Unit II

Computational Genomics and Transcriptomics

Computational genomics :

Computational genomics refers to the use of computational and statistical analysis to decipher biology from genome sequences and related data, including both DNA and RNA sequence as well as other "post-genomic" data (i.e., experimental data obtained with technologies that require the genome sequence, such as genomic DNA microarrays). These, in combination with computational and statistical approaches to understanding the function of the genes and statistical association analysis, this field is also often referred to as Computational and Statistical Genetics/genomics. As such, computational genomics may be regarded as a subset of bioinformatics and computational biology, but with a focus on using whole genomes (rather than individual genes) to understand the principles of how the DNA of a species controls its biology at the molecular level and beyond. With the current abundance of massive biological datasets, computational studies have become one of the most important means to biological discovery.

Genome comparison :

Computational tools have been developed to assess the similarity of genomic sequences. Some of them are alignment-based distances such as Average Nucleotide Identity.[7] These methods are highly specific, while being computationally slow. Other, alignment-free methods, include statistical and probabilistic approaches. One example is Mash,[8] a probabilistic approach using minhash. In this method, given a number k, a genomic sequence is transformed into a shorter sketch through a random hash function on the possible k-mers. For example, if �=2 , sketches of size 4 are being constructed and given the following hash function

(AA,0)	(AC,8)	(AT,2)	(AG,14)
(CA,6)	(CC,13)	(CT,5)	(CG,4)
(GA,15)	(GC,12)	(GT,10)	(GG,1)
(TA,3)	(TC,11)	(TT,9)	(TG,7)

the sketch of the sequence

CTGACCTTAACGGGAGACTATGATGACGACCGCAT

is {0,1,1,2} which are the smallest hash values of its k-mers of size 2. These sketches are then compared to estimate the fraction of shared k-mers (Jaccard index) of the corresponding sequences. It is worth noticing that a hash value is a binary number. In a real genomic setting a useful size of k-mers ranges from 14 to 21, and the size of the sketches would be around 1000.[8]

By reducing the size of the sequences, even hundreds of times, and comparing them in an alignment-free way, this method reduces significantly the time of estimation of the similarity of sequences.

Transcriptomics:

Transcriptomics has paved the way for a comprehensive understanding of how genes are expressed and interconnected. Over the last three decades, methodological breakthroughs have repeatedly revolutionized transcriptome profiling and redefined what is possible to investigate. Integration of transcriptomic data with other omics is giving an increasingly integrated view of cellular complexities facilitating holistic approaches to biomedical research (Lowe *et al.*, 2017b).

Normalization of transcriptomic data is an essential preprocessing step aimed at correcting unwanted biological effects and technical noises prior to any downstream analysis. Normalization methods shall be chosen according to the undertaken technology and can be platform-specific. While there is no consensus on the best normalization methods across different transcriptomic technologies, several efforts have been taken to develop additional robust and effective normalization techniques and to systematically assess their performance on individual data sets.

DNA sequence analysis:

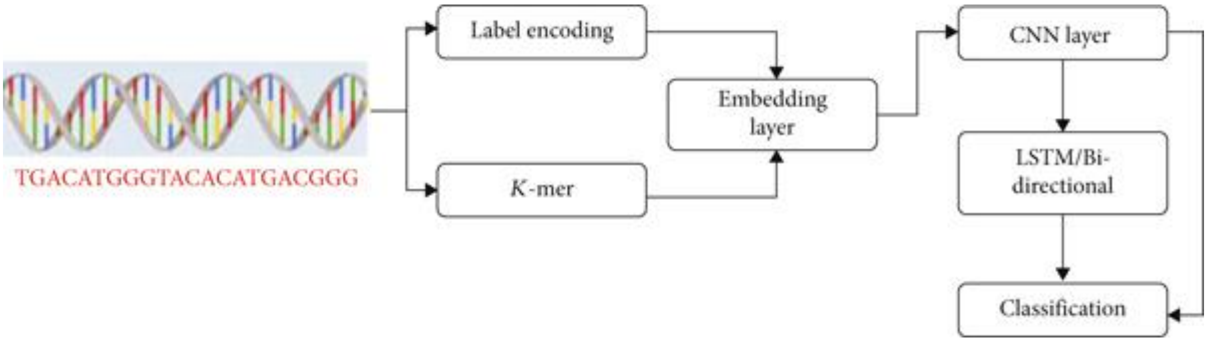
Sequencing is the operation of determining the precise order of nucleotides of a given DNA molecule. It is used to determine the order of the four bases *adenine (A), guanine (G), cytosine (C)* and *thymine (T),* in a strand of DNA.

DNA sequencing is used to determine the sequence of individual genes, full chromosomes or entire genomes of an organism. DNA sequencing has also become the most efficient way to sequence RNA or proteins.

Classification Models

In this work, three different classification models CNN, CNN-LSTM, and CNN-Bidirectional LSTM are used for DNA sequence classification.

The label encoding and -mer techniques are used to encrypt the DNA sequence, which preserves the position information of each nucleotide in the sequence. The embedding layers is used to embed the data from the above two techniques. The CNN layer is used as the feature extraction stage, and it is given as the input for LSTM and bidirectional LSTM for classification. The workflow for the proposed work is shown in Figure



Intron :

A segment of DNA or RNA that does not code for proteins is known as an intron. Introns interrupt the sequence of genes. Introns act as hot spots for recombination.

An intron is a region of a gene that does not code for amino acids that make up the protein encoded by that gene and does not stay in the final mature mRNA molecule following transcription. Exons and introns make up the majority of protein-coding genes in the human genome.

From this perspective, introns are extremely important. In the age of new exon mixtures, they serves as recombination hotspot.

Overall, they are in our characteristics because they have been used to build new quailities more quickly during development.

Functions of introns:

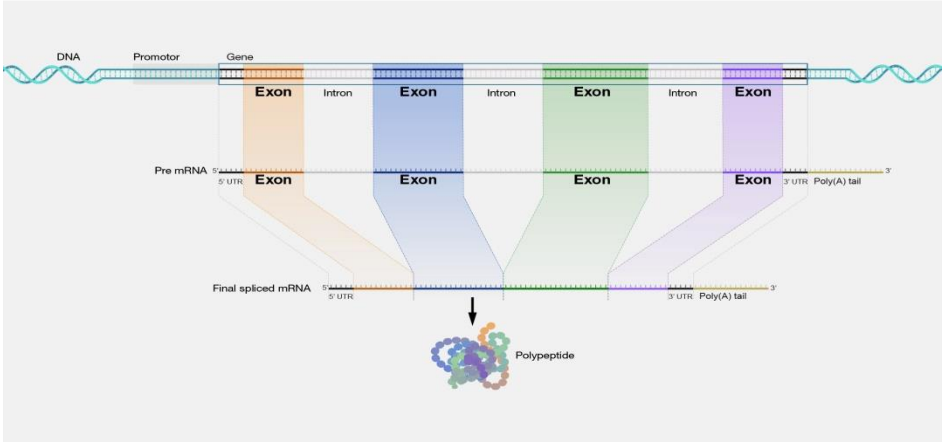
It helps in the regulation of transcription where it protects mRNA that leads to protein synthesis.

They control some genes that are involved in the transcription.

It helps in gene expression and gene regulation that is it helps carry the information through generations, especially in humans, which have more introns than exons.

Exons :

Exons are that part of the RNA that code for proteins. Now, RNA, when it first gets transcribed, is a very, very long piece of RNA molecule. And really, the important parts of that RNA are the exons. There are large, large chunks of RNA that get excised out. Now, it's important to remember that because I use the term excised doesn't mean that exons go away. The exons are what stay in the mature mRNA and eventually code for amino acids. Many times, including medical students like my wife, will forget whether it's the exons that code for the amino acids or the introns that code for the amino acids. Let me set the record straight that it's the exons that code for the amino acids, because sometimes people try to remember that exons get excised, but that's not true. It's that introns interfere. So you always have to remember that introns interfere, and the introns get excised out of the RNA to leave a string of exons together that will eventually code for the amino acids.



Difference between Introns and Exons :

Introns	Exons
Found in Eukaryotes only	Found in both prokaryotes and eukaryotes
Non-coding areas of the DNA	Coding areas of the DNA
Introns are the non-coding part of hnRNA, which are removed before translation by RNA splicing to form mRNA	Exons are the nucleotide sequence in mRNA, which codes for proteins
The sequence of the introns frequently changes over time. In other words, they are less conserved	Exons are highly conserved
DNA bases found in between exons	DNA bases that are translated into proteins
Introns are removed in the nucleus before the mRNA moves to the cytoplasm	Mature mRNA contains exons and moves to the cytoplasm from the nucleus

Microarray :

Microarray technology is a general laboratory approach that involves binding an array of thousands to millions of known nucleic acid fragments to a solid surface, referred to as a “chip.” The chip is then bathed with DNA or RNA isolated from a study sample (such as cells or tissue). Complementary base pairing between the sample and the chip-immobilized fragments produces light through fluorescence that can be detected using a specialized machine. Microarray technology can be used for a variety of purposes in research and clinical studies, such as measuring gene expression and detecting specific DNA sequences (e.g., single-nucleotide polymorphisms, or SNPs).

Microarray technology. Microarrays were revolutionary. They really allow genomic analysis without sequencing, which tremendously reduced the cost of doing large studies across a wide area of biology and biomedicine. Two things that you were able to do there. On the one hand, you're able to look at gene expression or the amount of gene product, RNA, from any given gene that you found in a cell. And the second thing you were able to look at easily was single nucleotide polymorphisms, or SNPs, which were useful for genome-wide association studies, or GWASs. Both of those approaches were used across all of the major human diseases, a large number of less common human diseases, and to also in our model organisms and in other organisms in this world of ours.

RNA Sequencing :

RNA sequencing (RNA-Seq) is revolutionizing the study of the transcriptome. A highly sensitive and accurate tool for measuring expression across the transcriptome, it is providing scientists with visibility into previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions, and across a wide range of other study designs.

RNA-Seq allows researchers to detect both known and novel features in a single assay, enabling the identification of transcript isoforms, gene fusions, single nucleotide variants, and other features without the limitation of prior knowledge.

Benefits of RNA Sequencing :

RNA-Seq with next-generation sequencing (NGS) is increasingly the method of choice for scientists studying the transcriptome.

- Covers an extremely broad dynamic range
- Provides sensitive, accurate measurement of gene expression
- Captures both known and novel features; does not require predesigned probes
- Generates both qualitative and quantitative data
- Reveals the full transcriptome, not just a few selected transcripts
- Can be applied to any species, even if a reference sequence is not available

Genome Annotation :

Genome annotation is the process of identifying functional elements along the sequence of a genome, thus giving meaning to it. It is necessary because the sequencing of DNA produces sequences of unknown function. In the last three decades, genome annotation has evolved from the computational annotation of long protein-coding genes on single genomes (one per species), and the experimental annotation of short regulatory elements on a small number of them, into the population annotation of sole nucleotides on thousands of individual genomes (many per species). This increased resolution and inclusiveness of genome annotations (from genotypes to phenotypes) is leading to precise insights into the biology of species, populations and individuals alike.

Gene prediction :

- In computational biology, gene prediction or gene finding refers to the identification of the genomic DNA regions that encode genes.
- This includes both protein-coding and RNA genes, but may also include the prediction of regulatory regions and other functional elements. Once a species' genome has been sequenced, gene discovery is one of the first and most crucial stages in comprehending its genome.
- Initially, "gene finding" was founded on painstaking experiments conducted on living cells and organisms.
- Statistical analysis of the rates of homologous recombination of several different genes could determine their order on a particular chromosome, and information from many such experiments could be combined to produce a genetic map indicating the approximate relative location of known genes. With a complete genome sequence and potent computational resources at the disposal of the scientific community, gene discovery is now primarily a computational problem.
- Differentiate between determining the function of a gene or its progeny and determining the functionality of a sequence. Although the frontiers of bioinformatics research are making it possible to predict the function of a gene based on its sequence alone, predicting the function of a gene and confirming that the gene prediction is accurate still requires in vivo experimentation through gene knockout and other assays.
- Gene prediction is one of the most important stages in genome annotation, following sequence assembly, non-coding region filtering, and repeat masking.

Gene Prediction Methods :

The process of identifying the locations and boundaries of genes within a genome is known as gene prediction. Understanding the genetic information contained in an organism's DNA is a crucial step. Among the many methods and algorithms used for gene prediction are the following:

1. **Ab initio prediction:** These methods predict genes based on the properties of DNA sequences using computational algorithms. They analyse coding potential, sequence motifs, splice sites, and start/stop codons, among other characteristics. The ab initio gene prediction algorithms GeneMark, Fgenesh, and Glimmer are examples.
2. **Homology-based prediction:** These techniques rely on comparing the DNA sequence to known sequences from organisms with similar characteristics. If a sequence has a high degree of similarity to a known gene, it is likely to be a gene. Homology-based methods seek for similarities using tools such as BLAST (Basic Local Alignment seek Tool). Evolutionarily conserved genes are assumed to have comparable sequences in closely related species.
3. **EST-based prediction:** ESTs (Expressed Sequence Tags) are brief sequences derived from cDNA libraries that represent segments of expressed genes. EST-based gene prediction entails aligning ESTs to genomic sequences and identifying overlapping regions. This method is especially beneficial when working with organisms for which genomic data is scarce.
4. **Transcriptome-based prediction:** These methods identify gene regions based on RNA sequencing (RNA-seq) data. RNA-sequencing provides information regarding the transcripts present in a particular tissue or condition. By aligning RNA-seq reads to the genome, gene locations and alternative splicing patterns can be inferred.
5. **Comparative genomics:** This method entails comparing the genomes of various species to identify conserved regions that are likely to correspond to genes. By aligning genomes, researchers can identify regions that may represent functional genes that are shared between species.
6. **Machine learning and deep learning:** Using advanced computational techniques, machine learning and deep learning train models on large datasets of known genomes. By recognising patterns within these datasets, the models are able to predict new genomic sequences. Machine learning and deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in gene prediction tasks.

Differential gene expression analysis :

Differential expression analysis means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

Methods for differential expression analysis :

There are different methods for differential expression analysis such as edgeR is based on negative binomial (NB) distributions or baySeq and EBSeq which are Bayesian approaches based on a negative binomial model. It is important to consider the experimental design when choosing an analysis method. While some of the differential expression tools can only perform pair-wise comparison, others such as edgeR, limma-voom, DESeq and maSigPro can perform multiple comparisons.

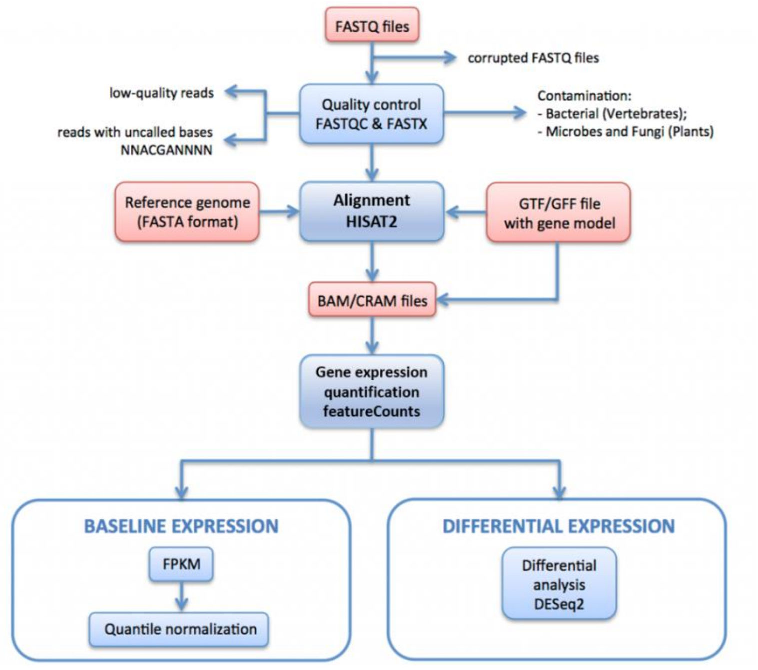

FASTQ files
corrupted FASTQ files
low-quality reads
reads with uncalled bases NNACGANNNN
Quality control FASTQC & FASTX
Contamination:
- Bacterial (Vertebrates);
- Microbes and Fungi (Plants)
Reference genome (FASTA format)
Alignment HISAT2
GTF/GFF file with gene model
BAM/CRAM files
Gene expression quantification featureCounts
BASELINE EXPRESSION
FPKM
Quantile normalization
DIFFERENTIAL EXPRESSION
Differential analysis DESeq2

Figure - RNA-seq processing pipeline used to generate gene expression data in Expression Atlas.

In this pipeline raw reads (FASTQ files) undergo quality assessment and filtering. The quality-filtered reads are aligned to the reference genome via HISAT2. The mapped reads are summarised and aggregated over genes via HTSeq. For baseline expression, the FPKMs are calculated from the raw counts by iRAP. These are averaged for each set of technical replicates, and then quantile normalised within each set of biological replicates using limma.

Finally, they are averaged for all biological replicates (if any). For differential expression, genes expressed differentially between the test and the reference groups of each pairwise contrast are identified using DESeq2.

NCBI and its role :

NCBI stands for the National Center for Biotechnology Information. It is a part of the United States National Library of Medicine (NLM), which is a branch of the National Institutes of Health (NIH). NCBI is a renowned resource in the field of bioinformatics and molecular biology, providing a wide range of tools and databases that facilitate research and information retrieval in these areas.

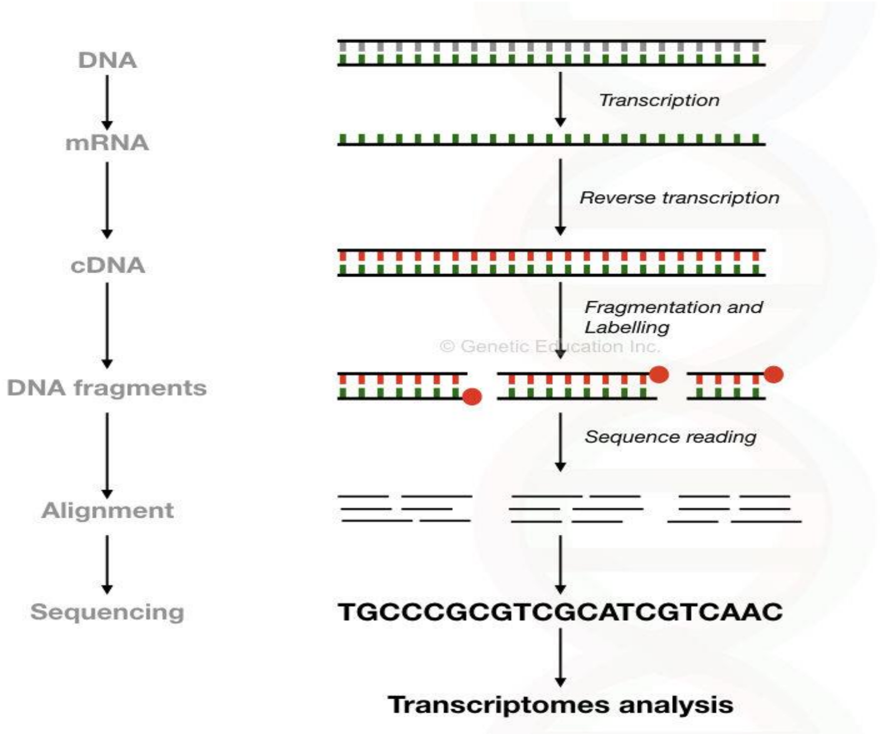
The National Center for Biotechnology Information (NCBI) plays a pivotal role in advancing scientific research and knowledge in the fields of biotechnology, molecular biology, genetics, and related areas. Its primary role revolves around providing access to vast amounts of biological and biomedical information through various databases, tools, and resources. Here are some key roles of NCBI:

- Data Repositories
- Research Facilitation
- Genomic Information
- Bioinformatics Tools
- Data Analysis
- Biomedical Research.

RNA Sequencing Process:

Explanation: RNA sequencing (RNA –Seq.) is a powerful technique used to study gene expression and transcriptomics. It provides valuable insights into the types and quantities of RNA molecules present in a biological sample at a given time. The RNA Seq. process involves several steps, from sample preparation to data analysis.

RNA sequencing (RNA-Seq) uses the capabilities of high-throughput sequencing methods to provide insight into the transcriptome of a cell. Compared to previous Sanger sequencing- and microarray-based methods, RNA-Seq provides far higher coverage and greater resolution of the dynamic nature of the transcriptome. Beyond quantifying gene expression, the data generated by RNA-Seq facilitate the discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele-specific expression. Recent advances in the RNA-Seq workflow, from sample preparation to library construction to data analysis.



Detail genome annotation :

Explanation : Genome annotation is the process of identifying functional elements along the sequence of a genome, thus giving meaning to it. It is necessary because the sequencing of DNA produces sequences of unknown function

It consists of three main steps:

- Identifying portions of the genome that do not code for proteins
- identifying elements on the genome, a process called gene prediction, and
- Attaching biological information to these elements.

Automatic annotation tools try to perform all of this by computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline (process). The basic level of annotation is using BLAST for finding similarities, and then annotating genomes based on that. However, nowadays more and more additional information is added to the annotation platform. The additional information allows manual annotators to deconvolute discrepancies between genes that are given the same annotation.