3G1 Introduction to Molecular Bioengineering

Course Notes

Mrinank Sharma

April 17, 2018

1 Biology Fundamentals

1.1 Evolution

Natural Selection

Over time, **heritable traits which promote survival and reproduction become more common.** Diversity within a population is driven by genetic variation (due to mutation) and **competition** provides a selective force. Selection may be environment (i.e. for local adaptation) or sexual (fitness or ornamentation).

Evolutionary Cycle

In short, mutations within the genotype (the genetic constitution of an organism) result in different phenotypes (observation characteristics resulting from the genotype). The most advantageous phenotypes are selected and the organisms which such genes reproduce (and hence further mutation occurs) i.e. **selection increases the frequency of useful traits**.

Where do we come in?

Humans interfere with the evolutionary cycle. We **either** carry out mutagenesis **or** design specific mutations **or** introduce new functions to direct generation of diversity and then **artificially select** for traits which are useful for our own purposes.

1.2 Human Genetic Variation

Why is Genetic Variation Masked in Humans?

Humans have 2 copies of most genes, **masking** latent genetic variation (i.e. dominant/recessive genes). DNA acts as a base-4 digital information store and variations in DNA result in genetic variation and hence different phenotypes.

1.3 DNA

What is DNA?

Deoxyribosenucleic acid or DNA for short is a key chemical in living organisms and can be considered to be a base-4 digital information store.

What is the structure of DNA?

DNA is a double stranded molecule with antiparallel strands. A single strand of DNA can be considered to be a series of nucleotides linked by phosphodiester bonds. The chain of deoxyribose sugar and phosphate groups is known as the **sugar-phosphate backbone** and note that there are **four bases**, adenine **(A)**, cytosine **(C)**, thymine **(T)** & guanine **(G)**.

Base pairs

Adenine and guanine bond with 2 hydrogen bonds whilst cytosine and thymine bond with 3 hydrogen bonds (and hence this bond is slightly stronger). As a result, DNA is able to form double stranded molecules with the well-known double helix structure. Note that the width of DNA is constant because each Watson-Crick base pair is between one pyrimidine and one purine.

Directionality

Please note that **DNA strands are directional.** There is a 3' (hydroxyl bearing) and a 5' (phophate bearing) end and whenever a double strand molecule (*complex*) forms each strand is arranged in an anti-parallel manner. When a sequence of DNA is written, it is always written 5' to 3'.

1.4 RNA

RNA

Ribonucleic acid or RNA is very similar to DNA. They are composed from similar building blocks and in a similar fashion but uridine replaces thymine in DNA. Also note that the sugar used is different.

RNA Diversity & Stability

In the cell, RNA often exists as single strands (this is uncommon for DNA). There is significant structure diversity since each strand is able to base pair with itself. Furthermore, note that RNA is relatively unstable compared to DNA.

1.5 Amino Acids

Sub-unit

Amino acids are the sub-units used to build proteins. Their structure can be seen in Fig. 2.

The R group (side-chain) is different for each amino acid and effective determines the function of each amino acid. Table 1 outlines some amino acids which are worth learning.

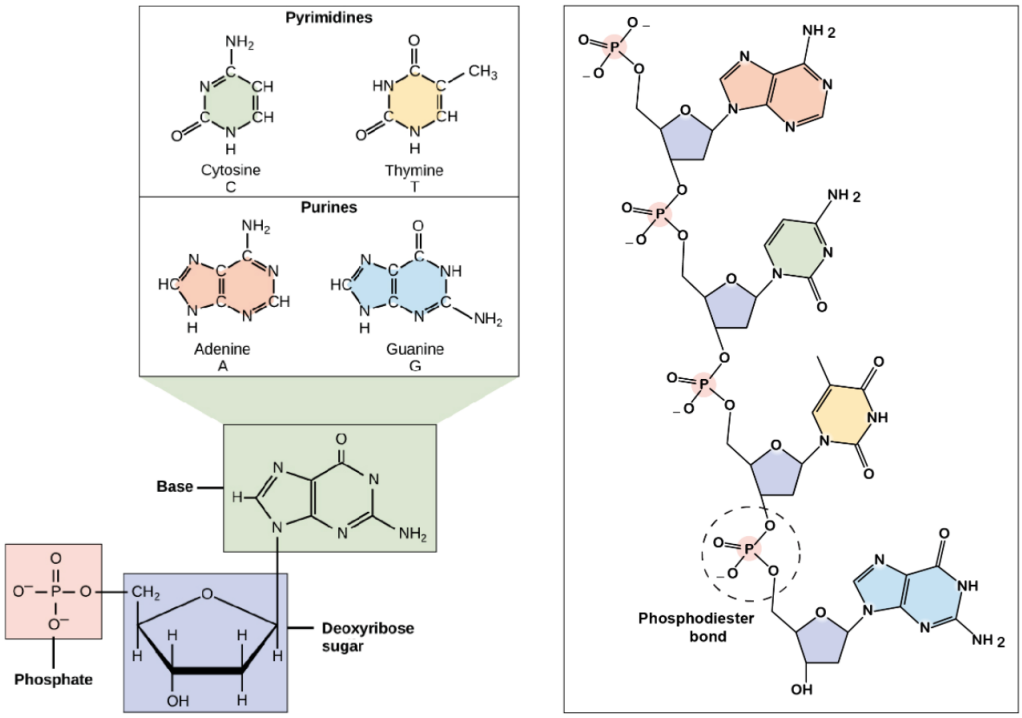


Figure 1: The structure of DNA. (Taken from Khan Academy)

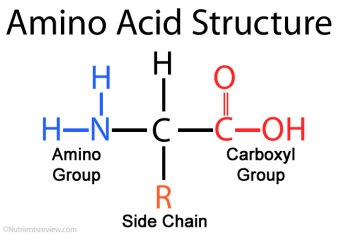


Figure 2: The structure of Amino Acids. The central carbon is referred to as the **alpha carbon** and the amino group is alkaline whilst the carboxyl group is acidic.

Type	Description	Examples
Aliphatic	Do not contain especially stable carbon rings (non-aromatic).	Alanine, Glycine
Aromatic	Cyclic, planar with a ring of resonance bonds (e.g. benzene ring).	Tyrosine
Acidic	-	Aspartic Acid
Basic	-	Arginine
Sulfur-containing	Cysteine can form disulfide bonds; important for structure!	Cysteine
Amidic	Containing a $CONH_2$ group.	Asparagine

Table 1: Common Amino Acids (worth learning).

Peptide Bonds

Amino acids form proteins through polymerization; through a dehydration reaction, multiple amino acids are able to join together forming peptide bonds. Even though the peptide bond itself is planar, the bonds around each alpha carbon are able to rotate giving great flexibility.

1.6 Proteins

Protein Structure

Secondary Structures: What & Why?

Hydrogen bonding between the sub-units in each polypeptide chain result in secondary structure. There are two main forms of secondary structure in proteins as seen in Fig. 3.

1. Hydrogen bonding between `N-H & C=O` four residues away cause **alpha helices** to form.
2. Hydrogen bonding between adjacent parallel or anti-parallel strands cause **beta sheets** to form.

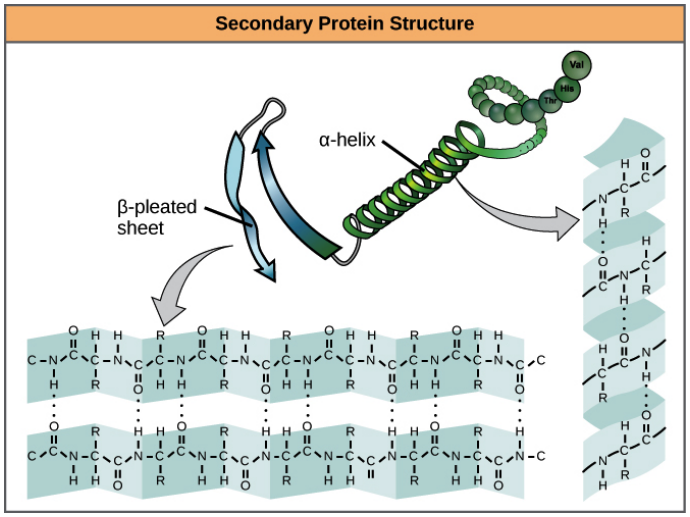


Figure 3: Protein secondary structure; hydrogen bonds shown.

Protein Structure

The structure of proteins is hierarchical. Primary structure refers to the sequence of amino acids whilst secondary structure refers to local folded structures (see above). The overall 3D structure of a polypeptide is the tertiary structure and is primarily due to interactions between different R groups. The quaternary structure only applies for proteins made up of several polypeptide chains and outlines how these chains interact (i.e. the complex formed).

Protein Function

Proteins carry out a wide range of roles in the cell such as acting as enzymes and structural roles with a high degree of specificity. For example, transmembrane transporters can distinguish between Na^+ & K^+ ions.

How?

Molecular recognition is bases on complementarity of shape and chemical properties (e.g. charge) and hence **protein shape is critical**.

1.7 Phospholipids

What are Phospholipid?

A class of lipids (soluble in non-polar solvents) which are amphiphilic i.e. possessing parts which are both hydrophilic (head) and hydrophobic (tail). Because of this property, structures which are energetically favourable can spontaneously form (self-assemble) such as those in Fig. 4.

Bilayer Sheet

The bilayer sheet is especially important; cells are surrounded by a lipid bilayer membrane which isolates the cell contents from the environment as it is impermeable to most ions and water soluble molcules.

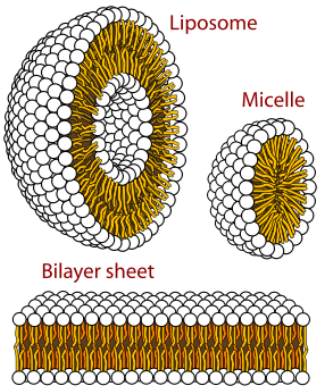


Figure 4: Structures formed from self-assembly

Self-Organisation

Self-Organisation!

In fact, living systems in general are highly sophisticated and self-organising. In addition to phospho-

lipids, RNA and proteins also self-assemble into complex, defined functional forms as does DNA. In fact, this self-assembly of DNA can be used to build complex structures.

1.8 The Three Domain System

Three Domains

When organising species, the domain is the highest rank of species used. Bacteria & archea are prokaryotes i.e. single celled and lacking a membrane bound nucleus whilst eukaryotes have nuclei and other organelles with membranes.

Archea

Archea were originally classed as bacteria. They are though of as extremophiles which inhabit as diverse environments as bacteria. They have similar metabolic pathways to eubacteria (bacteria with rigid cell walls) but some enzymes are more similar to those of eukaryotes.

Relevance

They play important roles in the carbon and nitrogen cycle and do not act as parasites. Their robustness and metabolism is useful for **biotechnology**.

Bacteria

Bacteria are found almost everywhere and are a significant proportion of the worlds biomass. They have no nucleus.

Relevance

They are critical to nutrient cycles. Bacteria are important for good health but also responsible for many diseases.

Eukaryotes

Eukaryotes have both single and multicellular forms with a large array of different physical forms. They inhabit diverse diverse environments and include plants, animals and fungi. Animals and plants are both eukarya and Fig. **??** shows some of the differences between their cells.

Relevance

We are eukarya! They are capable of significant environmental impact and many are of great importance to our survival.

Genetic Organisation

Prokaryotes

Prokaryotes typically have a single circular chromosome with extrachromosomal elements known as plasmids which form the basis for cloning vectors.

Eukaryotes

Eukaryotes typically have several linear chromosomes. Within their cells, mitochondria and chloroplasts have their own DNA and some 'lower forms' can have plasmids.

Chromosome Packing

DNA is wrapped around **histones** (proteins which bind with DNA through electrostatic interactions) to form a nucleosome which is the building block of chromatin. Nucleosomes coil to form chromatin fibres which are condensed into chromosomes during cell division. Note that the packing is dynamic according to the needs to the cell. See Fig. 5 to understand the hierarchical packing of chromosomes.

Centromeres and Telomere

Also note that chromosomes have centromeres (where chromosomes attach to the spindle during cell division - the central part of the chromosome) and telomeres at the end.

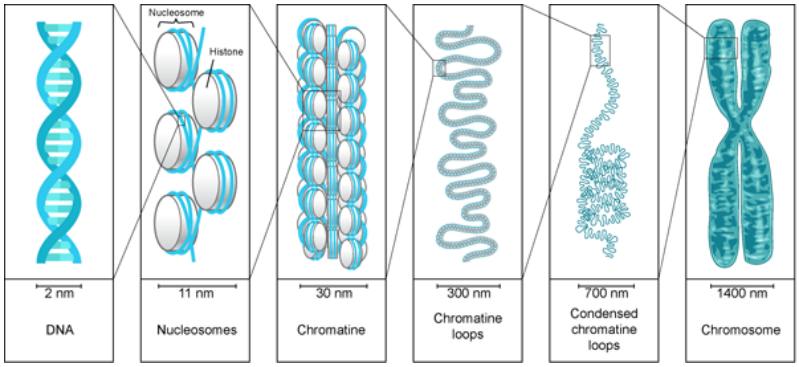


Figure 5: Chromosome Packing

1.9 Antibodies

Antibobies are proteins which are used naturally to neutralise pathogens. They have several useful properties.

Properties

- They are able to recognise antigens (protein coatings) which a very high degree of specificity.
- They are highly diverse, aided by the combination of light and heavy chains.
- They are robust and long-lived; disulfide bonds help to prevent degradation.
- They are bivalent.

Antibody Structure

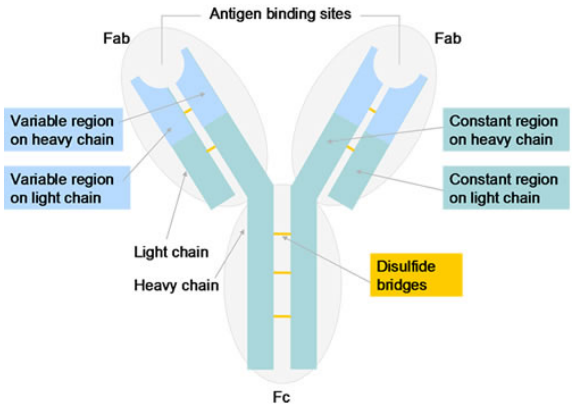


Figure 6: Structure of Antibodies. Fab: Fragment antigen binding. Fc: Fragment crystallisable.

Isotypes

There are antibody isotype determined by the constant regions which determine how antibodies interact with each other. Each class is linked to a specific gene. For example, `IgA` is produced using gene α and `IgG` is produced using gene γ (the remaining types follow this pattern). Each isotype is found in different places in the body and they act in different ways.

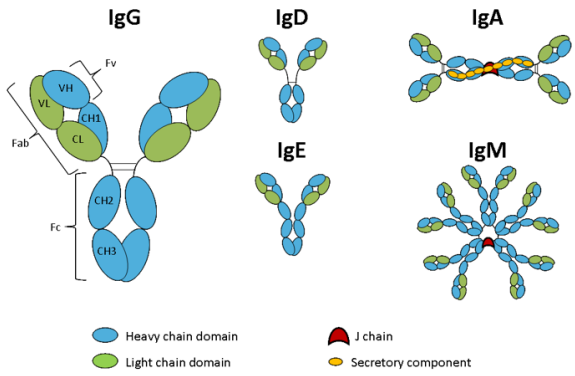


Figure 7: Different antibody isotypes. They contain different constant regions, leading to different complexes.

CDR

The complementarity-determining regions are part of the variable chains where the molecules specifically bind to their antigen and as the name suggests, variable chains change between antibodies.

Somatic/VDJ

There is significant antibody diversity due to **somatic recombination**. The variable region of each chain encoded in several gene segments of which there are three types, V (variable), D (diversity) & J (joining) (note, no D segments in light chains). Each B cell (a type of white blood cells of the lymphocyte subtype which secrets one type of antibody) will assemble a variable region by **randomly selecting and combining one V, D & J segment.** This leads to a very large number of different antibodies.

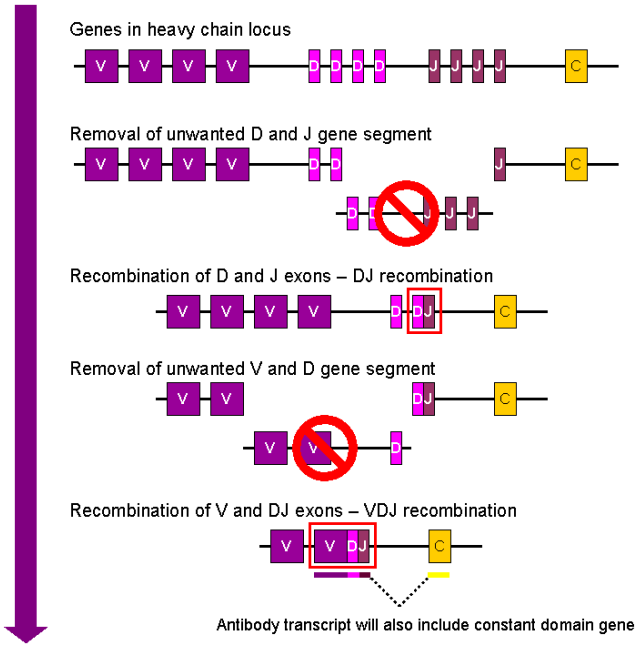


Figure 8: Somatic/VDJ recombination.

Class Switching

B cells are able to switch their class, changing the isotype of antibody they produce e.g. from `IgM` to `IgG`. The **constant-region of the heavy chain is changed but the variable region remains constant** meaning that **class switching does not affect antibody specificity**. Please note that only one heavy chain and one light chain is expressed within each B cell.

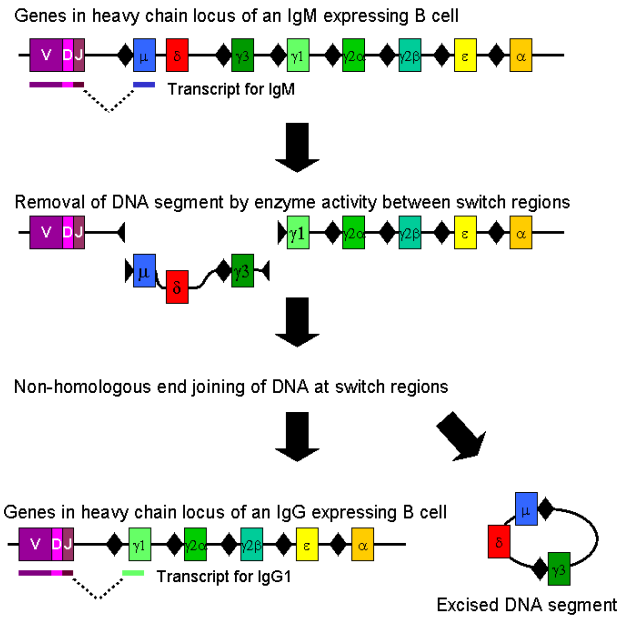


Figure 9: Recombination for B cell class switching.

1.10 Eukaryotic Cell Division

Cell division is essential for life. There are two primary forms of eukaryotic cell division.

Mitosis

Mitosis

Mitosis creates clones of an original diploid cell by copying each chromosome and then splitting the cell. NB: The chromosomes are seperated using a microtubule spindle.

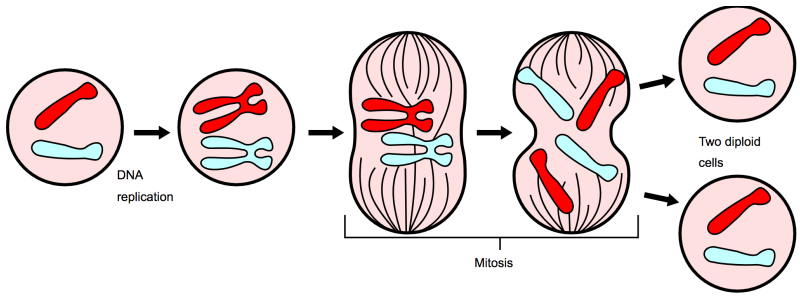


Figure 10: Mitosis.

Meiosis

Meiosis

Meiosis is a *reductive* type of cell division which forms **haploid gametes** - cells with single copies of DNA which are able to unite with another cell of the opposite sex during reproduction. Hence meiosis is part of the sexual life cycle. Fig. 11 outlines how meiosis occurs; note that crossing over occurs.

Recombination

Gamete recombination is important as it creates new combinations of alleles (different forms of the same gene).

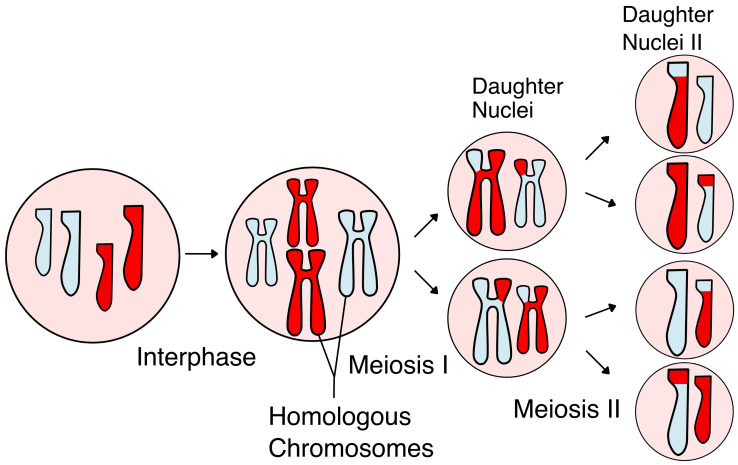


Figure 11: Meiosis.

2 The Central Dogma

2.1 The Central Dogma of Molecular Biology

The Central Dogma

The central dogma of molecular biology is a proposed **information flow** which explains how the cell transforms encoded information into the proteins required for development and metabolic demands.

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

Experimental Methods Used

It had been known that there existed heritable traits for some time (e.g. Mendel 1860 & Griffith 1928) but the molecule of heredity was unknown. It was discovered that **DNA is the molecule of heredity** by fractionating cell-free extract of S-strain cells and testing for transformation of R-strain cells (Avery, MacLeod & McCarthy, 1944) and this was confirmed in 1952 using radioactively labeled phage; bacteria was infected with phage with radioactive S on the phage protein coating and radioactive P in phage DNA. Radioactivity was recovered primarily on phage ghosts in the first case and in the bacteria in the second case (Hershey & Chase, 1952).

2.2 DNA Replication

If DNA is the molecule of heredity, DNA must be able to replicate.

Adding Deoxynucleotides

Using a template strand, an additional nucleotide can be added in two stages. First, a nucleotide forms hydrogen bonds with the exposed base on the template strand and then **DNA Polymerase** creates a phosphodiester bond between the adjacent nucleotides on the non-template strand. This process only occurs in a buffer solution at the correct pH. However, note that **DNA Polymerase only**

A caveat...

works from 5' to 3'! Hence synthesis of the lagging strand is discontinuous and there are **okazaki** fragments.

How does DNA replicate?

A double helix replicates by separating the helix into each strand and then cloning each strand, which can be shown by growing bacteria in a heavy `N` media and switching to grow in a normal `N` and observing that there exist DNA molecules with both heavy and light `N` (Meselson & Stahl, 1958). **N.B. DNA can be seperated using density gradient centrifugation.**

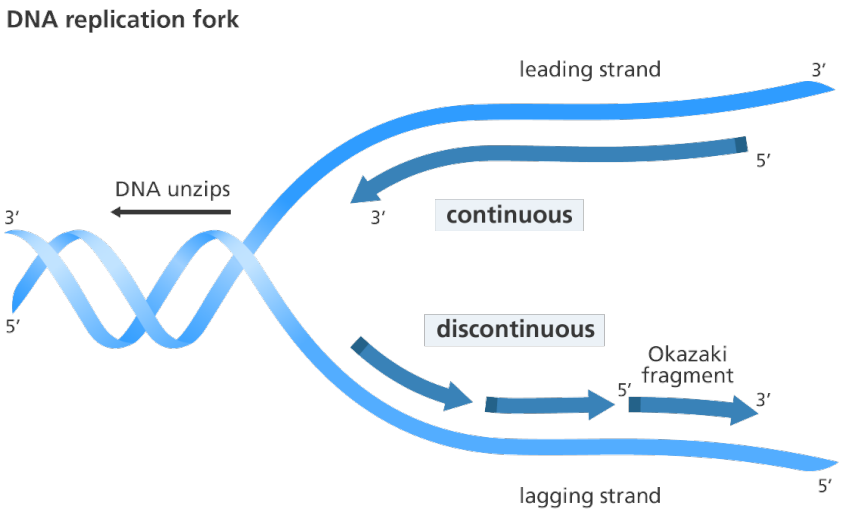


Figure 12: DNA Replication

Enzymes involved

Several enzymes are essential for DNA replication including but not limited to the following:

- DNA Polymerase III is essential for DNA replication and adds phosphodiester bonds between nucleotides.
- DNA Helicase is responsible for seperating the double-stranded DNA.
- RNA Primase places RNA primer as DNA Polymerase III requires a double stranded region to begin replicating DNA.
- Topoisomerase is responsible for unwinding the DNA.
- DNA Ligase repairs breaks in DNA strands (e.g. due to okazaki fragments).

2.3 RNA as a Messenger

Transcription

Short-lived RNA was observed and was hypothesised to encode for information. **DNA is transcribed to messenger RNA** mRNA which conveys genetic information to the ribosome. RNA polymerase synthesises RNA using a template strand but note that only one strand is used; hence there is a template strand and a coding strand (but note that `U` replaces `T`).

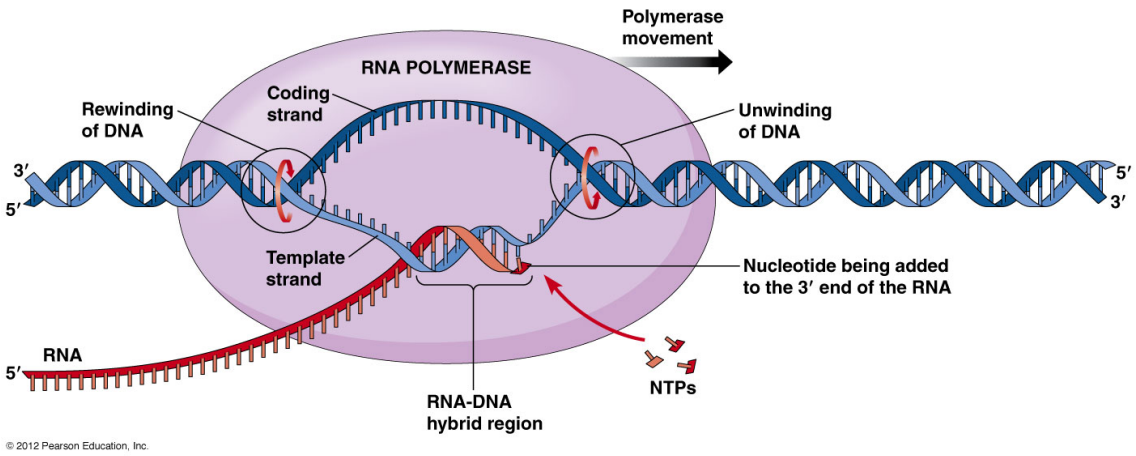


Figure 13: RNA Transcription

2.4 Translation

Code Used

The triplet code was suggested since this gave $4^3 = 64$ different amino acids and there needed to be at least 20 (i.e. the 20 common amino acids). Each **triplet is called a codon** and there are multiple codons coding for the same amino acid. Note that `AUG` is the most common start codon (which also codes for methionine) and `UAG`, `UAA`, `UGA` are stop codons (which do not specify amino acids).

Ribosomes

Ribosomes are the structures where polypeptides are build anad are made up of protein and ribosomal RNA (rRNA). The ribosome not only acts as an enzyme for protein synthesis but also structurally aids each tRNA to find its matching codon.

Transfer RNAs

Transfer RNAs are molecular bridges that connect mRNA codons to the amino acids they encode. Each tRNA has an anticodon which is able to bind to the mRNA and the other end of the tRNA carries the amino acid specified by the codon. One way of observed this is to radioactively label different tRNAs (which would then bind to different codons) and add to washed ribosomes complexed with known mRNA, then filtering out the ribosome and noting where radioactivity is found(Nirenberg & Leder, 1964).

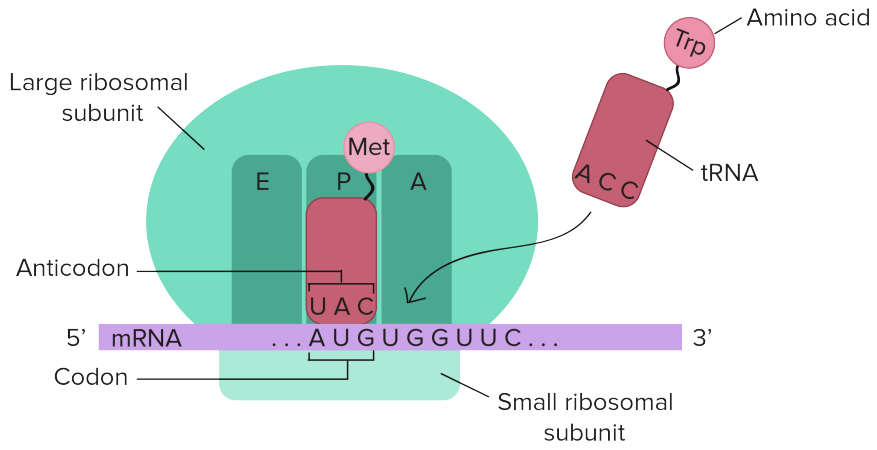


Figure 14: tRNA use.

Wobble Base Pair

Each tRNA can read one **or multiple codons**; Crick suggested that the third base was less spatiailly contrained than the previous two base pairs and hence can have non-standard (wobble) base pairs. Hence whilst there are 61 possible tRNAs, many cells have fewer due to the wobble base pairs with a minimum of 31 for unambiguous translation.

Prokaryotic Translation

In prokaryoes, translation occurs simultaneously with transcription. However, in eukaryotics the translation product may pass into the endoplasmic retriculum for cleavage of signal peptides. Once passed to the Golgi complex, further processing occurs and eventually the peptide can be secreted.

3 Gene Control

3.1 General Structure

The structure of DNA sequences in the genome is important. There are several components:

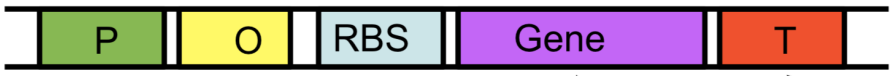


Figure 15: The structure of **genes? Is this the right word?**

- The promotor is the region of DNA that initiates transcription of a gene and where RNA polymerase is able to find. They can also be regulated.
- The operator is a region of DNA made up of binding sites for repressor protein and other transcription factors / regulatory elements.
- The ribsome binding site is the location in mRNA where the the ribosome binds and translation starts.
- The gene is the protein coding sequence.
- The terminator is the mRNA location where transcription terminates e.g. a hairpin in mRNA.

3.2 The *E. Coli lac* operon

For many bacteria, glucose is the prefered carbon source and the ability to use other sugars is carefully controlled. The cell is able to use lactose and the *lac* operon controls the use of lactose depending on the amounts of both lactose and glucose present.

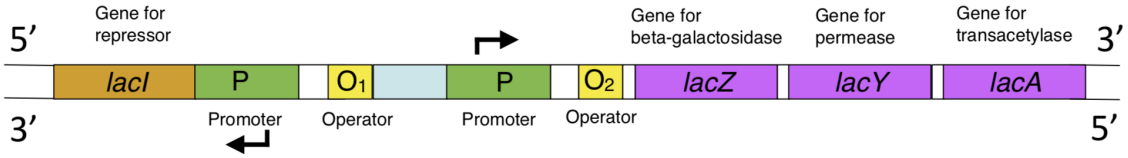


Figure 16: The structure of the *lac* operon, noting that ribosome binding sites and other encoding sequences are left out.

It is important to understand what each gene codes for:

- *lacI* codes for the repressor protein. *lacI*
- *lacZ* codes for beta-galactosidase which hydrolyses lactose into allolactose as well as glucose & galactose. *lacZ*
- *lacY* codes for permease which allows more lactose to enter the cell. *lacY*
- *lacA* codes for transacetylase whose biological role remains unclear. *lacA*

Note a few key elements of the lac operon:

- The **promotor** is where RNA polymerase binds and hence where transcription begins. They may be *constitutive* or regulated.
- The **operator** regions are bound together by the *lac* repressor protein if present. When this occurs, RNA polymerase is unable to bind to the promoter.
- Cyclic AMP (adenosine monophosphate) binds to CRP (cAMP receptor protein, also known as catabolite activator protein) allowing the CAP binds to the CAP binding site (blue, note that this is only possible when cAMP is bound to CAP). When CAP is bound to this site, transcription is promoted by aiding the binding of RNA polymerase.

The conditions which the *lac* operon operators under are very important.

- Without lactose, since the operator regions are bound by the repressor there is only a **low level 'leaky' expression** of *lacZ* e.t.c since this complex is in fact in equilibrium with an unbound complex. *Leak*
- In the presence of Lactose, allolactose created by hydrolysis of lactose is able to induce the *lac* operon by binding to the repressor protein. There is positive feedback as this results in the synthesis of permease which allows further allolactose into the cell. *Lactose*
- The amount of cyclic AMP is inversely proportional to the glucose in the bacterial cell. Hence as glucose levels rise, the amount of cyclic AMP in the cells fall and hence the activity of the lac operon falls; this is negative feedback. With high levels of glucose, transcription only occurs at a low level. *Glucose*

Hence, the lactose operon serves to regulate the use of lactose in the cell depending on the amount of glucose (and lactose) present.

3.3 The *ars* operon

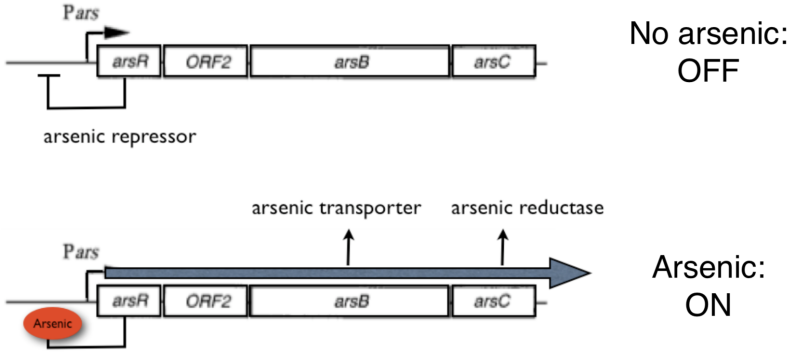
The ars operon is an example of negative auto regulation i.e. self-regulation. The Arsenic repressor respresses its own transcription in the absence of arsenic.

4 Genetic Engineering

4.1 Molecular Cloning

Cutting & Pasting DNA

Restriction Enzymes, a bacterial defense mechanism, are able to cut out DNA. They cut at specific *Cutting DNA*



Sato et al. (1998) J. Bacteriol 180:1655-1661

recognition sites whose sequences are often palindromic. Some Restriction enzymes create DNA with *sticky ends* (longer sections of overhanging DNA) whilst other enzymes produce blunt ends (no overhanging DNA). Note that chromosomal DNA is protected by a DNA *methylase*

Ends

Chromosomal DNA

DNA Ligase repairs breaks in double stranded DNA and is hence used to 'paste' DNA segments together producing recombinant DNA.

Pasting DNA

Separating DNA

Gel electrophoresis can be used to separate DNA molecules.

Technique for Separation

Since DNA is negatively charged in solutions at pH $7-8$, it will move in an electric field. Smaller molecules will migrate more rapidly. DNA can be visualised under UV light by staining with dyes e.g. ethidium bromide.

Methodology

Transforming DNA

Modified DNA needs to be taken up by the cell. There are two methods for doing this and in each case the cells are allowed to recover before selection occurs.

1. Chemical Transformation. Cells are chilled in $\mathtt{CaCl}_2$ to permeabilise the membrane. A heat shock $(\sim 42°C, 30s)$ prompts the uptake of new DNA.

   Chemical Transformation

2. Electroporation. Cells are purified to remove ions and a high voltage shock is used to punch holes in the membrane.

   Electroporation

Growing cell cultures on agar plates allows for isolation of each colony.

Overall Method

We seek to isolate, propagate and and clone large quantity of particular DNA sequences. We use cloning vectors which include endogenous plasmids, bacteriophages and bacterial artificial chromosomes to store DNA which will be taken up by the cell. The cloning vectors may include a multiple cloning site (MCS) which contains many restriction sites. Selectable markers are often introduced to allow for easier artificial selection.

Cloning Vector

Overall Method

1. Cut cloning vector using restriction enzymes.
2. Isolate fragments of interest via gel electrophoresis.
3. Ligate foreign DNA into vector.
4. Transform recombinant DNA into the host.
5. Select for growth of transformed bacteria. This is often aided through the use of selectable markers which confer traits suitable for artificial selection e.g. antibiotic resistance.

   Selectable Markers

6. Screen for clones with the correct construct and culture these on a large scale.

Often **blue-white** screening is used where cells transformed with vectors containing recombinant DNA produce white colonies and cells transformed with non-recombinant vectors grow in blue colonies e.g. by inserting DNA which can restore the activity of defective genes which produce a blue product e.g. the *LacZ* gene in the *lac* operon. These screenable markers are an alternative to selectable markers.

Blue-White Screening

The main limitations of this method are that it requires restriction sites in the correct places and very highly purified (and hence expensive) enzymes.

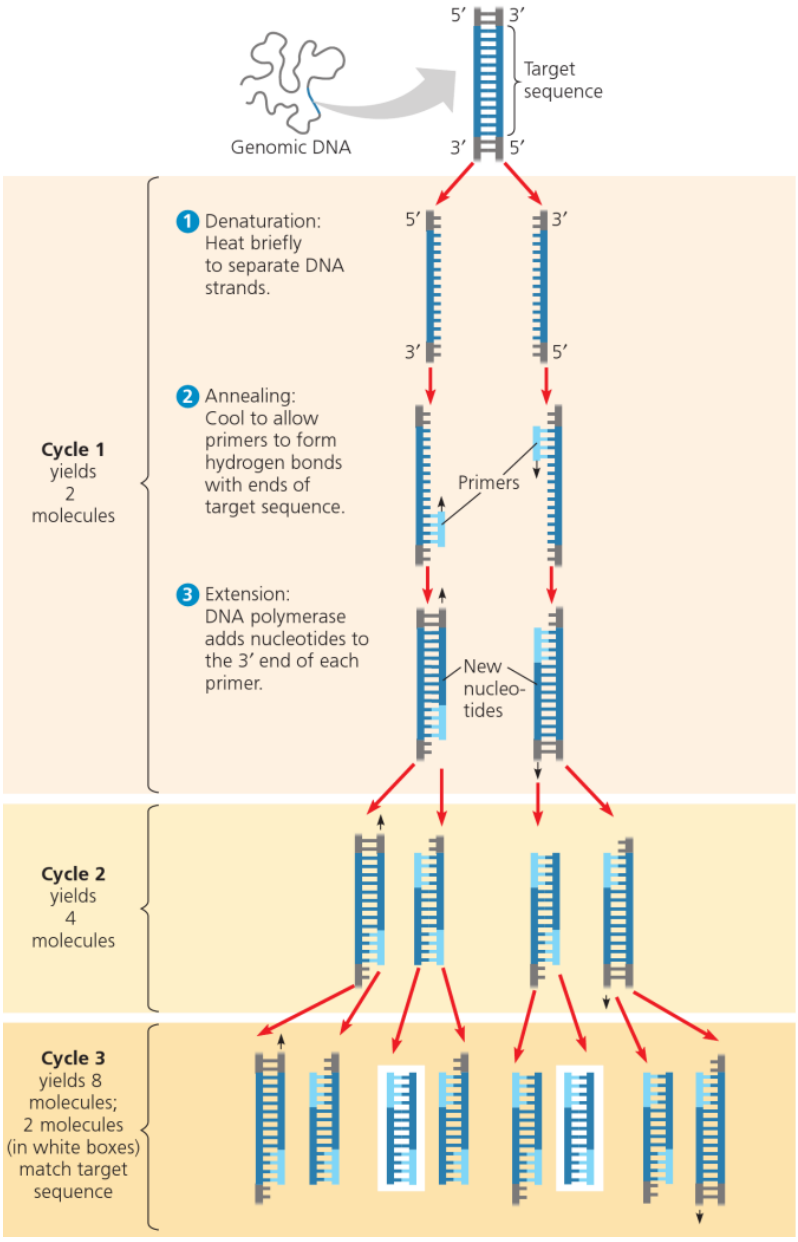
Limitations

4.2 Molecular Construction

Polymerase Chain Reaction (PCR)

Requirements

PCR is a technique which can be used to amplify single DNA molecules which extreme sensitivity. Note that **primers must be made** in order for PCR to work and **thermotolerant DNA polymerase** and dNTPs must be present under the correct conditions (a buffer including $MgCl_2$ is used).



Applications
Advantages & Limitations

PCR is useful for assays and also provides an alternative to restriction enzymes to forming parts of DNA. PCR is able to isolate fragments independently of restriction sites but there are length constraints on the sequence which PCR can amplify due to the processivity of the polymerase (how quickly the polymerase falls off). Furthermore, the GC content can be an issue for the thermocycling and the polymerase does have an error rate.

Seamless Joining

PCR can be used to seamlessly join different fragments by *extending primers to create overlaps between PCR reaction products* and then carrying out a PCR reaction on the products (noting that primers are still required because DNA polymerase only works in one direction). Whilst this does allow for seamless joining of arbitrary DNA fragments, only two fragments can be joining at a time and a size limit is imposed by the use of PCR in the final stage.

What can go wrong?

Several things can go wrong with PCR. Contamination can always be a problem (a laminar air flow hood and filter tips should be used) but assuming fresh, correct buffers & enzymes and that the thermocycler is correctly programmed, the following are common problems:

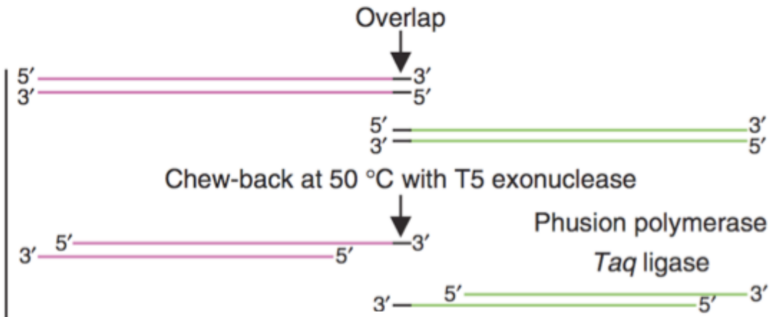
- No Product: Check primer design, decrease the annealing temperature (more likely to form base pairs) and try varying $MgCl_2$ levels.

- Multiple Products: Check primer design! Try using a hot start enzyme (only works at higher temperatures), increasing the annealing temperature and also varying $MgCl_2$ levels.
- Primer Dimers: The primers self-prime or prime on each each - redesign them!
- Incorrect Primer Concentrations: if too low, the PCR reaction will not work well.

Gibson Assembly

Gibson Assembly Advantages

The Gibson Assembly is a technique for joining DNA molecules which overlap by $\sim 20-40$ base pairs. It is able to combine several DNA fragments with very large lengths ($100+$ kilo-bases) though it is expensive for large lengths. It is cheaper than *de novo* synthesis and fast since it requires few steps and reagents.



De Novo Synthesis

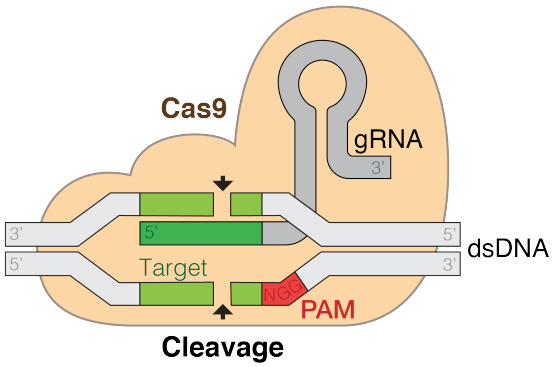
Advantages

It is also possible to synthesis DNA *de novo* completely independently of restriction sites and ability to PCR. The main limitation of this method is simply cost but this is dropping rapidly. In fact, PCR is not possible without primer synthesis and synthesis of 1kb is cost effective.

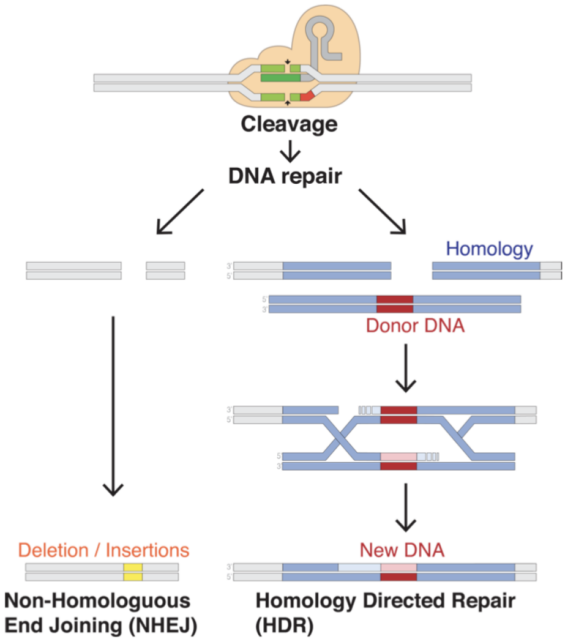
CRISPR/Cas9

How to program cut locations?

CRISPR is a family of DNA sequences in bacteria which contain DNA from viruses which have attacked the bacterium and are used by the bacteria to detect and destroy similar viruses. The CRISPR/Cas9 protein is able to find and cut a DNA target specified by the guide RNA (gRNA) with a specifity of 20 bases and **hence by changing the sequence of the gRNA cleavage locations can be programmed** noting that the target must be followed by a *protospacer adjacent motif(PAM)*.



This allows for homology directed repair.



4.3 Codon Optimisation

Justification?

Since the amino acid code is redundant, a choice of codons is made when engineering a sequence to express a protein. The codon usage of different organisms is also drastically different e.g. the codon usage in algae and jellyfish is 32% different .

Which codons to use?

A study in 2009 found that **favourable codons were predominantly those read by tRNAs that are most highly charged during amino acid starvation**, not the codons which are the most abundant in highly expressed proteins.

4.4 Case Study: Engineering Insulin

Insulin is an important human hormone and producing insulin in large quantities is desireable.

Insulin, a hormone produced in the pancreas controls cells through surface receptors. It increases liver, muscle and fat tissue to take up glucose for storage as glycogen and increases fatty acid synthesis.

Producing insulin naturally

Naturally, insulin creation occurs as follows:

1. Translation into a pro-protein.
2. Folding and oxidiation to form di-sulfide bridges occurs as does signal cleavage.
3. Transportation through the endoplasmic reticulum and the Golgi complex, following by packaging.
4. Central linking fragment is clipped resulting in the two other fragments bound by disulfide bonds, leaving a 51 amino acid monomer.

There are several issues when making insulin:

- The protein precursor (proprotein) is processed into a mature protein.
- The signal peptide is cleaved leaving no initiator methionine.
- The central fragment is cleaved with the polypeptides held with disulfide bonds which need to be formed.
- Mature insulin is a hexamer with Zn ions; the monomer-hexamer transition can be an issue.

The approach taken is to make the two polypeptide chains independently, creating *lacZ-A/B* fusion plasmids which are transformed into *E. coli*. After the beta-galactosidase is removed (from the *lacZ*), each chain is purified, mixed and cross-linked with disulfide bridges.

4.5 Engineering Antibodies

Requirements

It is desirable to produce a specific antibody that can be produced with a high yield. The following is required:

- High target specificity.
- Correct self-assembly i.e. folding into an active conformation.
- The antibody must not itself trigger an immune response.
- The antibody must be secreted.

The Mouse Method

One method for producing monoclonal antibodies is as outlined in Fig. 17. Once the antigen is injected into the mouse, the mouse will produce B cells which secrete the correct anti-bodies. Note that myeloma cells are immortal and the spleen cells contain required B cells meaning that after screening, large batches of the required antibody can be produced.

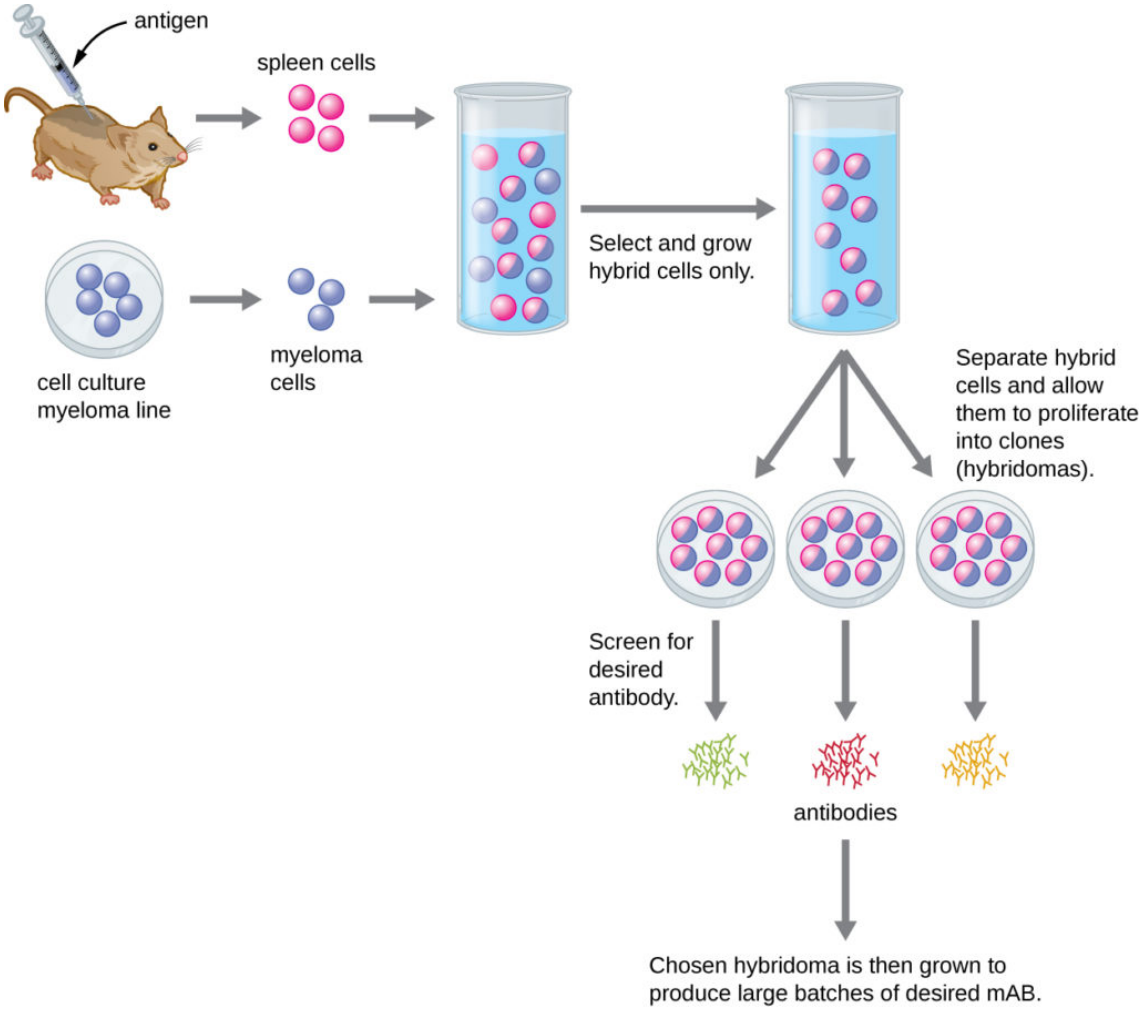


Figure 17: Producing monoclonal antibodies.

Humanising Antibodies

The antibodies produced must be humanised in order to prevent a immune response. For a humanised anti-body, the hypervariable/CDR regions of the mouse antibody must be combined to the rest of a normal human antibody. There are several ways to humanise antibodies.

1. CDR grafting uses the above approach, **replacing the CDR regions** of a human antibody.
2. Alternatively, a mouse carrying human antibody genes can be engineered and the produced antibodies will be human.
3. A more modern technique is to create a **phage display library** and then select the correct antibody from the library.

5 Metabolic Engineering

5.1 Preliminaries

Metabolism

The metabolism is the network of chemical reactions responsible for the breakdown of molecules for energy (catabolism) and the synthesis of new molecules for cellular function (anabolism). Catabolic bioprocesses are likely to generate energy and produce simpler, smaller building blocks whilst anabolic bioprocesses required energy and produce complex entities.

Precursor Metabolites

Precursor metabolites are intermediates formed by the catabolism and used by anabolic bioprocesses to form larger molecules. There are 12 precursor metabolites, including: `glucose-6-phosphate, fructose-6-phosphate, ribose-5-phosphate, pyruvate, ...`

Metabolic Pathway

A metabolic pathway is a set of reactions interacting under given physiological conditions via (apparently) simple intermediates. An intermediate is simple if there is a single pair of production and consumption reactions.

Role of Enzymes

Catalysis/Enzymes

Metabolic pathways are catalysed by enzymes which decrease the activation energy of reactions by providing an alternative reaction pathway. The active site of an enzyme is the region where substrate molecules are able to bind and undergo a chemical reaction.

There are two models for the action of enzymes. The lock and key model supposes that the enzyme and substrate are complementary to each other in shape and hence the substrate is also to fit in the enzyme. The induced fit model, there is some distortion of the enzyme as well as the substrate to allow the fit.

Linking Between the Catabolism & Anabolism

Activated carrier molecules

Activated carrier molecules serves as energy shuttles, linking the breakdown of food molecules to the energy-requiring synthesis of molecules required by the cell. There are a number of activiated carriers, including the following:

Activated Carriers	Group Carried in High-Energy Linkage
`ATP, GTP`	Phosphate
`NADH, NADPH,` $FADH_2$	Electrons, Hydrogens

Table 2: Some common activated carrier molecules

Redox Balance

Note that there must be *redox balance* in the cell; every time an electron is lost from a compound (oxidation), an electron must be gained elsewhere (reduction).

Important Pathways

Glycolysis

Glycolysis invests two molecules of `ATP` and a glucose molecule to give a net-gain of two `NADH & ATP` molecules as well as two pyruvate molecules. Glucose is converted to fructose which is then cleaved and processed.

Citric Acid/Krebs cycle

The Citric Acid/Krebs cycle is very important. The bond energy of a pyruvate oxidiation product, acetyl `CoA`, is harvested to form `NADH,` $FADH_2$ `& ATP` molecules and the reduced electron carriers can be used to form additional `ATP`.

In combination, eight of twelve precursors for secreted metabolites are formed from glycolysis and the citric acid cycle and hence these two cycles like at the core of the metabolism.

Regulation

Types of Regulation

Metabolic pathways can be regulated in different ways.. Hierarchial changes are those caused by changes in enzyme concentration through changes in mRNA processed (i.e. changes in enzyme activities, modifiers) while metabolic changes are those caused by changes in the concentrations of substrates, products or modifiers (changes in substrates, products).

1. Mass Action Regulation
   The law of mass action proposes that the rate of a chemical reaction is directly proportional to the product of the concentrations of the reactants.

2. Allosteric Regulation
   Allosteric regulation is the regulation of an enzyme by binding an effector molecule at a site other than the enzyme's active site.

3. Transcriptional Regulation
   A *transcription factor* is a protein that controls the rate of transcription by binding to a specific DNA sequence. Transcriptional regulation is where active transcription factors are modified and hence gene expression of enzymes is altered. Can be positive or negative feedback.

Enzyme Kinetics

Enzyme reactions are modeled using the Michaelis-Menten model i.e. a two stage reaction in which the substrate binds first and then the catalytic reaction occurs. Please see the course notes for 3G2. *Michaelis-Menten Model*

5.2 Metabolic Engineering

Purpose

> The purpose of metabolic engineering is to alter the metabolism of cells to enhance production of native metabolities (metabolism products) or to endow cells with the ability to produce new products.

Cells are either thought of as a factory which produce products that are deemed to be desirable. Alternatively, the modified cell may be the final product itself. These are a natural example of control loops with feedback and feedforward loops.

Building Cell Factories

When building cell factories, the aim is to achieve maximum high quality product secretion (which may require by-product control) at high yield. Several hosts can be used, such as microbes (bacteria and yeast), plants, algae and mammalian cell cultures. *Goal*

Metabolic Flux is defined as the rate of turnover of **molecules** through a metabolic pathway and hence in essence, we seek to control metabolic fluxes. In order to control fluxes, there are a number of different strategies which can be used: *Flux* *Strategies*

- Alter genes to remove feedback inhibition.
- Increase the concentration of key enzymes.
- Block metabolic pathways to divert fluxes.
- Introduce genes from other specifies which can create new pathways which can e.g. obtain non-endogenous products.

Note that if exogenous pathways are introduced, it is important to consider the implications of intermediates e.g. whether they are toxic.

Flux Balance Analysis

Flux balance analysis treats metabolic engineering as a constrained optimisation problem. The solution approach is as follows: *FBA*

1. Curate metabolic reactions from experimental evidence.
2. Formula the $\mathbf{S}$ matrix detailing reaction stoichiometry.
3. Apply the mass-balance constraints (i.e. multiple by a reaction rate vector, $\mathbf{v}$ and the steady-state condition).
4. Define an objective function, Z, as the dot product between $\mathbf{v}$ and a row vector detailing which reactions should be maximised.
5. Optimise Z using linear programming. Note that is it rate to have a single optimal solution point in $\mathbf{v}$.

This approach is not perfect; only one rate-limiting step is present under this regime but in fact, the flux of a pathway can depend on all of the rate constants. Measuring the slowest step (i.e. the step least able to go faster) is not easy. Each step has some degree of control over the flux.

Metabolic Control Analysis

An alternative to flux balance analysis, metabolic control analysis is a powerful tool. The approach is to use control coefficients which quantity the degree of control each step in a pathway has on the total flux and hence is a system variable; depends on the properties of all enzymes in the system.

1. Define the **flux control coefficient** as the fractional change in the pathway flux (J) caused by a fractional change in the activity of an enzyme (v) in that particular pathway. (Note that logs have been taken below). *FCC*

$$C^J_{v_i} = \frac{d \text{ ln } J}{d \text{ ln } v_i}$$

2. Define the **concentration control coefficient** as the fractional change in metabolite concentration (S) cause by a fractional change in the activity of an enzyme producing or consuming that metabolite (v). *CCC*

$$C^S_{v_i} = \frac{d \text{ ln } S}{d \text{ ln } v_i}$$

The above definitions lead to the following summation theorems:

1. Total control over a metabolic pathway flux should be distributed across all enzymes. *FCC Summation Theorem*

$$\sum_i C^J_{v_i} = 1$$

2. Control exerted by enzymes whose reactions produce a substrate should be equal to the control exerted by enzymes whose reactions consume that same substrate. (Note different signs) *CCC Summation Theorem*

$$\sum_i C^S_{v_i} = 0$$

Now, define additional coefficients.

1. The elasticity coefficient measures the fractional change in the rate of a reaction (v) to a fractional change in the concentration of one of its substrates, products or effectors (S). *Elasticity Coefficient*

$$\epsilon^v_S = \frac{\partial \ ln \ v}{\partial \ ln \ S}$$

2. The response coefficient measures how the pathway flux (J) responds to an effector (P) *Response coefficient*

$$R^J_P = \frac{\partial \ ln \ J}{\partial \ ln \ P} \qquad R^J_P = C^J_{v_i} \cdot \epsilon^{v_i}_P$$

The connectivity theorems link the kinetic properties of individual reactions to the system properties of a pathway. For enzyme, i, which responds to a metabolite, S, *Connectivity Theorem*

$$\sum_i C^J_i \cdot \epsilon^i_S = 0 \qquad \sum_i C^{S_n}_i \cdot \epsilon^i_{S_m} = 0 \quad n \neq m \text{ otherwise -1}$$

The FCCs of a network cannot be measured directly but are rather inferred by using the connectivity theorem with elasticity measurements. *Obtaining FCCs*

Metabolomics

The metabolome refers to the complete set of metabolites present. The transcriptome, proeome and metabolome are all context dependent and vary according to the state of the cell. The metabolome can be profiled by considering what is outside the cell (the exometabolome) but also what is inside the cell (the endometabolome). *Metabolome*

The metabolome can be investigated as follows:

1. Sample collection and preparation.
2. Instrumental analysis e.g. NMR. There are many here; sizes vary over 2 orders of magnitude, abundance varies over many more orders of magnitude. There are also a large number of structural isomers and many metabolites have very different chemical properties (e.g. polarity, volatility). *Instrumental Difficulties*
3. Preprocessing (e.g. noise removal) following by pattern recognition (may be supervised or otherwise).
4. Validation, both statistically and biologically.

Overall Method

In effect, the overall method can be interpreted as follows. A cell culture is analysed (transcriptomics, proteomics and metabolomics) and flux analysis is carried out. Information learn from here is used to alter gene targets and cell strains are then selected and transformation occurs and the cycle repeats. There are some caveats:

5.3 Biofuels

There are several generations of Biofuels.

1. First generation biofuels refer to fuels which have been derived from sources such as starch, sugar, animal fats and vegetable oil e.g. biodiesel. *1st Gen.*

   These types of biofuels threaten food supply and biodiversity and result in feedstock price volatility. There is also significant issue with the amount of land required. *Issues*

2. Second generation biofuels are produced from lignocellulosic (plant dry matter), agricultural residues (material left after crops have been harvested) or other forms of waste. The process to produce biofuel with these sources is more complicated but avoids some of the issues with 1st Gen. biofuels. This type of biofuel is very young and there tend to be high capital costs with some domestication issues with certain feedstocks. *2nd Gen.*

3. Third generation biofuels focus on non-arable land e.g. algae or cyanobacteria which require sunlight, CO_2 and water to produce biomass. Similarly there are domestication issues with high capital costs. *3rd Gen.*

4. Fourth generation biofuels are similar in technology to the previous generation but **biomass is not destroyed** e.g. photosynthesis mimicry to produce hydrocarbons. *4th Gen.*

The choice of feedstock for biofuel creation is important. Different feedstocks will have different proportions of their main sugar substrates and hence will affect the metabolic engineering work required.

6 Genome Sequencing

Genome sequencing aids species investigation; it allows finding lesions underlying disease, identifying individual sequence variations and comparing organisms.

6.1 Genome Structure

The genome is an organisms complete set of DNA. The structure of the genome is important. The genome of a typical prokaryote is circular with all regions coding for proteins; there is a very high gene density. Eukaryotes have much larger genomes and larger parts of the DNA do not code for RNA or protein. *What is the Genome?*

95% of the human genome is intergenic sequences. 50% is repetitious regions whilst there are also mobile elements (i.e. selfish DNA which form additional copies of itself within the genome whilst giving no survival advantage). The GC content also varies throughout and is linked to the gene density. In terms of components, within the genome there are: *What's in the Human Genome?*

- Protein coding genes.
- Structural RNA genes e.g. rRNA, tRNA.
- Regulatory sequences that affect transcription, replication and splicing. Such sequences are controlled by signals which are small $7 - 20$bp and these signals are often conserved between mechanisms.

A genome is made up of exons which code for a protein or peptide sequence and introns which do not code for proteins and interrupt gene sequences. *Exons and Introns*

Across different species, there tends to be conservation of: *Conservation*

1. Gene order.
2. Regulatory Motifs.
3. Protein sequence. Hence most changes tend to be **conservative** e.g. changing the codon which is used but not the amino acid sequence.

Often in smaller genomes of different species, significant portions of the repeating regions have been lost. Note that when comparing different genomes, gene number is a poor method of assessing complexity.

6.2 Sequencing Techniques

Sanger DNA Sequencing

Sanger DNA sequencing is able to sequence up to about 900 base pairs. The technique requires the following ingredients:

Sanger DNA Sequencing ingredients

1. A DNA Polymerase enzyme.
2. DNA Nucleotides, `dATP, dTTP, dCTP, dGTP`.
3. A DNA Primer.
4. Population of the DNA template to be sequenced; a population of a molecule only gives one read.
5. **Dideoxy** versions of each nucleotide e.g. `ddATP` **each labeled with a different dye**.

ddNTP

Figure 18: Dideoxy nucleotide versions.

Once a dideoxy nucleotide has been added to a chain, there is no hydroxyl group available and hence no further nucleotides can be added. The chain ends with a **dyed nucleotide**.

Hence, a the DNA sampled is heated to anneal the DNA and cooled to allow the primer to bind. The temperature is raised to allow DNA polymerase to synthesise new DNA and it is virtually guaranteed that a dideoxy nucleotide will be at every position of the DNA. Gel electrophoresis is used to separate out the molecules and hence the sample can be sequenced.

Method

Whilst this method gives high-quality sequencing, it is expensive, time consuming and inefficient for sequencing larger projects e.g. the whole genome.

Evaluating Sanger

Next Generation Sequencing

Next generation sequencing, such as Illumina, is able to read far faster than Sanger sequencing. See https://www.youtube.com/watch?v=fCd6B5HRaZ8 for an excellent video. Fig. 19 outlines the general method for this technology but the important points are important:

- Adapters to each DNA fragment are added which not only allow reuse of the same primer but also allow sample identification.

Adapters: Why?

- After bridge amplification, only the forward strands are kept. After the read, bridge amplification occurs again and the same strand is then read again backwards.
- Many sequences are sequenced at once.
- There are about 1000 copies of DNA in each cluster; this aids digital imaging.
- Computer controlled imaging is used to sequence millions of clusters in parallel.
- **Crucially, the bases are extended by one base each step** after the primers have annealed with different colours corresponding to different bases. This is achieved by using dyed `NTP`s which cannot be extended. After imaging, the terminator and dye are cleaved from the `NTP` and then extended by an additional base pair.

Extending one bases at a time

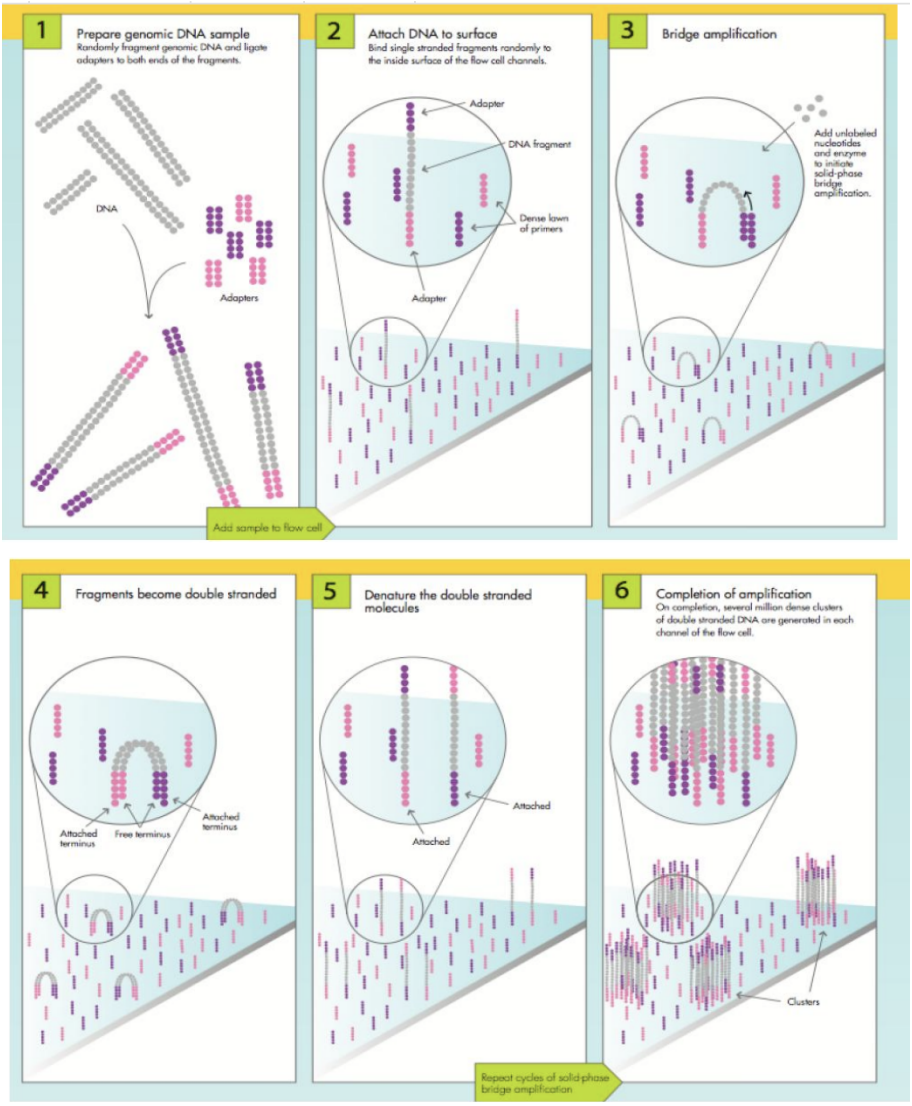


Figure 19: Illumina: A Next Generation sequencing technology.

Compared to Sanger sequencing, Illumina sequencing does have a shorter read length but there are $3000 \cdot 10^6$ reads per run and $300 \cdot 10^9$ bases per single end run. Unlike the Sanger method, each sequence corresponds to a read. It is expensive and time consuming; the 'sequencing by synthesis step' can take up to 8 days and the software analysis up to 2 days.

6.3 Whole Genome Sequencing

A genome of say 10^9 bases is hard to sequence when only a limited number of bases can be read at a time. One approach is to:

1. Fragment the genome.
2. Sequence the fragments.
3. Assemble the genome by searching for overlapping fragments.

Whole Genome Shotgun Sequencing

Shotgun Sequencing

Initially controversial, this is now the standard approach. DNA is broken up randomly into numerous small sequences which can then be sequenced e.g. using the Sanger method. **Several rounds of fragmentation and sequencing** are performed leading to multiple overlapping reads for the target DNA and the overlapping reads can then be assembled into a continuous sequence (through software).

Problems?

There are issues: there is a cloning bias since different DNA fragments have different affinities for cloning and assembly has potential for huge mistakes with difficult repeating systems and was computationally hard.

6.4 Libraries

Library

A library is a collection of **cloned** DNA fragments; each clone is fused with a vector (often a plasmid) and can be transformed into host cells. Vector sequences can be used for selection of cells which contain cloned DNA. A 'good' library has the following properties:

- Even source material representation.
- Few empty vectors.
- DNA not rearranged.
- No vectors with multiple inserts.

Production

Libraries are made by isolating the desired fragments and then ligating them into a vector. Two common libraries are cDNA libraries (gene libraries produced using a reverse transcriptase enzyme which copies mRNA to DNA) and genomic DNA. Source fragments can be synthesised, creating using different restriction enzymes or randomly sheared.

There is a maximum insert size since a larger insert tends to decrease stability and give a poor transformation efficiency. Cosmid inserts can give ~ 30kb inserts whilst more modern bacterial artificial chromosome inserts give around $100 - 150$kb inserts.

6.5 Finding Genes

Approaches

There are multiple approaches to finding genes once a genome has been sequenced.

- Human and vertebrate `cDNA` (complementary DNA formed from single stranded `mRNA` in a reaction catalysed by reverse transcriptase) can be aligned to the genome.
- The genome can be compared to known protein, finding similarities.
- *Ab initio* i.e. using statistical gene finders.

Sequence Alignment

Why important?

Sequence alignment is very important across molecular bioengineering. It it often used to

- match mRNA to genome sequence and match protein sequences to a genome.
- find overlapping sequences required for genome sequence assembly.
- find where PCR primers may be mis-priming.
- identify families of related protein sequences.

To develop methods of aligning sequences, the presence of gaps and amino-acid substitutions must be considered.

Amino-acid substitutions are scored using the log-likelihood with a large set of high quality ungapped protein sequence alignments.

$$\text{score}(a,b) = \log(\frac{p_{ab}}{p_a p_b})$$

The log likelihood score is positive if aligned more often than expected and vice versa. The score is often rounded to the nearest integer for computational efficiency.

Gaps tend to occur in runs. There are different penalities which can be used to score gaps. For a gap length of g, the following scoring methods are often used:

$$\begin{aligned} &\text{Linear: penalty} = -d \cdot g \\ &\text{Affine: penalty} = -d - e(g-1) \qquad e < d \end{aligned}$$

In the linear case, d acts a per-residue gap penalty. In the affine case, there is a larger price to pay for 'opening' a gap and hence an extension penalty is used which is lower.

Dynamic Programming

Dotter

It is possible to compare every residue in one sequence to every residue in another i.e. check every possible alignment. In time and memory, for sequences of length M and N, the complexity is $O(M \times N)$. Diagonal lines indicate matching sequences but this does not allow for gaps. An alignment with gaps will have a downwards and rightward trend but note that there **may be a gap in either of the sequences**.

Algorithm Steps

To account for gaps, dynamic programming is used.

1. Fill out a matrix of scores with each cell maximising its score by examining the three already evaluated neighbours. It will inherit the neighbours score and either pay a penalty for a gap or increase in score due to a good alignment.

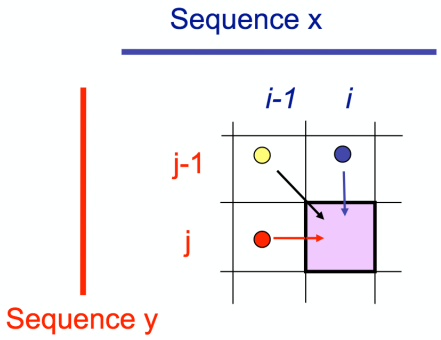


Figure 20: $F(i,j)$ is the score of the best alignment of subsequences $(x_1, x_2, ...x_i)$ and $(y_1, y_2, ...y_j)$. The algorithm is recursive and depends on previously evaluated values.

With the top most and left most scores filled, the remainder of the matrix is filled in:

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + \text{score}(x_i, y_j) \\ F(i, j-1) - d \\ F(i-1, j) - d \end{cases}$$

For each cell, a traceback pointer records from which parent the best score was inherited.

2. To generate the alignment, each traceback point is followed backwards from the final cell evaluated.

This algorithm has the same complexity as the dotter. There are many variants such as the best overlap, best local alignment and there are linear memory variants of the algorithm. An alignment can be considered to be a path through a finite state machine.

Using this in practice

This technique is used to find genes by aligning `cDNA` and can also be used to search for proteins. However, there are huge gaps and the method must be tolerant. Since this would be costly to solve via dynamic programming, dynamic programming is used to piece exons **once they have been roughly located using heuristic methods**.

Ab Initio Gene Finding: GenScan

Complexity

GenScan simultaneously finds forward and reverse genes. It is a probabilistic model which includes models for exons with $O(\text{States} \cdot M)$ (which is very fast). However, nested genes are missing and there is no alternate splicing.

Similarly to dynamic programming, this model can be represented as a finite state machine but is far more complex. The FSM is symmetric since the model is able to read genes from the forward and reverse strand simultaneously. The model effective chooses genes which align best with model evidence and many heuristics are used.

6.6 Investigating a Gene

Investigating Genes

There are many methods for investigating genes. Directed approaches include removing the gene, over-amplifying it or directing mutation of the gene. Alternatively, random mutagenesis can be used (i.e. change the genome and screen for phenotypes) or error-prone PCR used to randomly alter specific regions of DNA. In fact, RNA can be used to repress genes through RNA interference; `dsRNA` is expressed as smaller interfering RNA binds to `mRNA`. Alternatively, CRISPR/Cas9 can also be used for gene knockout.

RNAi