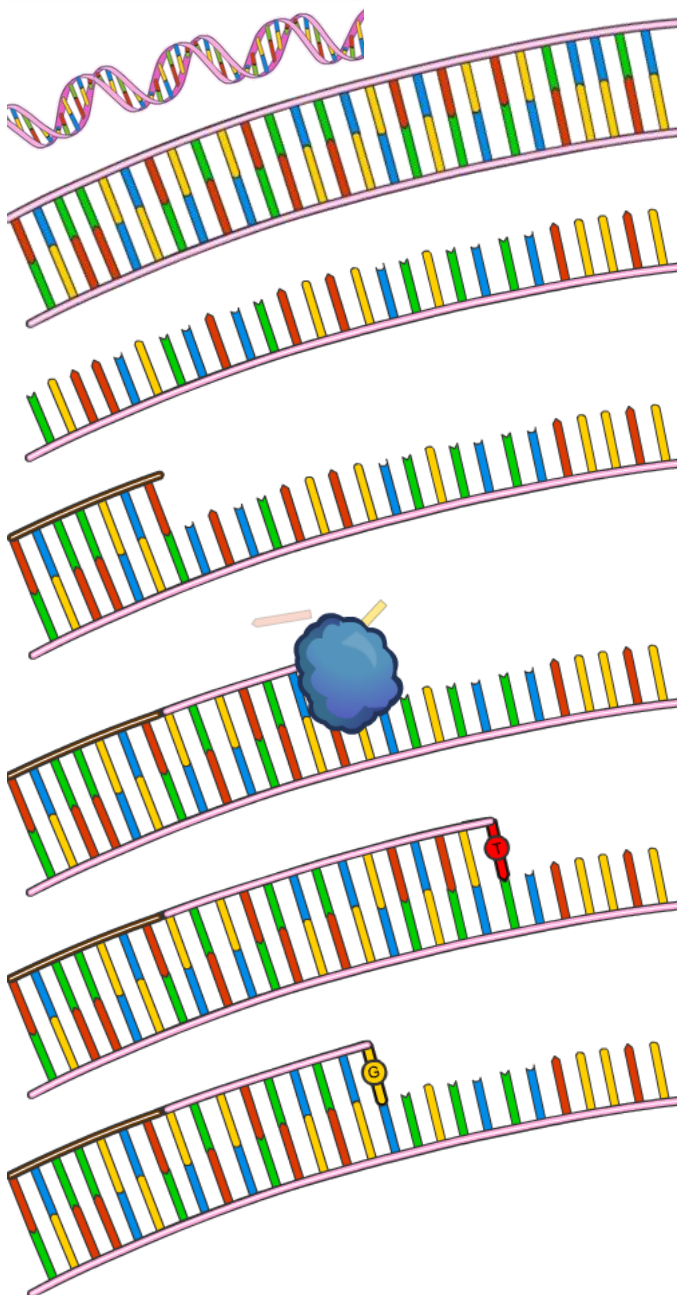
The reaction is initiated at either end of the target length of DNA. The target DNA shown in this example here is also shown in its unwound form, which makes it easier to demonstrate the Sanger method.

The two strands separate, and a short length of DNA called a **primer** (a length which has a base sequence which is complementary to the start of the target length) binds to the template.

The enzyme **DNA polymerase** starts to build an entire new double-stranded molecule by binding free nucleotides to the exposed bases on the target DNA. The DNA polymerase continues to construct a second, complementary strand until it places a **fluorescently-marked** nucleotide (called a **dideoxynucleoside triphosphate** – ddNTP) in the strand. When Sanger sequencing takes place, a mixture of standard free nucleotides (99%) and dideoxynucleosides (1%) are placed into the reaction chamber, so the fluorescently-labelled base can bind to its complementary base.

Dideoxynucleosides lack a *hydroxyl group* (OH) in their structure, which therefore prevents DNA polymerase from binding further bases to the sequence after the dideoxynucleoside has joined. The example shown in the diagram is *thymidine triphosphate*. This molecule prevents the chain from continuing, and therefore this length of DNA is complete.

This is repeated many times, using the same original template DNA, generating a large number of DNA fragments terminating in different areas with fluorescently-labelled markers – which are known as either adenine, cytosine, guanine or thymine. Once many fragments have been generated, they need to be collected and separated using **electrophoresis**.

Generating target DNA fragments

For the above process to work, the target DNA needs to be mass produced: in order to generate enough DNA fragments by Sanger sequencing to be able to decipher a genetic code, *millions* of copies of the original DNA sequence must be produced. This can be done by a process called *polymerase chain reaction* (PCR) which is explained later on.

Separating DNA fragments using electrophoresis

The process of **electrophoresis** is used to separate fragments of DNA according to their size. It is an accurate enough process that it can separate and distinguish fragments that are different by only one base in length. The procedure involves using a gel plate containing **agarose** (a sugar) gel, which is covered in a **buffer solution**. Electrodes are attached at each end of the gel so a current can be passed through it. The separation occurs because longer strands get caught up in the agarose gel and move slowly, whereas shorter strands can move much more quickly through the gel.

wells for
samples
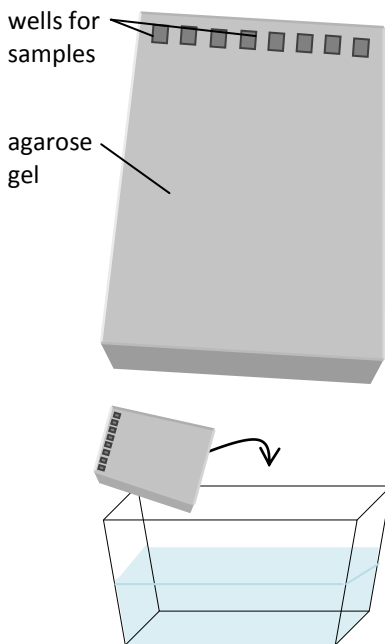
agarose
gel

- the DNA samples are treated with **restriction enzymes** to cut them into fragments (restriction enzymes slice the backbone of DNA molecules in specific places, separating it into smaller fragments)
- an agarose gel is prepared by mixing agarose sugar powder with a buffer solution, and the gel is allowed to set, leaving hardened dips (**wells**) for the DNA samples
- DNA samples are placed into the wells at the negative electrode end of the gel
- the gel is immersed into a deep tank full of buffer solution, called the *gel electrophoresis tank*
- an electric current is passed through the tank for a fixed period of time, usually an hour or so
- in terms of electronegativity, DNA is negatively charged due to the **phosphoryl groups** the nucleotides contain, and so the fragments are attracted to the positive electrode at the other end of the tank
- as was explained above, shorter fragments move more quickly, and so at the end of the fixed time will have travelled further along the gel plate, and vice versa with the longer fragments
- the position of the fragments can be shown by staining the DNA molecules with a specialised dye

A technique known as **Southern blotting** is used to further analyse the fragments. A **nylon sheet** is placed over the gel plate, which is then covered in absorbent paper towels and pressed, and left overnight (this is *blotting*). The DNA fragments are transferred to the sheet (drawn up to the nylon sheet membrane by **capillarity**) and can be analysed.

Even though the fragments are drawn to the nylon sheet, they do not become directly visible this way. They can be seen easily however if all of the samples are labelled with a **radioactive marker** during the electrophoresis process, and then placing photographic film over the sheet, or shining ultraviolet light on the sheet, shows the samples positions in the gel.

If one particular fragment or sequence of DNA is being searched for, such as a particular gene, a radioactive **DNA probe** may be used to check for the presence of that particular gene or sequence.

DNA probes

A DNA probe is a short, single-stranded piece of DNA around 50 bases long which can be used to test for the presence of a particular sequence of DNA. The probe is complementary to the piece of DNA being investigated. A probe can be labelled in a small number of ways:

- using a radioactive isotope marker (usually on the phosphoryl groups) so that the location is revealed when exposed to photographic film
- using a fluorescent marker that emits a colour when exposed to ultraviolet light (the same markers as used above)
- sometimes also using an '*antigen*' which is attached to the DNA probe, that can be analysed to identify the sequence by finding the matching '*antibody*' (which itself will be marked with a radioactive or fluorescent marker)

DNA probes can be added to any sample of DNA fragments, and since it is a single-stranded molecule, it will bind to any fragment which has the complementary target sequence. This binding to complementary sequences is called **annealing**.

Polymerase chain reaction

As mentioned above, the method in which millions of copies of the target DNA fragment are produced for Sanger sequencing to take place is using the **polymerase chain reaction** (PCR). This is essentially just artificial DNA replication. It is carried out on a sample where millions of identical copies can be produced in a short series of events.

The 'ingredients' for the PCR broth are: the piece of DNA being copied, DNA **primers** (short sequences of DNA with base sequences complementary to the starting sequence at one end of the target DNA), the DNA polymerase enzyme and free nucleotides to make the new copies of DNA.
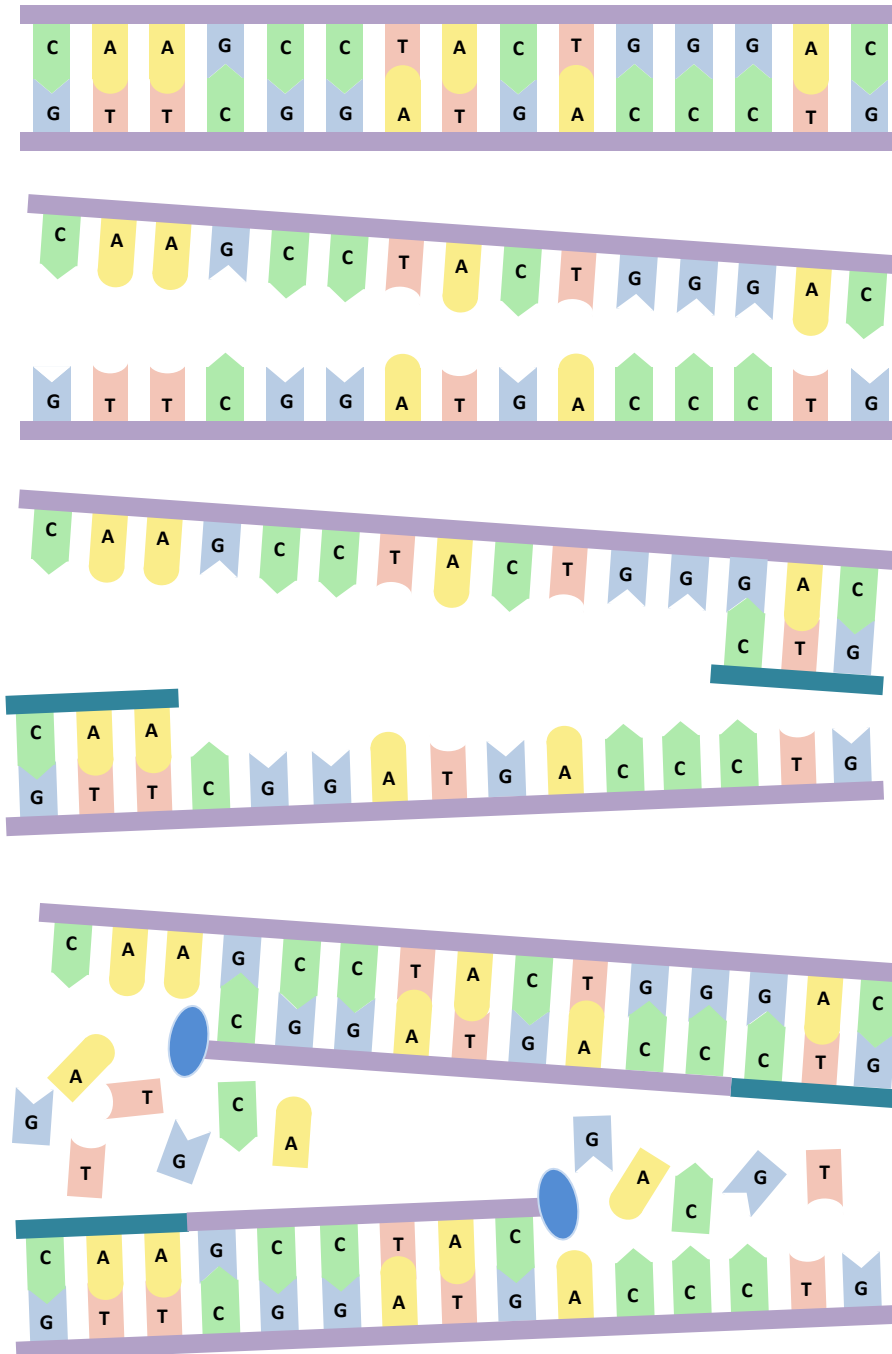
1 The process begins with the double-stranded DNA sample which is to be copied

2 The mixture is heated to 95°C for around 20 seconds, in order to break the hydrogen bonds holding together the two strands, producing two antiparallel single-stranded molecules from the target DNA

3 Short-length DNA *primers* which are complementary to the ends of the target DNA are added, and the mixture is cooled to around 55°C to allow **annealing** to take place (hydrogen bonds form between the primers and the single strands) to form small sections of double-stranded DNA

4 The mixture is then heated to 72°C which is the optimum temperature for **taq polymerase** (a special form of DNA polymerase which functions best at high temperatures), which continues to build the parallel strands on each of the original strands by binding the free nucleotides to the backbone, just as in natural DNA semi-conservative replication

5 When taq polymerase reaches the end of its length, it detaches itself, and the process can be repeated

Once the strand has been replicated and taq polymerase is released, the process can be repeated: the two identical strands which have been produced are heated to separate the strands and can be replicated again. This way, replication is **exponential**, and the amount of copies made doubles each turn of the cycle. After 25 times, the single strand at the beginning of the process has been replicated to produce just over 33 million lengths of DNA.

The enzyme used in polymerase chain reaction, *taq polymerase*, is **thermophilic**. It is described in this way because it is not denatured by the extreme temperatures used in the process, which makes it a very useful enzyme. This special form of DNA polymerase actually has an optimum temperature of around 72°C. The reason it can survive such temperatures is because it derived from a thermophilic bacterium which thrived in hot springs at temperatures of around 90°C.

Sanger sequencing comprises PCR, the sequencing reaction and electrophoresis

Now that you understand the processes of the polymerase chain reaction, DNA sequencing, and electrophoresis, it is easier to understand how Sanger sequencing really works:

- a target length of DNA that is to be sequenced is artificially replicated thousands, or millions, of times using the polymerase chain reaction, to produce identical copies for the sequencing reaction
- the physical sequencing reaction then takes place as described before: primers attach to the beginning of each strand, and then nucleotides are bound to the target strands by the enzyme DNA polymerase, until a dideoxynucleoside triphosphate molecule is bound, which forces the DNA polymerase to detach and the sequence stops
- the resultant products are lots of different lengths of DNA, each terminating at different points with a fluorescently-tagged dideoxynucleoside triphosphate
- the DNA samples which have been produced are then inserted into the wells of an electrophoresis agarose slab and immersed into a tank of buffer solution, so that electrophoresis can take place
- as electrophoresis separates the sequences by length, a machine with a laser runs across all the samples as they move through, and a computer reads and displays the colour sequence: starting with the sample with one nucleotide only, then the one with two, then the one with three and so on, decoding the whole target DNA