Chapter 1

Introduction

The past decade has seen an explosive growth in the amount of biological data being stored in databases. However, this data is essentially useless until analysed. Bioinformatics is a cross-disciplinary field that emerged in the 1960s, pioneered by Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others, that is focused on managing and interpreting this data. This field combines aspects of biology, computer science and statistics. Research in bioinformatics focuses on interpreting, processing, analyzing and developing algorithms that can make predictions based on biological data such as microarray/RNA-seq data and draw biologically and clinically meaningful conclusions. Data mining has become a popular solution to this problem, as it uses efficient and reliable computational and mathematical techniques. This has led to improvements in critical disease diagnostics, biomarker identification and medicine discoveries.

1.1 Central Dogma of Molecular Biology

A cell is the basic unit of life. Cells are highly organized and complex structures made up of smaller components, including DNA, proteins, carbohydrates, and lipids. DNA, or deoxyribonucleic acid, is the genetic material that carries the instructions needed for the cell to function properly. It is composed of two strands of nucleotides twisted into a double helix. Each strand is composed of four nucleotides i.e., adenine (A), thymine (T), guanine (G), and cytosine (C) which are the building blocks of DNA as presented in Figure 1-1. DNA is the fundamental building block of genes. Gene is a sequence of nucleotides that contains the instructions for the development of a particular organism. Each gene consists

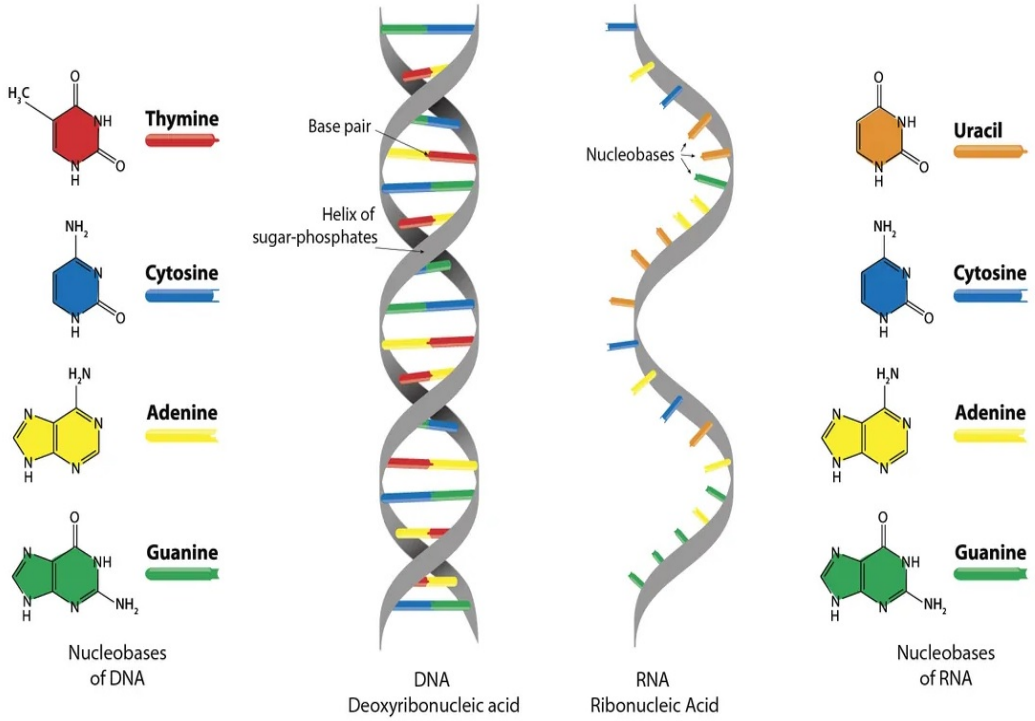


Figure 1-1: Structures of DNA and RNA (Credit: Technology Networks, https://www.technologynetworks.com/genomics/lists/, accessed on 21/07/2023)

of two parts: the coding region, which contains the information for a particular protein, and the non-coding region, which is the regulatory region that controls the expression of the gene. Genes are present in all cells and control the traits of various characteristics in an organism. These traits can be related to either diseases or regular growth. The process of how genes influence characteristics is known as the Central Dogma shown in Figure 1-2. The Central Dogma of Molecular Biology explains how genetic information is passed from one generation to the next, how it is expressed and regulated, and how different proteins are created. It begins with the process of DNA replication, in which the double-stranded DNA is unwound and each single strand acts as a template for a new double-stranded DNA. This new DNA is identical to the original. The next step is transcription, in which the sequence of DNA is copied into a complementary RNA molecule, using the enzyme RNA polymerase. This process is known as transcription. The RNA produced by this process is known as messenger RNA (mRNA). The mRNA is then translated into a protein molecule by the process of translation. During this process, the mRNA molecule is read codon by codon, and each codon is translated into an amino acid. These amino acids are then linked together to form a protein molecule. Finally, the newly formed protein molecule carries out the cellular activity that corresponds to the genetic code. This process is known as gene expression.

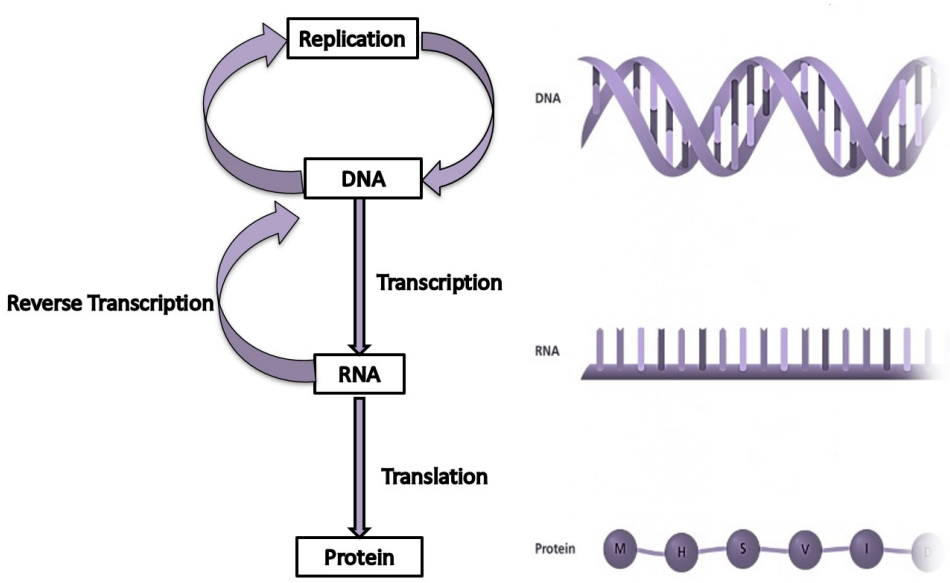


Figure 1-2: Central Dogma

1.2 Gene Expression Data (GED)

Gene expression is the process by which genetic information is used to synthesize proteins and other gene products. Measuring gene expression entails analyzing the amount of gene products (i.e. proteins, mRNA, etc) present in a sample. There are a variety of techniques used to measure gene expression, including DNA microarray or sequencing technologies such as bulk RNA-sequencing (RNA-seq) and Single-cell RNA-seq (scRNA-seq). The data generated from DNA microarrays or sequencing technologies is called as gene expression data. Gene expression data is typically represented as a table as shown in Figure 1-3, with each row being a gene and each column being a sample or cell.

Samples →	S1	S2	S3	S4	...	Sn
G1	0.1	0.7	0.2	0.8	...	0.9
G2	0.4	0.1	0.1	2.8	...	0.4
G3	3.6	0.2	0.7	1.5	...	0.5
G4	1.1	0.1	0.9	0.8	...	1.7
...	...	...	...	...	...	
Gn	0.2	0.9	0.6	0.2	...	1.2

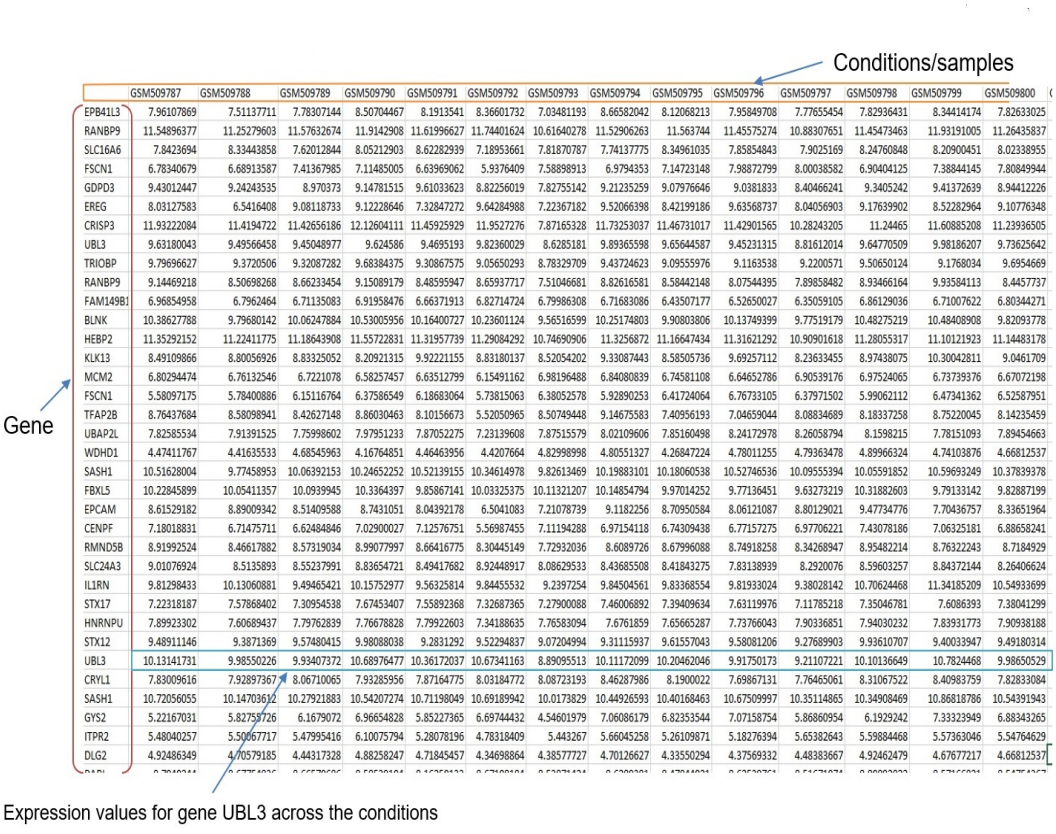
Genes →

Gene Expression Value

Figure 1-3: An example of Gene Expression Dataset

1.2.1 DNA Microarray Data

Microarrays are a type of DNA chip technology measures the expression levels of thousands of genes simultaneously [1]. Microarrays use DNA probes attached to a surface to detect the presence and quantity of specific RNA molecules. Microarrays generate data in the form of two-dimensional grids, or "spots", which represent the expression levels of each gene. Each spot contains information about the expression level of a particular gene. The intensity of the spot is proportional to the amount of gene expression. The spots can be labeled with fluorescent dyes or with radioactive isotopes, depending on the type of microarray used. The spots are arranged in an array, with the rows representing different genes and the columns representing different samples. The data generated by the microarray is usually presented as a heatmap, which highlights the relative expression levels of the different genes. Microarray data (figure 1-4) can be used to compare gene expression levels between different conditions or treatments, and can be used to identify genes involved in particular biological processes, disease, discover potential drug targets, and develop diagnostic tests.



	GSM509787	GSM509788	GSM509789	GSM509790	GSM509791	GSM509792	GSM509793	GSM509794	GSM509795	GSM509796	GSM509797	GSM509798	GSM509799	GSM509800
EPB41L3	7.96107869	7.51137711	7.78307144	8.50704467	8.1913541	8.36601732	7.03481193	8.66582042	8.12068213	7.95849708	7.77655454	7.82936431	8.34414174	7.82633025
RANBP9	11.54896377	11.25279603	11.57632674	11.9142908	11.61996627	11.74401624	10.61640278	11.52906263	11.563744	11.45575274	10.88307651	11.45473463	11.93191005	11.26435837
SLC16A6	7.8423694	8.33443858	7.62012844	8.05212903	8.62282939	7.18953661	7.81870787	7.74137775	8.34961035	7.85854843	7.9025169	8.24760848	8.20900451	8.02338955
FSCN1	6.78340679	6.68913587	7.41367985	7.11485005	6.63969062	5.9376409	7.58898913	6.9794353	7.14723148	7.98872799	8.00038582	6.90404125	7.38844145	7.80849944
GDPD3	9.43012447	9.24243535	8.970373	9.14781515	9.61033623	8.82256019	7.82755142	9.21235259	9.07976646	9.0381833	8.40466241	9.3405242	9.41372639	8.94412226
EREG	8.03127583	6.5416408	9.08118733	9.12228646	7.32847272	9.64284988	7.22367182	9.52066398	8.42199186	9.63568737	8.04056903	9.17639902	8.52282964	9.10776348
CRISP3	11.93222084	11.4194722	11.42656186	12.12604111	11.45925929	11.9527276	7.87165328	11.73253037	11.46731017	11.42901565	10.28243205	11.24465	11.60885208	11.23936505
UBL3	9.63180043	9.49566458	9.45048977	9.624586	9.4695193	9.82360029	8.6285181	9.89365598	9.65644587	9.45231315	8.81612014	9.64770509	9.98186207	9.73625642
TRIOBP	9.79696627	9.3720506	9.32087282	9.68384375	9.30867575	9.05650293	8.78329709	9.43724623	9.09555976	9.1163538	9.2200571	9.50650124	9.1768034	9.6954669
RANBP9	9.14469218	8.50698268	8.66233454	9.15089179	8.48595947	8.65937717	7.51046681	8.82616581	8.58442148	8.07544395	7.89858482	8.93466164	9.93584113	8.4457737
FAM149B1	6.96854958	6.7962464	6.71135083	6.91958476	6.66371913	6.82714724	6.79986308	6.71683086	6.43507177	6.52650027	6.35059105	6.86129036	6.71007622	6.80344271
BLNK	10.38627788	9.79680142	10.06247884	10.53005956	10.16400727	10.23601124	9.56516599	10.25174803	9.90803806	10.13749399	9.77519179	10.48275219	10.48408908	9.82093778
HEBP2	11.35292152	11.22411775	11.18643908	11.55722831	11.31957739	11.29084292	10.74690906	11.3256872	11.16647434	11.31621292	10.90901618	11.28055317	11.10121923	11.14483178
KLK13	8.49109866	8.80056926	8.83325052	8.20921315	9.92221155	8.83180137	8.52054202	9.33087443	8.58505736	9.69257112	8.23633455	8.97438075	10.30042811	9.0461709
MCM2	6.80294474	6.76132546	6.7221078	6.58257457	6.63512799	6.15491162	6.98196488	6.84080839	6.74581108	6.64652786	6.90539176	6.97524065	6.73739376	6.67072198
FSCN1	5.58097175	5.78400886	6.15116764	6.37586549	6.18683064	5.73815063	6.38052578	5.92890253	6.41724064	6.76733105	6.37971502	5.99062112	6.47341362	6.52587951
TFAP2B	8.76437684	8.58098941	8.42627148	8.86030463	8.10156673	5.52050965	8.50749448	9.14675583	7.40956193	7.04659044	8.08834689	8.18337258	8.75220045	8.14235459
UBAP2L	7.82585534	7.91391525	7.75998602	7.97951233	7.87052275	7.23139608	7.87515579	8.02109606	7.85160498	8.24172978	8.26058794	8.1598215	7.78151093	7.89454663
WDHD1	4.47411767	4.41635533	4.68545963	4.16764851	4.46463956	4.4207664	4.82998998	4.80551327	4.26847224	4.78011255	4.79363478	4.89966324	4.74103876	4.66812537
SASH1	10.51628004	9.77458953	10.06392153	10.24652252	10.52139155	10.34614978	9.82613469	10.19883101	10.18060538	10.52746536	10.09555394	10.05591852	10.59693249	10.37839378
FBXL5	10.22845899	10.05411357	10.0939945	10.3364397	9.85867141	10.03325375	10.11321207	10.14854794	9.97014252	9.77136451	9.63273219	10.31882603	9.79133142	9.82887199
EPCAM	8.61529182	8.89009342	8.51409588	8.7431051	8.04392178	6.5041083	7.21078739	9.1182256	8.70950584	8.06121087	8.80129021	9.47734776	7.70436757	8.33651964
CENPF	7.18018831	6.71475711	6.62484846	7.02900027	7.12576751	5.56987455	7.11194288	6.97154118	6.74309438	6.77157275	6.97706221	7.43078186	7.06325181	6.88658241
RMND5B	8.91992524	8.46617882	8.57319034	8.99077997	8.66416775	8.30445149	7.72932036	8.6089726	8.67996088	8.74918258	8.34268947	8.95482214	8.76322243	8.7184929
SLC24A3	9.01076924	8.5135893	8.55237991	8.83654721	8.49417682	8.92448917	8.08629533	8.43685508	8.41843275	7.83138939	8.2920076	8.59603257	8.84372144	8.26406624
IL1RN	9.81298433	10.13060881	9.49465421	10.15752977	9.56325814	9.84455532	9.2397254	9.84504561	9.83368554	9.81933024	9.38028142	10.70624468	11.34185209	10.54933699
STX17	7.22318187	7.57868402	7.30954538	7.67453407	7.55892368	7.32687365	7.27900088	7.46006892	7.39409634	7.63119976	7.11785218	7.35046781	7.6086393	7.38041299
HNRNPU	7.89923302	7.60689437	7.79762839	7.76678828	7.79922603	7.34188635	7.76583094	7.6761859	7.65665287	7.73766043	7.90336851	7.94030232	7.83931773	7.90938188
STX12	9.48911146	9.3871369	9.57480415	9.98088038	9.2831292	9.52294837	9.07204994	9.31115937	9.61557043	9.58081206	9.27689903	9.93610707	9.40033947	9.49180314
UBL3	10.13141731	9.98550226	9.93407372	10.68976477	10.36172037	10.67341163	8.89095513	10.11172099	10.20462046	9.91750173	9.21107221	10.10136649	10.7824468	9.98650529
CRYL1	7.83009616	7.92897367	8.06710065	7.93285956	7.87164775	8.03184772	8.08723193	8.46287986	8.1900022	7.69867131	7.76465061	8.31067522	8.40983759	7.82833084
SASH1	10.72056055	10.14703612	10.27921883	10.54207274	10.71198049	10.69189942	10.0173829	10.44926593	10.40168463	10.67509997	10.35114865	10.34908469	10.86818786	10.54391943
GYS2	5.22167031	5.82753726	6.1679072	6.96654828	5.85227365	6.69744432	4.54601979	7.06086179	6.82353544	7.07158754	5.86860954	6.1929242	7.33323949	6.88343265
ITPR2	5.48040257	5.50667717	5.47995416	6.10075794	5.28078196	4.78318409	5.443267	5.66045258	5.26109871	5.18276394	5.65382643	5.59884468	5.57363046	5.54764629
DLG2	4.92486349	4.70579185	4.44317328	4.88258247	4.71845457	4.34698864	4.38577727	4.70126627	4.33550294	4.37569332	4.48383667	4.92462479	4.67677217	4.66812537

Figure 1-4: Screenshot of a micoarray dataset

1.2.2 RNA sequencing (RNA-seq) Data

RNA-seq technology is used to quantify the expression levels of all genes in a sample of tissue. It is similar to microarray analysis, but instead of measuring the presence and amount of specific DNA probes, RNA-seq measures the presence of amount of all RNA molecules in the sample. RNA-seq is based on next-generation sequencing (NGS) technology, which uses high-throughput sequencing to determine the order of the nucleotides in a given RNA molecule [2]. It determines the exact sequence of nucleotides (A, C, G, and U) in a given RNA molecule. RNA-seq data is used in a wide range of applications, including the study of gene expression, the identification of novel transcripts, the detection of gene regulation, and the study of post-transcriptional processing. RNA-seq can also be used to detect mutations and to provide insight into the regulation of gene expression. One common approach to analyzing RNA-seq data is by creating a count matrix that represents the gene counts per sample (figure 2-4) . This matrix is then analyzed using count-based models, which are often constructed based on the negative binomial distribution. The data can also include additional information such as the type of tissue or cell the sample was taken from, the age of the sample, and the environment in which the sample was taken. This additional information helps to understand the gene expression levels better in the sample. Count data can be used to identify differentially expressed genes and pathways, as well as to identify changes in gene expression over time. Count data can also be used to quantify gene expression levels, allowing the identification of potential biomarkers and the development of therapeutics. Count data can also be used to compare expression levels between different tissue types or developmental stages. Count data is an invaluable tool in the study of gene expression and can provide valuable insights into gene regulation and disease.

1.2.3 Single-cell RNA sequencing (scRNA-seq) Data

scRNA-seq technology allows researchers to measure the expression levels of all genes in a single cell instead of an entire population [3]. Unlike other RNA-seq techniques, scRNA-seq does not require the sample to be homogenized, so it can be used to analyze the gene expression profiles of individual cells (Figure 1-6). scRNA-seq technique generated count data (figure 1-7) can be represented as a matrix, where rows correspond to genes and columns represent individual cells. The elements of the matrix indicate the expression levels of genes in each cell. This high-throughput technique allows researchers to detect and analyze rare cell types,

	SRR26781	SRR26782	SRR26781	SRR26781	SRR26781	SRR26781	SRR26781	SRR26781	SRR26781	SRR26781
TSPAN6	487334	230175	755395	218735	594013	203803	664028	30823	607115	301120
TNMD	112	6456	1796	90	48	0	400	0	200	0
DPM1	136067	211354	217312	232908	85679	294578	143579	28671	123227	236740
SCYL3	108222	100620	66834	95177	54686	77272	82816	33331	51244	178516
C1orf112	53729	49231	27830	100881	26806	53459	32015	25674	33208	148618
FGR	33130	214516	57784	16410	22689	83222	63777	29546	17222	120581
CFH	847044	1554904	377713	61958	585824	286227	617931	796034	367293	243260
FUCA2	89570	341181	165196	180166	116103	206711	83445	114527	85872	489335
GCLC	215471	185091	199275	54752	56163	343843	192132	48493	80959	592239
NFYA	76475	66384	35952	102050	29133	137503	73524	18113	25154	217990
STPG1	34929	14858	24034	59440	27644	105327	26984	16414	36852	157836
NIPAL3	186828	56228	160763	140235	155540	66789	133236	39284	147123	171077
LAS1L	92300	135241	108714	304423	82360	182414	110260	149210	107593	427382
ENPP4	34348	68509	40635	2300	24910	34229	26384	17575	18619	76788
SEMA3F	214580	78107	185093	76068	178335	234922	451588	176560	234422	522379
CFTR	18917	867	22755	55	4292	3509	5644	0	3148	5704
ANKIB1	199254	234024	113331	116170	48746	259979	84899	14156	47035	174066
CYP51A1	142065	118713	237043	247544	130928	247156	308330	50933	131501	340706
KRIT1	205150	158209	70091	46962	43050	173460	70020	18889	34930	150663

Figure 1-5: Screenshot of a RNA-seq count data

gene expression variations between cell types, and the heterogeneity of complex tissues and measure alterations in gene expression over time or in response to an environmental stimulus. scRNA-seq is more expensive and time-consuming than traditional RNA-seq, but provides a much more detailed view of gene expression in individual cells.

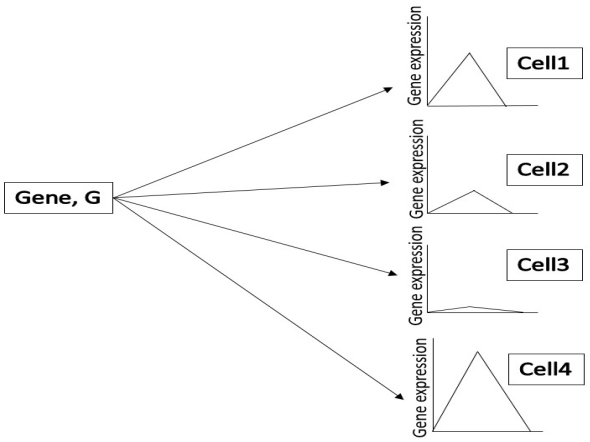


Figure 1-6: In different types of cells within the same tissue, a gene can be expressed at different levels and regulated by different control mechanisms.

Gene_names	SRR3579502	SRR3579503	SRR3579504	SRR3579505	SRR3579506	SRR3579507	SRR3579508	SRR3579509	SRR3579510	SRR3579511	SRR3579512
A1BG	0.192859	0.211661	3.55604	0	0.401412	0.418335	0.405444	0.523657	11.28916	0.3493	0.072501
A4GALT	10.504582	0.773935	3.715087	0	1.054334	0	0	0	1.297082	1.294943	4.462896
AAAS	48.951829	13.84899	45.174011	39.946515	43.560397	1.013684	58.434871	29.633865	26.521199	25.392367	21.284159
AACS	22.217662	11.977955	33.459868	53.372515	12.116112	19.715317	2.860383	14.372082	18.838499	14.135527	22.638465
AAGAB	39.517958	72.703458	35.182839	44.250956	46.378565	39.47878	17.598419	52.656031	62.282032	20.783167	31.188088
AAK1	52.09774	42.708846	46.559193	25.916817	25.253448	54.435199	57.494816	33.649443	33.664663	68.158958	53.628838
AAMDC	144.826506	107.434854	150.823233	157.339122	121.284586	80.318793	63.006427	144.285432	71.151353	186.144802	142.544412
AAMP	129.215682	95.372441	146.797158	119.283556	61.644767	107.991105	65.969273	58.13599	159.060605	101.25939	140.003466
AAR2	5.589466	21.09468	36.007997	24.135579	15.911111	7.006115	20.468391	36.907712	9.621016	28.972602	16.9755
AARD	0	3.542869	0.872424	5.840985	1.18377	0.823942	8.213304	0.159437	7.491501	8.145478	11.218729
AARS1	48.090528	47.888851	29.495752	38.226956	57.47265	20.339344	74.211613	46.288428	28.482507	25.267587	38.116794
AARS2	4.667899	12.921247	1.689467	10.626074	14.927309	11.802244	2.314339	9.32945	1.209465	9.424335	3.540992
AARSD1	59.888556	37.506091	52.471751	27.136306	60.495912	85.26328	37.371126	21.691603	68.790125	139.172657	106.621989
AASDH	0.957921	7.140927	3.776621	4.13682	5.237583	1.306455	0	5.640572	2.155313	4.650174	8.23874
AASDHPPT	15.994185	16.941843	25.964272	10.928822	31.878405	26.683659	36.275853	24.104862	22.524238	17.789242	51.069216
AATF	79.464873	49.154692	79.071922	62.478139	106.153541	64.551695	69.044752	119.071379	64.908246	84.495662	33.063324
ABAT	2.499635	0.255089	2.581853	6.767813	6.856265	0.718548	0.152603	0	6.513614	2.564187	0
ABCA2	2.346874	1.711131	4.761079	7.984226	8.307341	5.951667	1.619915	2.073505	18.386521	2.875179	8.949066
ABCA5	6.774078	13.077227	0	39.031999	20.707156	3.949336	4.544848	1.165851	5.749682	6.073585	28.000378
ABCA7	24.364671	21.166474	15.538966	12.369079	8.07234	4.010337	14.120329	6.366498	45.086304	55.791649	68.030182
ABCB10	1.908987	8.169665	0.503502	3.135993	7.764868	12.294037	0.04426	6.277526	8.606598	0.051903	0
ABCB6	22.651755	48.301889	33.743603	38.796391	23.703894	19.054547	21.613972	34.396242	9.283182	9.987846	10.424255
ABCB7	34.074408	40.688008	41.282732	30.976323	34.867599	12.56416	32.68343	49.217807	33.281148	19.563709	32.367798
ABCB8	59.421594	33.730901	40.190206	10.640185	19.440431	3.89409	10.179389	56.190167	2.259351	48.961104	31.883583
ABCB9	0.196685	5.969372	9.274591	1.103959	15.857321	8.357322	13.83864	1.105176	1.953963	4.034054	8.287256
ABCC1	8.818969	6.559603	6.200729	13.156986	4.143249	14.166628	11.171353	10.74096	13.412867	10.284897	7.512505
ABCC10	6.5646	8.887373	6.277798	0	13.509573	3.085832	0.726301	14.80424	2.057119	11.442354	14.175067

Figure 1-7: Screenshot of a scRNA-seq count data

1.3 Gene Expression Data Analysis

The amount of messenger RNA (mRNA) produced by a gene in a particular situation is represented by its expression profile. With the advent of newer transcriptomic technologies, it has become possible to measure the gene expressions inside millions of cells in a single experiment. A biologist generates the biological data whereas a computational expert has the job of mining the information hidden inside the high dimensional biological data using efficient algorithms. Gene expression data such as DNA microarray, RNA-seq, and scRNA-seq are high dimensional. Data analysis in biological research is performed in a systematic way and the validation of the final findings is an absolute necessity. Gene expression data analysis is a systematic approach to understand the functions and interactions among genes. Datasets have different characteristics and the type of dataset determines which tools and techniques can be used to process and analyze the data.

Gene expression data is used to study gene expression levels in various organisms, including humans, animals, plants, and bacteria. It is also used to study the effects of gene expression on phenotypes, diseases, and other traits of interest. The data is used in a variety of fields, including bioinformatics, genetics, and genomics. Gene expression data can provide valuable insight into the functioning of a gene. For example, it can reveal whether a gene is upregulated or downregulated in response to a particular stimulus. It can also indicate the level of expression of a particular gene in a given cell type or tissue. Furthermore, it can be used

to compare the expression of a gene in two different samples, such as normal and cancerous tissues. It can also be used to make predictions about gene function and to identify potential therapeutic targets.

Extracting a set of modules (i.e., groups of closely associated genes) with high biological significance from a biological (co-expression or regulatory) network supports identication of interesting behavior of a set of participating driver or causal genes across the states of a disease. Identifying causal or driver genes or interesting biomarkers for a given disease across conditions with high precision is the initial step in isolating genetic causes of diseases. To identify such biomarkers, the process exploits co-expression, diferential co-expression, or diferential gene expression analysis approach that uses statistical, data mining or machine learning techniques to analyze gene expression data [4].

1.4 Data Mining

Machine learning is a type of data analysis that uses algorithms to identify patterns in data sets and make predictions or decisions based on the data. Data Mining, a crucial part of Machine Learning, is the process of exploring large datasets to discover patterns, trends, and correlations that may be used to make predictions or decisions [5]. It involves the use of sophisticated algorithms and software to analyze large datasets to identify patterns, find relationships, and uncover trends. Supervised learning and unsupervised learning are two main categories of machine learning techniques, and they differ in their approach to handling labeled or unlabeled data. In supervised learning, the algorithm is trained on a labeled dataset, where each input is paired with the corresponding correct output. The goal is to learn a mapping from inputs to outputs. Unsupervised learning is a type of machine learning where the algorithm is presented with data that has no predefined labels or target outputs. The goal is to discover patterns, structures, or relationships within the data without explicit guidance on what to look for. Unsupervised learning is particularly useful for exploratory analysis and uncovering hidden insights. Clustering is one type of unsupervised learning techniques. Data mining algorithms are used to identify patterns in data that may be too complex or too large for humans to identify. These algorithms can be used to classify data, cluster data, and detect anomalies. Data mining is an important tool in bioinformatics and has been used to uncover new knowledge and insights into biological systems. With the increasing volume of biological data, the relevance of data mining techniques has also been increasing simultaneously. Data mining tasks are

the specific objectives or goals that a data mining process aims to achieve. These tasks define what the analyst or data scientist intends to discover or extract from the data. Data mining tasks can broadly be categorized into predictive and descriptive tasks [6]. Data mining techniques, on the other hand, are the methods or algorithms used to perform these tasks. These techniques are the tools that enable the extraction of valuable patterns, insights, or knowledge from the data.

1.4.1 Predictive Data Mining Tasks

Predictive data mining tasks involve forecasting future trends or outcomes based on historical data. Algorithms analyze past observations to make predictions about unknown future events. Examples include predicting stock prices, customer churn, or medical diagnoses. Examples of data mining techniques used in predictive task include:

Classification: In classification tasks, data mining algorithms assign predefined labels or categories to new data based on the characteristics of previously labeled data. Classification is used to predict the class label of objects for which the the class label is unknown. This is useful for tasks such as spam email detection, sentiment analysis, disease detection, or disease diagnosis. Classification analysis uses popular methods such as Decision Trees (e.g., C4.5, CART), Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Naive Bayes.

Regression: Regression tasks involve predicting a continuous numerical value based on input variables. Algorithms build models to estimate relationships between variables and make predictions. Examples include predicting house prices based on features like location and size or forecasting sales revenue based on marketing expenditure. Linear Regression, polynomial regression, Ridge regression, and Lasso Regression are some popular methods under regression analysis.

1.4.2 Descriptive Data Mining Tasks

Descriptive data mining tasks focus on summarizing and understanding the characteristics and patterns present in the data. These tasks aim to provide insights into the underlying structure of the data without making predictions about future outcomes. Examples of data mining techniques used in descriptive task include:

Clustering: Clustering tasks involve grouping similar data points together based on their attributes or features. This helps identify natural groupings or clusters within the data, revealing patterns or relationships that may not be apparent initially. The formation of clusters ensures that objects or data points within the same cluster exhibit high similarity to each other but are notably dissimilar to objects in other clusters. Each cluster can be interpreted as a distinct class of objects, from which rules can be derived. Clustering methods include K-Means, hierarchical clustering, DBSCAN, Gaussian Mixture Model, Agglomerative Clustering. Examples: disease biomarker discovery, social network analysis, traffic flow analysis etc.

Association rule mining: Association mining tasks discover relationships and associations between variables in large datasets. These tasks identify frequent patterns, co-occurrences, or correlations among variables, which can be used for market basket analysis, recommendation systems, or cross-selling strategies. Algorithms for association mining include Apriori and FP-Growth (Frequent Pattern Growth).

Outlier analysis: Outlier analysis, also known as anomaly detection, is a data mining technique that focuses on identifying data points or instances that deviate significantly from the general pattern or behavior of the dataset. Outliers are observations that are rare, unusual, or different from the majority of the data. The goal of outlier analysis is to recognize these exceptional cases, which could represent errors, anomalies, or valuable insights depending on the context of the data. Example: Credit card fraud detection. Isolation Forest, One-Class SVM, and Local Outlier Factor (LOF) are some examples of outlier analysis methods.

1.5 Motivation

Interesting biomarker identification for a given disease is a challenging task. Generally, differential and (differential) co-expression analysis of gene expression data are commonly used to find interesting biomarker(s). Integration of multiple data sources or methods during such analysis increases accuracy and robustness of the results. There are many tools available for differential and (differential) co-expression analysis of GED. But, they have many limitations and handling of these limitations is important before potential biomarker(s) identification. From our study, we observe that although the researchers have been able to identify biomarkers for some diseases, some diseases still need serious attention for biomarker(s)

identification. These problems motivate to develop effective differential and (differential) co-expression analysis for effective interesting genes finding. Generally, all genes present in a dataset are not interesting. An optimal group(s) of genes play key roles in disease developmental stages or progression or tissue types differentiation. Have all the data mining techniques for gene expression data analysis reported so far, been able to extract interesting novel and crucial genes that can be used as disease biomarkers? In our opinion the answer is no and this has motivated us to carry out our present study.

1.6 Research Objectives

The objective of my Ph.D. work is to analyze gene expression data to identify potential biomarker(s) and to evaluate the analysis results using ground truth knowledge and gold standards available with the external resources. In order to achieve this objective, the following steps have been carried out.

- The first objective is to study the performance of existing biclustering methods over a number of microarray datasets for ESCC disease and observe the variations in the performance obtained. Moreover, most of the biclustering techniques are limited to finding biclusters but it is important to perform gene-based analysis to identify relationships and potential biomarkers among genes to establish the association with disease in progression. Hence, it is aimed to develop a biclustering approach to support identification of interesting biomarkers for a disease dataset. It was expected that this multi-objective study will enable us to identify several interesting biomarkers for ESCC dataset.

- It has been found that Esophageal Squoumous Cell Carcinoma (ESCC) disease has not been clearly understood using only microarray data. So, next objective is to find interesting and novel disease biomarker(s) using DE analyses on RNA-seq datasets for ESCC disease.

- The experiments carried out with a single source of data often have been found biased. It can be believed that a use of a combination of multiple sources of data will give unbiased results. Next objective is to develop an integrated framework to support handling of multiple sources of data to identify DE genes towards finding of potential gene biomarkers for microarray as well as RNA-seq datasets for ESCC disease.

- Development of a measure which can identify positive and negative shifting, scaling and shifting-and-scaling patterns at the same time is my next objective. It was also aimed to find important and enriched genes using the proposed measure to extract biollogically significant network modules.
- In recent gene expression data analysis research, an attention has been increased by scRNA-seq data due to its information rich contents. So, it is aimed to explore identifying potential biomarkers for scRNA-seq data of ESCC disease using an integrative method.

1.7 Research Contributions

To address issues associated with different problems in gene expression analysis mentioned above, a number of solutions have been introduced which are described, next.

1.7.1 Biclustering-based Biomarker Identification in ESCC Microarray Data

A potential biomarker identification method called PD_BiBIM has been developed. The method is based on biclustering approaches and uses microarray datasets of esophageal squamous cell carcinoma (ESCC) disease to extract insights relevant to ESCC. Here, several biclustering techniques have been considered and accepted those techniques which are found effective from enrichment perspective for subsequent analysis. Based on biclustering results, gene networks have been constructed and carried out a topological, pathway and causal analysis on the modules extracted from the networks. This method identifies several potential biomarkers for esophageal squamous cell carcinoma (ESCC) such as IFNGR1, CLIC1, CDK4, and COPS5, after applying a ranking scheme.

1.7.2 Identifying Crucial Genes in ESCC GED using Differential Expression Analysis

In this method an attempt has been made to identify a set of crucial genes for Esophageal Squamous Cell Carcinoma (ESCC) using Differential Expression analysis followed by gene enrichment analysis. To validate the method, RNA-seq

datasets are used. Initially, a subset of up-regulated and down-regulated genes are identified based on adjusted P-value and log-fold change value. Then, co-expression networks and PPI networks are constructed on selected genes to investigate the interactions and associations among these genes. Finally, enrichment analysis is performed to filter out the most crucial subset of genes which are also established to have strong association with the ESCC. Three genes, namely TNC, COL1A1, and FN1 are found most closely relevant to ESCC.

1.7.3 Identifying Crucial Genes in Esophageal Squamous Cell Carcinoma using an Ensemble Approach

This method has been introduced to remove the biasness of resulting DEGs given by a DEG tool towards the identification of critical genes for ESCC. It is a consensus function on which user can rely on the output generated by differential expression analysis methods applied on multiple sources of data. Both microarrays and RNA-seq data are used in the analysis. Initially, independent downstream analysis on each type of data using six differentially expressed gene identification tools followed by an integrative analysis supported by an effective consensus function is conducted to identify an unbiased set of differently expressed genes. Next, differential co-expression analysis is performed and identified a set of low preserved modules. Finally, hub genes are identified from the selected low preserved modules and validated both topologically and biologically. A set of hub genes are identified such as SOX11, COL27A1, TOP3A, BAG6, CDC6, EZH2, COL7A1, G6PD, and AKR1C2 which have been established to be critical for ESCC.

1.7.4 An Advanced Measure for Co-expression Network Analysis

A novel similarity measure called SNMRS has been developed based on existing NMRS measure [7]. SNMRS yields correlation values in the range of 0 to +1 corresponding to negative and positive dependency. To study the performance of our measure, internal validation of extracted clusters resulted from different methods is carried-out. Based on the performance, hierarchical clustering has been chosen and applied the same using the corresponding dissimilarity (distance) values of SNMRS scores, and utilized a dynamic tree cut method for extracting the dense modules. Modules are validated through literature search, KEGG pathway, and gene-ontology analyses on the genes representing the modules. Our measure can

handle all types of correlations and provides a better performance than several other measures in terms of cluster-validity indices. Also, SNMRS based module detection method have been found to extracts more biologically relevant and interesting patterns from gene microarray and RNA-seq data.

1.7.5 Identification of Potential Prognostic Biomarkers for ESCC using Single-cell RNA Sequencing Data Analysis

This chapter analyses the difference between parental cells and cells that acquired radioresistance using scRNA-seq data and investigates the dynamic changes of the transcriptome of cells in response to fractionated irradiation (FIR) towards the identification of potential biomarkers for Esophageal Squamous Cell Carcinoma (ESCC). We use an effective pipeline to investigate cell heterogeneity and to identify potential biomarkers in scRNA-Seq data using differential expression analysis (DEA). The divergence of gene expressions is analyzed in response to FIR and the dynamic changes in differentially expressed genes (DEGs) of KYSE-180 cells with two different cumulative doses of FIR (12-Gy and 30-Gy). We construct several biological networks and observe relative to control (0-Gy), 30-Gy induced higher variability of genes. Four hub genes TXN, IER2, PCNA, and CENPF are identified which are involved in ESCC progression.

1.8 Organization of the Thesis

Chapter 2 presents background of gene expression data analysis, Microarray, RNA-seq, scRNA-seq, ESCC disease, biomarker identification, and the use of data mining in these analyses are discussed in details. Various datasets and tools that are used in my research work are also discussed in the chapter. Chapter 3 discusses biclustering techniques and presents own biclustering approach called PD_BiBIM to find biomarkers from ESCC datasets. Chapter 4 presents an ensemble of differential expression analysis methods to identify potential biomarkers from ESCC microarrays and RNA-seq data. Chapter 5 presents a similarity measure named SNMRS and an approach to find biomarkers. Chapter 6 introduces and explains scRNA-seq data analysis techniques and explains the proposed framework to identify biomarkers from ESCC data. Finally, Chapter 7 presents concluding remarks and highlights the future directions of research.