# Chapter 1

# Introduction

A biomarker is a quantitatively measurable indicator or trait that can identify the presence, progression, and severity of a disease. It can be used to assess the biological effects of particular treatments on a disease. Genes, proteins, metabolites, and nucleic acids are a few examples of biomarkers. Other examples of biomarkers include physiological changes such as variations in heart rate, brain activity, or blood pressure. Biomarkers can play a significant role in guiding new therapeutic approaches and diagnostic methods by providing crucial information about the underlying biological mechanisms of a health condition or disease. Biomarkers play a significant role in critical diseases.

- Biomarkers make early detection of specific diseases a possibility, as opposed to traditional diagnostic methods.
- Biomarkers help in the prediction of the likelihood of recurrence and severity of a disease.
- The progression of a disease and the effectiveness of a treatment for that disease can be monitored and assessed using biomarkers.
- Biomarkers can be used to assess the efficacy of new therapeutics and treatments and find new targets. Moreover, biomarkers can help identify patients who are most likely to benefit from a specific course of treatment.

Finding biomarkers for critical diseases requires an understanding of the Central Dogma of molecular biology.

## 1.1 Central Dogma

The central dogma [109] of molecular biology defines the only direction of genetic information flow within a biological system. It asserts that Deoxyribonucleic Acid (DNA) [758] is transcribed into Ribonucleic Acid (RNA) [445, 16] which is then further translated into protein as shown in Fig. 1.1.

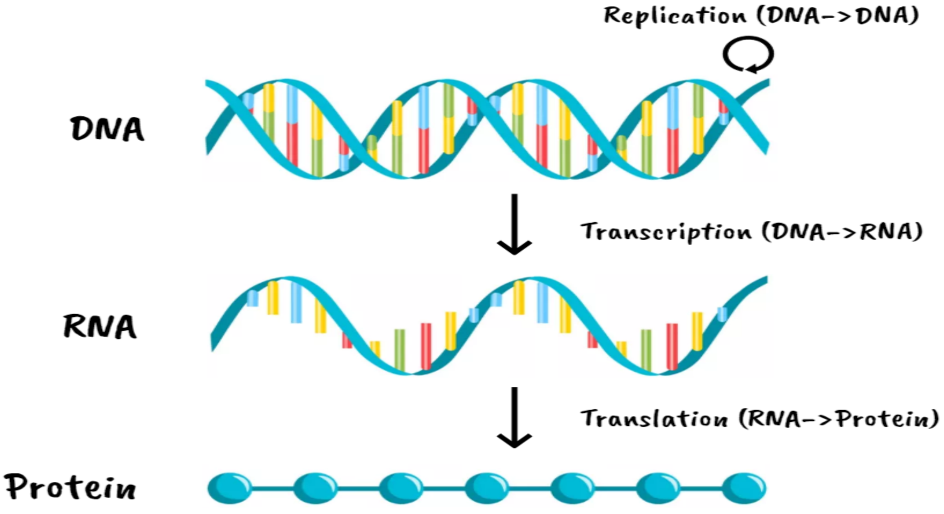

*Fig. 1.1:* Steps involved in the Central Dogma [109] of molecular biology

The amino acid sequence ultimately determines the structure as well as function of a protein thus affecting the characteristics and traits exhibited by an organism. Central dogma is essential to the comprehension of fundamental processes of life and helps in understanding gene regulations, genetic disorders, and evolutionary links. The following describes the fundamental steps included in central dogma as shown in Figure. 1.1.

(a) *Transcription*: During the initiation phase of transcription, the enzyme RNA polymerase latches on or binds to a specific section of the DNA strand known as the promoter, and separates the double strands of the DNA into two single strands. One of the DNA strands acts as the template while the enzyme builds an RNA molecule out of complementary nucleotides until it encounters a termination sequence. Through a process known as RNA splicing, the RNA is then converted to mRNA.

(b) *Translation*: The messenger RNA (mRNA) transports the genetic information from the DNA to the ribosome. This genetic information in the form of mRNA codons is read by the ribosome and utilized to assemble a specific sequence of amino acids that constitute a protein. Particular amino acids are transported to the ribosome and matched with mRNA codons with the help of transfer RNA (tRNA). When a stop codon is encountered, the translation process stops resulting in the formation of a long polypeptide chain of amino acids. As the three dimensional structure of a protein is crucial to its function, the polypeptide chain folds into a biologically active protein.

The fundamental premise of the central dogma [109] plays an important role in biomarker identification. Central dogma provides a foundation to comprehend expression of genes and indication of a disease state as variations in gene expression. Transcription and translation stages of the central dogma are significant for gene expression. In other works the central dogma provides a framework for the comprehension of the fundamental ideas underlying gene expression.

## 1.2 Gene Expression

The DNA within an organism consists of specific segments contain instructions for performing regulatory functions or for synthesis of proteins. Specific function or trait is a result of encoding of particular proteins by a gene. In the translation phase of the central dogma, mRNAs that act as templates for protein synthesis are transcribed from the genes. Gene expression is the process of creation of a function of a gene product such as an RNA molecule or protein through the utilization of the genetic information encoded in a gene. Gene expression plays a crucial role in living organisms by reacting to environmental changes or enabling cells to perform certain tasks. Gene expression often occurs either through RNA molecules that code for protein or non-coding RNA molecules that serve other functions [1]. Under varying conditions and cell types, gene expression tends to undergo substantial changes and thus is regulated carefully. Gene expression regulation refers to the intricate process that is responsible for controlling trancription and translation (Central Dogma, Section 1.1). Through observation of phenotypes associated to a gene or measurement of functional activity of a gene, gene expression can be analyzed.

### 1.2.1 Gene Expression Data

The information about the activity levels of a gene in a cell or tissue of an organism is referred to as gene expression data. In a particular biological context, gene expression data can provide insight into which genes are active and which genes are inactive. Thus, gene expression data facilitates understanding regulation and function of genes as well as their contribution to biological traits and processes.

Microarrays, bulk RNA sequencing (RNA-Seq), or single cell RNA sequencing (scRNA-Seq) are examples of various experimental techniques to collect gene expression data. The amount of RNA obtained from various genes under different conditions

[1] https://www.genome.gov/genetics-glossary/Gene-Expression

or samples can be measured using these techniques. Depending on the precise research objectives and experimental strategy, various types of gene expression data can be produced and analyzed. Typical forms of gene expression data include the following.

(a) *Microarray [610]*: The expression levels of thousands of genes within a biological sample can be measured simultaneously using microarray technology. A microarray chip with probes corresponding to thousands of genes of interest is attached with tiny fragments of DNA or RNA. The labeled DNA or RNA sample binds to the corresponding probes enabling the measurement of the abundance of specific RNA molecules (mRNA) in that sample. The information generated unravels the relative gene expression levels in various samples.

(b) *Bulk RNA Sequencing (RNA-Seq) [512, 753]*: Bulk RNA-Seq technology starts with conversion of the RNA molecules into fragments of complementary DNA (cDNA) followed by the use of high-throughput sequencing machines that results in the generation of millions of short sequence reads. Each read corresponds to a cDNA fragment. To determine the origin and abundance of the RNA molecules, the generated sequence data is processed and aligned to a reference genome or transcriptome. Subsequent analysis facilitates the determination of the absolute levels of each gene. as well as discovery of new splice variants and new transcripts.

(c) *Single-cell RNA Sequencing (scRNA-Seq) [672]*: scRNA-Seq technology provides insight into the diversity and heterogeneity of gene expression patterns in different cell types by enabling cellular level analysis of gene expression profiles. Isolation of the cells from a sample is followed by extraction and processing of the contents of each cell. cDNA fragments extracted from the RNA molecules are further amplified and sequenced using high-throughput sequencing. This results in the generation of millions of short sequence reads which are further analyzed for determination of gene expression profiles corresponding to each individual cell.

Unlike microarrays that have constrained dynamic range, bulk RNA-Seq has a wider dynamic range facilitating reliable identification of both highly expressed and lowly expressed genes. While microarray is strictly restricted to the probes of interest contained in the array, bulk RNA-Seq is able to cover both known and new transcripts. Owing to its low sensitivity microarray may overlook transcripts and isoforms with low abundance. scRNA-Seq requires specialized technology such as microfluidic devices and other added processes for separation of each individual cell, and tools that enable analysis regardless of its unpredictability and inherent noise. scRNA-Seq facilitates recognition of uncommon cell types, analysis of single-cell level dynamic changes, and identification of cell-to-cell heterogeneity. When compared to bulk RNA-Seq, scRNA-Seq has requires specialized knowledge and tools, stricter technical requirements, and has higher costs.

## 1.3 Gene Expression Data Analysis

The process of understanding the data present in a gene expression dataset is known as gene expression data analysis. Through gene expression data analysis, expression levels and patterns exhibited by genes across samples are analyzed and interpreted so as to learn more about biological processes. Analysis of gene expression data can be utilized in clinical settings to identify diagnostic or prognostic biomarkers and for therapy decision-making. Depending on the precise objective and the type of data at hand, there are numerous gene expression data analysis approaches. The following are some of the most prevalent gene expression data analysis approaches.

### 1.3.1 Gene Co-expression Analysis: Biclustering Approach

Through biclustering [96, 481, 553], a computational approach for gene expression data analysis, groups of genes are identified that are co-expressed across conditions or samples. While traditional clustering methods cluster genes based on their expression profiles across all conditions, biclustering takes a two-pronged approach and considers both genes and conditions simultaneously to find groups of genes that exhibit similar expression patterns. Biclustering aids in the prediction of gene functions, identification of co-regulated genes corresponding to specific biological processes, and identification of potential regulatory networks.

### 1.3.2 Differential Co-expression Analysis (DCA)

Through differential co-expression analysis (DCA) [853, 327], the variations in co-expression patterns exhibited by genes between two or more conditions or groups are analyzed. Co-expression, which can imply functional links between the genes, is the propensity for genes to be expressed in concert with one another [647]. As a response to varying conditions such as healthy vs disease, interactions among genes may change. These changes can be highlighted via DCA. The general pipeline for DCA includes: a) pre-processing of gene expression data through normalization and batch effect removal among others, b) construction of a co-expression network (CEN), c) employment of various statistical methods to compare the CENs under different conditions, and d) biological validation of the results through functional enrichment analysis and regulatory behavior analysis.

### 1.3.3 Differential Expression Analysis

Through differential expression analysis (DEA) [193], the levels of gene expression of individual genes under various conditions or groups are analyzed. In other words, DEA involves contrasting the levels of expression of each individual gene in two or more varying biological samples. Genes that exhibit differential expression between various conditions, such as disease and healthy tissue, or between various phases of development, can be found using DEA. This can shed light on the underlying biological mechanisms that contribute to these diseases. Statistical methods are often employed to find differentially expressed genes (DEGs) in DEA. DEA methods that often cater specifically to bulk RNA-Seq data or microarray data are frequently used for analysis.

### 1.3.4 Discussion

Machine learning techniques are often capable of handling large and complex datasets and can recognize patterns and relationships among the data that might be overlooked by conventional statistical methods thus leading to the rising popularity of these techniques in gene expression data analysis. In gene expression data analysis, dimensionality reduction, classification, regression, and clustering are few examples of machine learning techniques frequently employed. Machine learning techniques are capable of handling high-dimensional data and can provide novel biological insight as well as uncover intricate relationships and interactions between gene expressions. Predictive models for diagnosis, prognosis, and treatment of critical diseases can be constructed using machine learning techniques. Machine learning techniques are however vulnerable to overfitting, sensitive to missing values and biases, may be computationally intensive and may require prior knowledge. Weighing the possible advantages and disadvantages of each machine learning technique taken into consideration for implementation for analysis and interpretation of gene expression data.

Biclustering, DEA, and DCA are the most widely used methods for gene expression data analysis in genomics and transcriptomics research and each of these methods have unique advantages and disadvantages. While biclustering identifies groups of genes and samples that exhibit similar expression patterns, DEA primarily focuses on identifying individual genes that exhibit varying patterns under two or more conditions. By identifying biologically significant subsets of genes and samples, biclustering has the additional benefit of dimensionality reduction leading to more efficient and effective subsequent analysis. While biclustering can unravel co-regulated gene sets and detect sample subtypes, DEA is capable of detecting genes that are most closely related to a given condition or phenotype. Biclustering is robust to missing values and noise quite prevalent in genomic and transcriptomic datasets. Implementation of DEA is simpler and thus more widely used.

Although the primary focus of biclustering and DCA are co-expression patterns exhibited by genes, the final results and interpretations of the two analysis are quite different. Unlike biclustering, DCA helps in the identification of gene-gene pairs that exhibit distinct co-expression under two or more varying conditions. DCA sheds light on the changes in regulatory linkages and can identify key regulators in biological processes. Biclustering aids the discovery of shared biological processes and regulatory mechanisms by identifying sets of co-regulated genes that exhibit similar expression patterns. While biclustering can aid in the discovery of patient clusters or sample subtypes , DCA aids in the identification of genes that are crucial in the regulation of biological processes. Biclustering is more susceptible to false positives while interpretation and comprehension of revealed regulatory linkages by DCA is quite complex.

DCA and DEA different approaches and aim to accomplish different things. DEA seeks to find genes that exhibit noticeably different expression levels under varying conditions thus assisting in the identification of genes that are closely related to particular biological processes. DCA primarily focuses on finding gene pairings that exhibit varying levels of co-expression under varying conditions and seeks to discover regulatory links between genes that are significantly altered under different conditions. As DEA primarily focuses on individual genes, it fails to recognize significantly co-regulated gene sets as having different gene expression levels. DCA sheds light on changes in regulatory interactions that may be overlooked by DEA.

## 1.4 Motivation

This research is focused around the discovery of potential biomarkers for critical diseases, such as Esophageal Squamous Cell Carcinoma (ESCC). As a consequence, we must comprehend and investigate:

- the similar or dissimilar expression patterns exhibited by individual genes under different conditions, specifically normal and pathological states.
- the co-regulation of gene clusters or subsets that exhibit comparable patterns in these two states.
- the simultaneous activation as well as dynamic behavioral changes among associated genes under these conditions.

It is used to investigate three gene expression data analysis strategies that potentially satisfy all these three objectives.

We discover gene biclusters that are functionally connected, participate in the same pathways, or create modules that are possibly co-regulated by a limited group of transcription factors using bicluster analysis. Biclustering gene expression data analysis results in the discovery of local patterns in which a subset of genes exhibit similar expression levels across a subset of situations. DEA employs statistical techniques to identify quantifiable differences in gene expression levels between experimental groups. The correlation of expression patterns across several samples is used to determine whether CENs have the capacity to successfully find gene connections. DCA identifies genes that co-express in various ways under different conditions. By estimating the expression correlation change of each gene pair between circumstances, DCA investigates differences in gene interconnections.

Because this research is primarily concerned with gene expression data, it is aimed to investigate the applications of improved machine learning techniques in gene expression data analysis to identify potential biomarkers for ESCC.

## 1.5 Objectives

The main objective of this research work is to explore and develop effective machine learning technique(s) to analyze gene expression data of various types towards the identification of crucial genes as potential biomarkers for critical diseases like ESCC. It is aimed to achieve this objective through the following.

1. Explore the state-of-the-art biclustering techniques and analyze their pros and cons using gene expression data towards development of an effective biomarker identification framework for critical diseases.
2. Study the effectiveness of state-of-the-art DEA analysis methods for both microarray and RNA-Seq data and develop a framework for both microarray and RNA-Seq methods for unbiased and interesting biomarker identification.
3. Study the effectiveness of existing methods for biomarker identification based on topological analysis of CEN modules and DCA using microarray and RNA-Seq data.
4. The rapid development of single-cell RNA Sequencing (scRNA-Seq) data provides enormous scope to gain insight into the interplay between intrinsic cellular processes as well as transcriptional and behavioral changes in gene-gene interactions across varying conditions. It is aimed to explore the effectiveness of machine learning enabled biomarker identification framework in handling such high dimensional scRNA-Seq data for identification of crucial gene(s) for critical disease(s).
5. Identify genes that are most likely to be potential biomarkers for a critical disease.

## 1.6 Contributions

Through this time-based research, a number of contributions have been made to support diagnosis or prognosis of critical diseases such as ESCC using gene biomarkers. Next, the contributions have been reported, in brief.

### 1.6.1 Biclustering Analysis

We develop BicGenesis, a biomarker identification framework that incorporates existing biclustering algorithms from multiple approaches. Interesting biclusters were extracted using appropriate selection criteria for subsequent downstream analysis towards identification of crucial genes for critical diseases. Additionally, we employ DCA to examine the dynamic changes in gene connections in a bicluster under normal and disease conditions. To perform DCA, a robust pipeline is used which includes: (i) the construction of a CEN, (ii) the designation of each relevant bicluster as a CEN module, (iii) the extraction of connections related to each relevant bicluster, and (iv) the identification of relevant modules, which we term as 'Modules of Interest' (MoI). We use a well-known hub-gene finding approach to identify hub-genes in a MoI, with the premise that hub-genes in a module provide relevant information. All hub-genes of MoIs are considered as candidates for potential biomarkers, which we term as ‘Biomarker Candidate Gene (BCG)’. We validate these candidate genes as potential biomarkers for the relevant critical disease. Our biomarker validation criteria, include (i) biological validation and (ii) literature trace. We use functional enrichment analysis and regulatory network behavior analysis to biologically validate each candidate gene. Furthermore, we review established literature on wet-lab results related to the concerned gene(s) to substantiate our claim of finding crucial genes(s) for a given disease.

### 1.6.2 Differential Expression Analysis (DEA)

We investigate the efficacy of existing state-of-the-art DEA methods for both microarray and RNA-Seq data and developed a framework for biomarker identification that integrates selected approaches from both microarray and RNA-Seq methodologies. Individually, the approaches used to identify a set of Differentially Expressed Genes (DEGs). We propose a consensus function for seamless integration of various methods into the framework and determination of the suitable set of DEGs. We find the DEGs that are substantially functionally enriched from this list of DEGs. We term these DEGs as Top Enriched DEGs (TEDs). We consider TEDs as candidates for potential biomarkers, i.e, BCGs. Furthermore, we use DCA to analyze the dynamic changes in the interactions among the list of DEGs under normal and disease conditions. The proposed pipeline for DCA includes: (i) the construction of a CEN, (ii) the extraction of connections related to the list of DEGs, (iii) module extraction, and (iv) the identification of relevant modules or MoIs. To identify hub-genes in a MoI, a well-known hub-gene finding approach is used, with the premise that hub-genes in a module provide relevant information. These DEGs, which are hub-genes in significant CEN modules, are candidates for potential biomarkers, i.e., BCGs. All genes identified as candidates for potential biomarker by the proposed framework are evaluated both statistically and biologically using our biomarker criteria.

### 1.6.3 Differential Co-expression Analysis (DCA)

We investigate the efficacy of existing approaches for biomarker identification based on hub-gene discovery in CEN modules and DCA. In a network, a hub-gene is one that is highly coupled to other genes in the network. Hub-genes are essential for the formation and operation of biological networks and are known to influence network dynamics. The discovery of potential biomarkers for disease diagnosis, prognosis, and monitoring can be aided by the identification of hub-genes in disease networks. We develop a centrality-based hub-gene finding method, keeping in mind the relevance of a hub-gene in a CEN. This approach incorporates seven well known centrality measures and a consensus function to find an unbiased subset of hub-genes in a CEN module. It outperforms or performs on par with a handful of its well-known counterparts in majority of cases. Furthermore, the aforementioned hub-gene finding method is incorporated into our DCA framework, referred to as Centrality Based DCE Method (CBDCEM). CBDCEM is capable of constructing CENs from microarray and bulk RNA-Seq data and can extract relevant modules or MoIs from these CENs. Following that, the proposed hub-gene finding approach discovers hub-genes from these MoIs. These hub-genes are further analyzed as candidates for potential biomarkers,i.e., BCGs All genes identified as candidates for potential biomarker by CBDCEM are evaluated further using our biomarker criteria.

### 1.6.4 DCA on scRNA-Seq data

The scRNA-Seq technology allows researchers to learn about the relationships between intrinsic cellular processes as well as transcriptional and behavioral changes in gene-gene interactions under various conditions. DCA on scRNA-Seq data can reveal genes that are specifically expressed in one cell type but not in others by comparing the co-expression patterns of genes across various cell types. Genes that are co-expressed under varying conditions can be identified using DCA. These genes have the potential to be used as biomarkers for critical diseases. However, the huge bulk and sparsity of the cellular-level scRNA-seq data provide considerable difficulty in the analysis. We investigate the benefits and limitations of scRNA-Seq data, as well as the difficulties encountered while performing DCA on it. We present a framework, scDiffCoAM, that facilitates DCA on scRNA-Seq data while maintaining as close to the standard DCA process as possible. The three major steps in the framework are. (i) CEN construction, (ii) module extraction, and (iii) DCA. Because of the inherent sparsity of scRNA-Seq data, CEN construction and module selection are accomplished by identifying blocks of well-connected genes and extracting connections relevant to these blocks, which are now designated as modules. We propose a simple but effective criteria to establish MoIs. We employ the hub-gene finding method proposed in CBDCEM to find hub-genes from the identified MoIs. The hub-gene finding algorithm incorporates appropriate measures to find hub-genes. These hub-genes are further analyzed as candidates for potential biomarkers, i.e, BCGs. All genes identified as candidates for potential biomarkers by scDiffCoAM are evaluated further using our biomarker criteria.

### 1.6.5 Potential Biomarker Identification and Ranking

All BCGs identified by our proposed frameworks, BicGenesis, Integrative DEA, CBDCEM, and scDiffCoAM using a set of methods corresponding to the method, are further validated as potential biomarkers. We introduce a biomarker criteria which identifies potential biomarkers for a disease through literature evidence and biological validation. Biological validation involves functional enrichment analysis and analysis of gene regulatory behavior. We also use a biomarker ranking scheme that scores each potential biomarker based individual biological and literature evidence presented that associate them with the disease as well as the number of frameworks that identify that gene as a potential biomarker. All highly ranked genes are the most likely potential biomarkers for the disease.

## 1.7 Organization Of The Thesis

The rest of the thesis is organized as follows. Chapter 2 presents background of various problems that are addressed in this research. Gene expression analysis tools, statistical methods/ measures, and biological knowledge bases employed in the process of this research are discussed in detail in the chapter. Various statistical measures, analysis, tools, our biomarker criteria, the datasets used, and the pre-processing pipelines used for all data types across all the works are also elaborated in this chapter. Chapter 3 presents an overview of existing literature on biclustering analysis. The chapter also introduces and discusses our own BicGenesis, a method to identify ESCC biomarkers using the biclustering approach. Chapter 4 presents a discussion on existing microarray and bulk RNA-Seq DEA methods. Further, it introduces and establishes a DEA framework that employs integrative approach and a consensus function to identify and analyze DEGs with the aim to establish them as potential biomarkers. Chapter 5 presents an effective centrality based differential co-expression analysis method for critical gene finding. Also, this chapter provides an overview on seven centrality measures and introduces a hub-gene finding method that employs these centrality measures. Further, it also presents a detailed comparison and discussion of the proposed hub-gene finding method against four well known hub-gene finding methods. Chapter 6 reports a detailed discussion of DCA on scRNA-Seq data. It introduces a framework, scDiffCoAM, for DCA on scRNA-Seq data which incorporates the hub-gene finding method introduced by CBDCEM with network specific measures and presents a detailed comparison of the same with four other hub-gene finding methods. We also present a ranking scheme that scores all genes identified as potential biomarkers by all four frameworks. Based on the final rank we identify the likely-hood that a gene is a potential biomarker for a chosen disease. Finally, Chapter 7 presents concluding remarks and future direction of the research.