DNA Sequence and Genetic Variation (W1)

From DNA sequence to disease states (L1)

What is Bioinformatics?

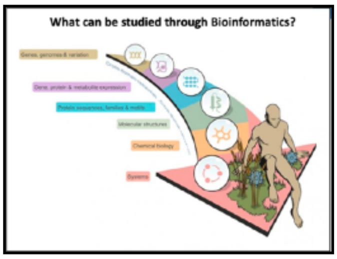
- In short Bioinformatics is “any application where computers are used to process, store and **analyse biological data.**”
- Bioinformatics is the application of computer technology to the understanding and effective use of biological and biomedical data
  - It involves the storage, analyses and interpretation of the big data generated by life-science experiments, or collected in a clinical context
- Bioinformatics = Biology + Statistics + Computer Science

Why do we need bioinformatics?

- The amount of **biological data** we’re trying to **analyse** is **huge** - sometimes called “big data”
  - Explosion of publicly available genomic information and other biological information DNA and RNA sequencing produce large amounts of data
- Bioinformatics helps us to identify patterns in biological data
  - We need tools to analyse the biological data
    - The human genome is made up of 6 billion bases - where are the important bits?
  - Bioinformatic insights often require validation through lab based experimentation

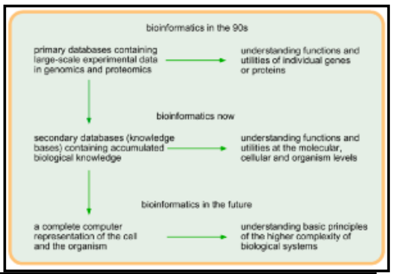
What can be studied through bioinformatics?

- Genes, genomes and variation
- Gene, protein and metabolite expression
- Protein sequences, families and motifs
- Molecular structures
- Chemical biology
- Systems


What can be studied through Bioinformatics?
Genes, genomes & variation
Gene, protein & metabolite expression
Protein sequences, families & motifs
Molecular structures
Chemical biology
Systems

How can bioinformatics help us understand molecular biology in a way that was previous impossible?

- Today we have access to huge amounts of rich biological data and the tools to extract meaning out of it
  - Used to be too costly and or didn’t have the techniques to sequence genomes, transcriptomes and proteomes of organisms
  - Structural biology techniques give us a high resolution picture of proteins and protein complexes
  - Computer power available today helps us make sense of the biological data (Extract meaningful information)


bioinformatics in the 90s

primary databases containing large-scale experimental data in genomics and proteomics → understanding functions and utilities of individual genes or proteins

bioinformatics now

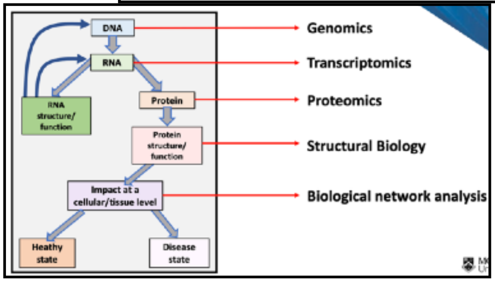
secondary databases (knowledge bases) containing accumulated biological knowledge → understanding functions and utilities at the molecular, cellular and organism levels

bioinformatics in the future

a complete computer representation of the cell and the organism → understanding basic principles of the higher complexity of biological systems

How can bioinformatics help us understand molecular biology?

- Genomics
- Transcriptomics
- Proteomics
- Structural Biology
- Biological network analysis



What is bioinformatics being used for?

- The capacity to analyse bioinformatic data has opened up a world of possibilities
- “The keys to the kingdom of molecular world”
  - Categorising genetic variants associated with disease
  - Enhancing our understanding and our ability to detect and treat disease
  - Investigating the evolution and spread of a microorganism
  - Enhancing our understanding of complex biological systems
  - Identify desirable properties of plants that could allow the development of environmentally sustainable solutions for food and energy production

  - Drug design enabled by bioinformatic tools
  - Genetic engineering, gene therapy, gene editing

How can genomics help us understanding molecular biology?

- **Genomics (DNA)**
  - **Starting point**: A single genome or multiple genomes from the same organism
  - **What is a genome?**: The complete set of genetic information in an organism that is housed in the chromosomes
- Bioinformatic analysis allows us to pose and answer questions like:
  - Where are the genes located in the genome?
  - Where are the mutations that cause disease located within the genome
  - Which variation in the genome actually matters?
  - Which genes or regions of genes do we share with other organisms?

How can transcriptomics help us understanding molecular biology?

- **Transcriptomics** (**RNA**)
  - **Starting point**: A single transcriptome or multiple transcriptomes from the same organism
  - **What is a transcriptome?** The sequence of each RNA transcript present in a cell or group of cells at a point in time
- Bioinformatic analysis allows us to pose and answer questions like:
  - How many different variations of a gene transcript exist?
  - What RNAs do different types of cells produce at different points in time
  - How is the transcriptome of disease and healthy tissues different
  - How is the transcriptome of a drug treated vs a non-drug treated tissue different?
  - How much of a particular mRNA, or set of mRNA, is being produced, is it within normal limits?

How can proteomics help us understanding molecular biology?

- **Proteomics** (**Protein**)
  - **Starting point**: A single proteome or multiple proteome from the same organism
  - **What is a proteome?**: The sequence of each protein present in a cell or group of cells at a point in time
- Bioinformatic analysis allows us to pose and answer questions like:
  - How many different variants of a protein exist?
  - What proteins do different types of cells produce at different points in time?
  - How is the proteome of diseased and healthy tissue different?
  - How is the protein of a drug treated vs a non-drug treated tissue different?
  - How much of a particular protein, or set of proteins, is being produced, is it within normal limits?

How can structural biology help us understanding molecular biology?

- **Structural Biology (Protein structure/function)**
  - **Starting point**: A protein or collection of proteins
  - **What is structural biology?**: The study of molecular structure and dynamics of biological macromolecules
    - Protein structures are the **most studied**
- Bioinformatic analysis allows us to pose and answer questions like:
  - What is the function of a protein? Can be deciphered from a protein's domain's, ligand binding site and 3D structure
  - Which regions of a protein's sequence are functionally important and conserved? Multiple sequence alignments, BLAST
  - What other molecules does the protein interact with?
  - How do common mutations that lead to changes in protein structure contribute to disease? How do these mutations alter the function or 3D folding of the protein?

How can biological network analysis help us understanding molecular biology?

- **Biological network analysis (impact at a cellular/tissue level)**
  - Not covered in great detail in this unit
- Everything that happens inside the cell relies on multiple components interacting together

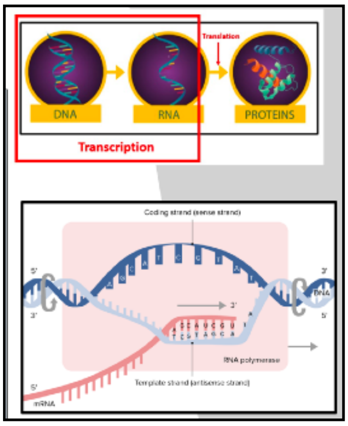
- Biological network analysis attempts to map and understand the interactions between the components of the cell that impact cellular function
  - Pulls data from many different data repositories to begin to decipher these networks
  - **Some of the most common types of biological networks are:**
    - Protein-protein interaction networks
    - Metabolic networks
    - Genetic interaction networks
    - Gene/transcriptional regulatory networks
    - Cell signalling networks

How can biological network analysis help us understanding molecular biology?

- **Knowledge of the relationship between DNA, RNA, protein and cellular function allows us to better understand health and disease states (Disease State)**
- Can attempt to answer questions such as:
  - What mutations always result in disease?
  - What mutations sometimes result in disease
  - How does the protein of interest in individuals with the disease differ from those who have the native protein?
  - What causes the disfunction in the protein?
  - How can we correct mistakes in the genetic material that cause the disease? Can we compensate for the dysfunction caused?
  - How can we determine who is likely to exhibit the disease?
  - How can our knowledge of the molecular basis of the disease inform management and treatment strategies?

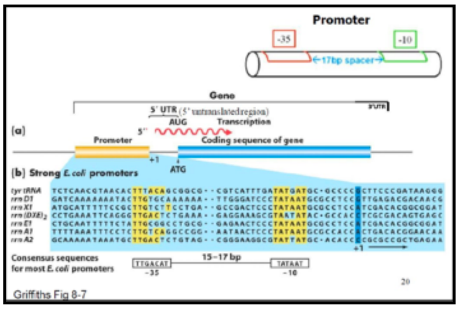
DNA contains the blueprint for every living thing on earth

- DNA is the instruction manual for how to make an organism
- Gene —(**transcription**)—> mRNA
  - Copy the chapter into a disposal copy
- mRNA —(**translation**)—> chain of amino acids
  - Translate the language of the book to make something useful
- Chain of amino acids —(**protein folding**)—> 3D folded protein
  - Amino acid interactions
  - Chaperones
  - Low energy state
- 3D folded protein —> Protein **executes function**
  - 3D structure allows protein to perform function
  - Helps maintain normal cellular activity
    - Thousands of different proteins executing their function



What is required for transcription to take place?

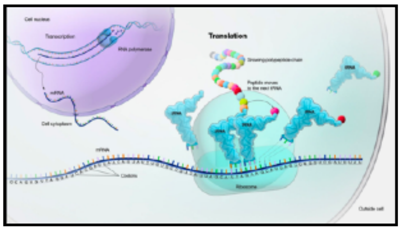
- DNA strands need to be **accessible** in the nucleus
  - **Epigenetic** mechanisms promote **decondensing** of **specific regions** of **DNA** to allow genes to be expressed (transcribed)
- **RNA polymerase** must bind to the **promoter region** of the gene to initiate **transcription**
  - **TATA boxes** at **-35 and -10** (in eukaryotes), similar binding sites in prokaryotes
  - Requires help of additional transcription factors that bind to TATA box first
- **Template** and **coding** DNA strand needs to come apart (**DNA helicase**)
  - Allows RNA polymerase II to make an mRNA copy of the template strand
    - The need strand is created in the 5' to 3' direction
- **Reservoir** of RNA **nucleotides**
  - RNA nucleotides sequentially recruited to **match** to the **coding strand** and fused together via RNA ligase to create the RNA transcript


Promoter
-35
-10
←17bp spacer→
Gene
5' UTR (5' untranslated region)
3'UTR
AUG
Transcription
(a)
5'
Promoter
Coding sequence of gene
+1
ATG
(b) Strong E. coli promoters
tyr tRNA
rrn D1
rrn X1
rrn (DXE)2
rrn E1
rrn A1
rrn A2
TCTCAACGTAACACTTTACAGCGGCG--CGTCATTTGATATGATGC-GCCCCGCTTCCCGATAAGGG
GATCAAAAAAATACTTGTGCAAAAAA--TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
ATGCATTTTTCCGCTTGTCTTCCTGA--GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
CCTGAAATTCAGGGTTGACTCTGAAA--GAGGAAAGCGTAATATAC-GCCACCTCGCGACAGTGAGC
CTGCAATTTTTCTATTGCGGCCTGCG--GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
TTTTAAATTTCCTCTTGTCAGGCCGG--AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
GCAAAAATAAATGCTTGACTCTGTAG--CGGGAAGGCGTATTATGC-ACACCCCGCGCCGCTGAGAA
+1
Consensus sequences for most E. coli promoters
TTGACAT
-35
15–17 bp
TATAAT
-10
20

Griffiths Fig 8-7

What is required for transcription to take place?

- Key points:
- **A gene** consists of:
  - Promoter
  - 5’ and 3’ Untranslated region (UTR)
  - Coding region (+ introns if eukaryotic)
  - Terminator sequence (within 3’ UTR)
- **An mRNA** consists of:
  - 5’ and 3’ UTR
  - Coding sequence (introns removed in mature mRNA)
  - STOP codon (AUG)
  - Terminator sequence (within 3’ UTR)

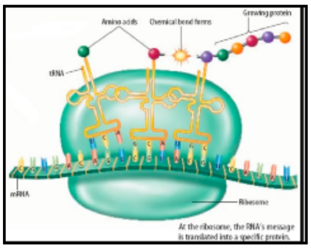


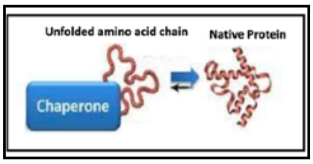
Connecting transcription to translation

- Mature mRNA (without introns) leaves the nucleus
- **Ribosomes** in the cytosol or endoplasmic reticulum (ER) **bind the mRNA** through recognition of the 5’ methyl-guanosine cap (in eukaryotes)
  - Translation of mRNA can begin to build the amino acid chain

What is required for translation to take place?

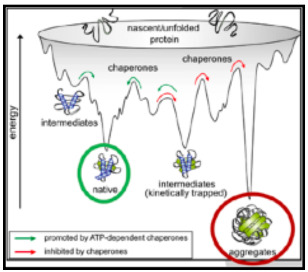
- **Mature mRNA to bind to a ribosome**
  - Ribosomes in cytosol and ER are the molecular machines that enable translation
- **tRNAs to recruit the “correct” amino acid to build the amino acid chain**
  - Anticodon of tRNA recognises mRNA codon - matched via complementary base pairing
  - tRNA’s with specific anticodons carry specific amino acids
    - Human cells have between 4-60 different types of tRNAs to recognise the 61 non-stop codons
- **A stop codon in the correct position**
  - Once the ribosome recognises a stop codon in the mRNA strand it disengages, ceasing translation
  - Vital that the STOP codon is in the correct position to form a functional protein
- What happens **after translation**?
  - The amino acid chain created through translation, aka polypeptide, is folded up into a specific 3D shape that allow the protein to perform its specific function


Amino acids
Chemical bond forms
Growing protein
tRNA
mRNA
Ribosome
At the ribosome, the RNA's message is translated into a specific protein.



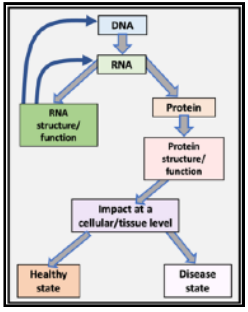
What happens after translation?

- The amino acid chain folds into a 3D protein through:
  - Interactions between amino acids
  - Molecular chaperones (prevents aggregation)
  - Sampling different conformation in an attempt to achieve a low energy state



What impact do changes have on other biological levels?

- Change that occur at one biological level are passed on or have consequences for the levels below



Changes at the DNA level are passed onto other biological levels

- Changes at **DNA level** passed onto other levels:
  - Mutations (base changes) are copied from DNA to mRNA (transcription)
  - Can also change intron-exon boundaries
- The **protein** generated in translation (unless it is a silent mutation) is **altered**
  - Called a mutant protein
- Mutant proteins may be:
  - Fully functional, partially functional, non-functional, constitutively active
- If the protein behaves differentially altering how it contributes to cellular processes -> **disease**

Changes at the DNA level are passed onto other biological levels

- Mutations in the regulatory region of a gene can alter **gene expression**:
  - Under expression of gene
  - Overexpression of gene
  - No expression of gene
- Altered amount of a specific protein produced can impact how it contributes to cellular processes -> **disease state**

Changes at the RNA and protein level are passed onto other biological levels

- RNA dysfunction can be caused by **mutations** in **RNA binding proteins** (stabilise mRNA during translation)
  - Reduced levels of functional proteins due to translation operating sub-optimally
  - Lead to impact cellular/tissue level -> disease state
- Proteins can undergo **misfolding** or **aggregation**
  - Misfolded and aggregated protein is non-functional
    - Can cause serious issues for the cells e.g. neurodegeneration

How can bioinformatics help combat disease?

- Identify where genetic variants associated with disease are located
- Establish how genetic variants actually cause disease
- Detect which individuals are susceptible or have a high likelihood of being susceptible to a disease
- Create treatment and management practices for diseases

DNA sequencing enables **genomics** where whole genomes are sequenced:

- Can identify genetic variations in comparison to a reference sequence
- Can identify pathogenic and predisposing mutations
- Not all variations cause or predispose to disease
- If interested in a specific gene or genes these can be sequenced rather than the whole genome

Susceptibility

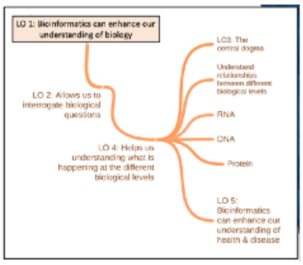
- Identification of pathogenic mutations
- DNA sequencing, Genomics
- Screening for mutations that contribute to polygenic diseases
  - How many of these mutations are present?

Causing Disease

- Compare healthy vs disease tissues
  - **Comparative transcriptomics**
  - **Comparative proteomics**
- How is the mutation impacting the mRNA and protein coded for by the gene and those it associates with through its normal function?
  - **Transcriptomics + proteomics**
- How has the mutation impacted the proteins structure and its function? What are the knock-on effects of cellular function?
  - Structural **bioinformatics**

Treatment of diseases

- Understand the mechanism of disease to a point where existing therapeutics can be applied effectively
- Screen drug candidates to assess their capacity to treat disease while minimising side effects
  - **Comparative transcriptomics**
  - **Comparative proteomics**
- Understand the mechanism of disease of designing drugs, protein therapeutics (biologics) or gene therapy
- Identify which existing medicines are most likely to be effective based on genomic variation profile present - precision medicine
  - **Genomics**


LO 1: Bioinformatics can enhance our understanding of biology
LO 2: Allows us to interrogate biological questions
LO3: The central dogma
Understand relationships between different biological levels
RNA
DNA
Protein
LO 4: Helps us understanding what is happening at the different biological levels
LO 5: Bioinformatics can enhance our understanding of health & disease

Why does the variation in our genetic material matter? (L2)

What is a DNA mutation

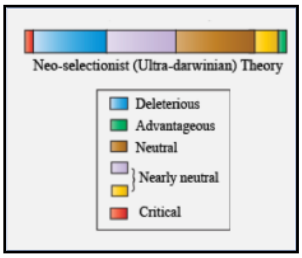
- A mutation is a change or alteration in the nucleotide sequence of an organisms DNA
  - This is a change from the most prominent DNA sequence found in the organism, often called wild-type
- Mutations are often referred to as being within genes but can also occur in other regions of the genome
  - Wild type gene contains the most frequently observed sequence in the organism
  - Mutant gene = a gene that contains one or more mutations changing the nucleotide sequence from the wild-type gene

At what biological level do mutations occur?

- **Mutations occur within the DNA of a single cell** = one or more nucleotides within the cell's genetic material changes (e.g. A to T)
  - **Somatic mutations** are not passed onto the next generation, but is passed onto other cells originating from the altered cell (mitosis)
  - **Hereditary mutations** are those passed from the parent to the child via the sperm and egg cells
    - Present in all of the parents cells - i.e. not mutations picked up through their lifetime
  - Mutations not found in the parents somatic cells but that originate in the child and usually preset in all cells = **de novo mutation**
    - Mutations can arise in the sperm or egg cells or after fertilisation
    - Mosaicism possible where some cells have mutation and some don't
  - **Non-hereditary mutations**
    - Environmental and spontaneous mutations

What is the significance of DNA mutation? Why does it matter?

- Most mutations in the genome are **neutral** or nearly neutral and do not have an impact
- But other mutations can be:
  - **Deleterious** >>> possibly pathogenic
  - **Critical** >>> definitely pathogenic (small no. of total)
  - **Advantageous** (small no. of total)



What causes new mutations to arise?

- Spontaneous mutations
  - Main source is errors during DNA replication
- Non-spontaneous mutations - induced by mutagens
  - UV radiation
  - X-ray radiation
  - Tobacco
  - Certain chemicals - carcinogens, mutagens
  - Nitrites - present in processed meats
  - Viruses and bacteria

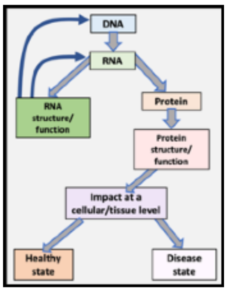
How likely is it that mutations will be repaired?

- Spontaneous mutations are the main source of new mutations
  - DNA polymerase makes a mistake in 1 in very 100,000 nucleotides during DNA replication = 120,000 mistakes per cell division
- Repair of spontaneous DNA mutations
  - DNA polymerase proof reading can correct 99% of errors
  - Mismatch repair mechanism also assists in correcting errors
- Unprepared mutations seen as normal in the next round of cell division = can't be repaired
- A 15 year old the somatic cells, the genes are predicted to have taken on:
  - 100-1000 spontaneous mutations in non-replicating cells
  - 1000-10,000 spontaneous mutations replicating cells
  - 4,000-40,000 spontaneous mutations replicating cells by the age of 60
- When mutations accumulate in proto-oncogenes or tumour suppressor genes the cells become cancerous

  - Single proto-oncogene — mutations cause gain of function — loss of cell cycle control = cancer
  - Two homologous tumour suppressor genes — loss of cell cycle control = cancer

<u>Why didn’t we evolve to not pick up new mutations?</u>
- Mutation is evolutionarily advantageous
  - Allow organism at the population level to adapt to a changing environment
    - If new mutations confer a reproductive fitness advantage, they are more likely to be passed onto the next generation
- Without new mutations appearing in the population, a population is vulnerable to changes in environmental conditions
  - e.g. appearance of a new deadly pathogen

<u>Mutation Pathway</u>
- Mutations within the protein coding region of a gene (DNA) are passed onto the RNA level through transcription
- Changes in the mRNA transcript alter the amino acid sequence (unless they are silent mutations)
- Changes to the amino acid sequence may alter protein folding and/or function



<u>What can a mutation potentially change?</u>
- **A single point mutation can cause a change to one of the amino acids in the protein sequence**, or a base may be added or deleted causing more drastic changes to the protein sequence
- Mutations can introduce a **premature stop codon** >>> gene produces a shorter mRNA which in turn generates a shorter (truncated) protein which is unlikely to be functional (**nonsense**)
- An **exon** region of a gene could be **excluded** or an **intron included** in mRNA transcript
- Mutations can alter the amount of mRNA being transcribed or the amount of the protein being translated. **Both mRNA and protein can be under or overexpressed.** In regulatory regions of DNA
- Results in a mutant protein

<u>What can a mutation potentially change?</u>
- Single point mutation can use a change to one of the amino acids in the protein sequence
- **DNA** level: Single point mutation changes the DNA sequence of a gene
- **RNA** level: Changes the mRNA sequence transcribed form the gene
- **Protein** level: changes the amino acid sequence of the protein created through translation (unless it was a silent mutation)

<u>What can a mutation potentially change?</u>
- **Mutant proteins often function differently to native (wild-type) proteins**
  - Mutant proteins result from mutation
- **A mutant protein may be:**
  - Non-functional
  - Partially functional
  - Gain a new or enhanced function
  - Function normally:
    - If the altered amino acid didn’t affect folding or a functionally important area
    - If a similar amino acid was added in place of the original amino acid
- When a mutant protein does not perform the role of the native protein or performs different function:
  - Knock on consequence for cellular processes
  - This can lead to disease state

<u>Example</u>
- The XPA is a DNA repair protein involved in nucleotide excision repair
- If it were mutated and became partially functional or non-function = unable or less able to contribute to DNA repair