# GENOME- AND TISSUE-WIDE ANALYSIS OF ALTERNATIVE POLYADENYLATION EVENTS USING CLUSTERING AND FEATURE LEARNING METHODS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

PINAR YILMAZER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2019

Approval of the thesis:

**GENOME- AND TISSUE-WIDE ANALYSIS OF ALTERNATIVE POLYADENYLATION EVENTS USING CLUSTERING AND FEATURE LEARNING METHODS**

submitted by **PINAR YILMAZER** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Tolga Can
Supervisor, **Computer Engineering, METU** _____

**Examining Committee Members:**

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU _____

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Assoc. Prof. Dr. Mehmet Tan
Computer Engineering, TOBB ETU _____

Date:

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Surname: Pınar Yılmazer

Signature :

# ABSTRACT

## GENOME- AND TISSUE-WIDE ANALYSIS OF ALTERNATIVE POLYADENYLATION EVENTS USING CLUSTERING AND FEATURE LEARNING METHODS

Yılmazer, Pınar
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Tolga Can

September 2019, 73 pages

Alternative polyadenylation (APA) is a biological process that takes places during gene transcription and recent studies relate APA events with gene expression regulation and with diseases such as cancer. Studying gene expression across tissues for various conditions is crucial guiding scientists to work on biomarker discoveries and treatment options of the transcriptomic diversity related diseases. In this thesis, several novel genes and tissues that are more prone to 3'UTR shortening, which is an APA event, in diseased state are presented by analyzing significant proximal APA events across human tissues. Most of the identified genes are also validated by existing studies. Furthermore, we demonstrate that hierarchically closer tissues share similar gene set or interactions according to APA events, and tissue hierarchy can be built by just considering top affected genes. Overall, this work covers creation of biological tissue hierarchy, comparison of tissue networks in normal/diseased states and feature learning analysis of protein-protein networks using APA events. To the best of our knowledge, no such human genome- and tissue-wide analysis, based on APA events, has been conducted before. Therefore, our multidisciplinary work may

guide researchers to the next step of genomics studies on diseases.

Keywords: Feature Learning, Hierarchical Clustering, Alternative Polyadenylation, 3'UTR Shortening, Cancer, Gene, Disease

# ÖZ

## ALTERNATİF POLİADENİLASYON OLAYLARININ KÜMELEME VE ÖZNİTELİK ÖĞRENME YÖNTEMLERİ İLE GENOM VE DOKU ÇAPLI ANALİZİ

Yılmazer, Pınar
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Tolga Can

Eylül 2019 , 73 sayfa

Alternatif poliadenilasyon (APA), gen transkripsiyonu sırasında gerçekleşen biyolojik bir süreçtir ve son çalışmalar APA olaylarını gen ekspresyonu regülasyonu ve kanser gibi hastalıklar ile ilişkilendirmektedir. Çeşitli durumlar için dokular arasında gen ekspresyonunun incelenmesi, bilim insanlarının, transkriptomik çeşitlilikle ilgili hastalıkların biyobelirteç keşifleri ve tedavi seçenekleri üzerinde çalışmak için çok önemli bir rehberdir. Bu tezde, hastalıklı bir durumda bir APA olayı olan 3'UTR kısalmasına daha yatkın olan bazı yeni genler ve dokular, insan dokularında önemli proksimal APA olayları analiz edilerek sunulmuştur. Tanımlanan genlerin çoğu, mevcut çalışmalarla da doğrulanmaktadır. Ayrıca, hiyerarşik olarak daha yakın dokuların APA olaylarına göre benzer gen setini veya etkileşimleri paylaştığını ve doku hiyerarşisinin sadece en iyi etkilenen genler göz önüne alınarak oluşturulabileceğini göstermekteyiz. Genel olarak, bu çalışma biyolojik doku hiyerarşisinin oluşturulmasını, normal / hastalıklı durumlarda doku ağlarının karşılaştırılmasını ve APA olaylarını kullanarak protein-protein ağlarının öğrenme analizini içerir. Bildiğimiz kadarıyla,

APA olaylarına dayanan böyle bir insan genomu ve doku çapında analiz yapılmamıştır. Bu nedenle, bu tezdeki disiplinerarası çalışmamız, araştırmacıları hastalıklar konusundaki genomik araştırmaların bir sonraki adımına yönlendirebilir.

Anahtar Kelimeler: Öznitelik Öğrenme, Hiyerarşik Kümeleme, Alternatif Poliadenilasyon, 3'UTR Kısalması, Kanser, Gen, Hastalık

To my family

## ACKNOWLEDGMENTS

I would first like to thank my supervisor Prof. Dr. Tolga Can for his supervision, constant support and guidance.

I am also grateful to my supporting family and friends for their help and patience which made it easier to overcome difficulties.

## TABLE OF CONTENTS

## LIST OF TABLES

TABLES

# LIST OF FIGURES

FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AHC | Agglomerative Hierarchical Clustering |
| APA | Alternative Polyadenylation |
| API | Application Programming Interface |
| BRENDA | Braunschweig Enzyme Database |
| BTO | Brenda Tissue Ontology |
| CBOW | Continues Bag Of Words |
| DNA | Deoxyribonucleic Acid |
| GEO | Gene Expression Omnibus |
| mRNA | Messenger Ribonucleic Acid |
| NJ | Neighbor Joining |
| OTU | Operational Taxonomic Unit |
| PAS | Polyadenylation Site |
| PCA | Principal Component Analysis |
| PPI | Protein Protein Interaction |
| RNA | Ribonucleic Acid |
| SLR | Short to Long Ratio |
| SPR | Subtree Pruning and Regrafting |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| UPGMA | Unweighted Pair Group Method with Arithmetic Mean |
| UTR | Untranslated Region |

# CHAPTER 1

# INTRODUCTION

Alternative polyadenylation (APA) has emerged as a novel mechanism which produces several isoforms of a gene with different 3' UTR length and increases diversity of gene expression. It is a commonly observed phenomenon affecting most of the genes in many different species. Although the role of RNA 3'end formation in regulatory variation remains mostly unexplored, it is known that it can affect the stability, translation efficiency and binding sites of microRNAs. Due to its critical regulatory effects, a lot of research is being conducted to characterize the underlying mechanism for different species and to advance controlling of gene activity.

Studies report that proliferation signals, differentiation factors and hormones lead to proximal APA. The resulting shorter mRNA isoforms are associated with rapid cell proliferation across various cell types and tissues, and may be the basis of some of the pathological events [11, 12, 13]. Although most of such events are unknown, proto-oncogene activation cases, observed in breast and lung cancer cells, and several disease signatures affecting heart, endocrine and hematology are likely to be correlated with alternative polyadenylation [14, 15, 16].

Polyadenylation events are also studied across tissues and tissue specific signatures are found in different species. Among many eukaryotes, we focused on researches working on human genome and performed this thesis study on a large gene expression dataset of normal and diseased human tissues. According to previous studies, some tissues were shown to produce overall longer or shorter mRNA isoforms. For instance, although tissues in central nervous system hierarchy express longer isoforms, placenta, blood, testis and ovaries are tend to show shorter isoforms [17, 18]

Because of all the reasons specified above, alternative polyadenylation has been receiving increasing interest in disease research. Moreover, prognostic predictions, treatment options and diagnostics are the top topics for current studies. In this thesis, we developed a strategy to analyze APA shortening events and identify several candidate genes in an organism-wide fashion using human tissues gene expression dataset with proximal to distal site ratios (SLR) for each gene. Since we were interested in shorter mRNA isoforms, we filtered the gene set, selected genes with higher SLR values compared to a threshold and organized tissue sets in different states accordingly. For this research, we performed three different analyses.

In the first analysis, normal, diseased, and mutual tissues were processed and dendrogram structures were constructed by using a distance matrix computed using SLR values. Comparing generated trees with a literature curated tissue hierarchy, we showed that hierarchically closer tissues contain similar gene sets filtered by SLR values and cluster similarly. Analogous tissue hierarchies can be constructed by just considering genes with high SLR values.

In the second analysis, PPI network of different tissues were processed and feature embeddings of each gene were created by modeling a multiscale tissue hierarchy. Two different embeddings have been created, one having all PPI data and other having filtered PPI data according to SLR values. Genes were mapped into low dimensional network to make an inference about similarity of tissues before and after filtering. We showed that proteins activated in similar tissues indicate correlation and clustered better in SLR based feature extraction. Although filtering performed with SLR information does not affect members of main tissue clusters, it changes surrounding neighbours.

In the third analysis, diseased and normal samples of tissues were processed according to gene-SLR based criteria. Tissues more prone to proximal APA were identified and several prominent genes showing significant 3'UTR shortening were proposed. We saw that tissues with similar APA differentiation in either disease or normal conditions are also positioned closer in the hierarchy. Also, we verified that many of the proposed genes were categorized as disease or cancer related.

## 1.1 Related Works

A lot of research conducted on quantitative and global analysis of APA try to develop a new reliable sequencing-based method for profiling RNA polyadenylation at the transcriptome level to overcome limitations of widely profiled RNA-seq and microarray methods [19, 20, 21, 22]. They follow various RNA-biochemical experimental steps to advance prevalent methods and promote prediction of greater number of expressed genes in diverse cell-types, stages, and species.

Apart from next generation sequencing methods, there are many studies working on previously detected APA sites and protein databases [23, 13, 24]. Achieving more comprehensive sets of genes, these methods generally focus on specific species or genes to characterize the diversity of polyadenylation and analyze different developmental or disease stages. Researches target not only mammals but also many other animal and plant species. However, they mostly do not make additional analysis in terms of tissue-specific gene expressions and differences across tissues and cell types. Our genome-wide APA analysis overcomes this limitation and provides a perspective by studying genes that undergo significant APA related changes in human tissues.

Some existing works, similar to our research, investigate alternative polyadenylation in tissue specific manner in mammalians[25, 17]. They compare genomic regions of tissues surrounded by polyadenylation sites and determine cis-regulatory elements. However, they experiment with a limited set of tissues and do not consider relationships between tissues. Among the ones modeling multiscale tissue hierarchy, Ohmnet provides a feature learning approach for multi-layer networks [26]. Yet, it mainly focuses on cellular function prediction and does not work on alternative polyadenylation events specifically.

## 1.2 Contributions

The main motivations of this thesis are to determine proteins showing APA shortening event across tissues and investigate whether hierarchically similar tissues share similar genes or interactions according to SLR values. Moreover, we aim to ana-

lyze mutual/differentiated genes by processing tissue samples of diseased and normal states and reveal common biological pathways, if there are any. To the best of our knowledge, such kind of genome-wide analysis covering construction of actual tissue hierarchies, comparison of tissue networks and feature learning analysis based on activated genes and SLR values has not been conducted before. Our work may lead to new insights for discovering the links between alternative polyadenylation and disease states of tissues sharing similar characteristics and may improve understanding in a genome-scale fashion.

## 1.3 Thesis Outline

The rest of the thesis is organized as follows. In Chapter 2, we briefly summarize biological and mathematical background related to our work. In Chapter 3, we present the structural comparison of the actual tissue hierarchy and the dendrograms, which were constructed by hierarchical clustering methods. In Chapter 4, we describe the feature learning approach to find feature embeddings of proteins in a multi-layer tissue network and the tissue hierarchy and then compare tissue similarities in different networks. In Chapter 5, we analyze diseased/normal tissues with respect to genes which demonstrates significant 3' UTR shortening and report some candidate genes. We empirically evaluate our results in Chapter 6 and conclude with directions for future work.

# CHAPTER 2

# BACKGROUND

## 2.1 Biological Background

High-throughput data-generating genomics experiments, development in the gene sequencing techniques and high demand of the analysis and interpretation of various types of genomic data have given rise to the interdisciplinary field, bioinformatics, which primarily combines molecular biology and computer science as well as various subfields. It processes the structure of biomolecules on a large scale by using wide range of computational techniques [27]. Although comprised fields have been increasing each day, bioinformatics mainly focuses on sequence alignment, protein structure and interaction predictions, noble gene identification and drug discovery. Gene expression data is one of the most fundamental and valuable resource to discover biological characteristics of the organism. It gives information regarding transcriptional, translational, folding and splicing phases. There are many techniques performing transcriptome analysis such as qPCR, expression microarrays and RNAseq [28]. They enable researchers tackling a wide range of biological problems by examining the expression levels of vast number of distinct genes simultaneously. Differentiating biologically critical isoforms, detection of genetic modifications, identification of novel transcripts and post transcriptional variations are main fields studied in gene expression analysis [29]. Alternative polyadenylation (APA) and alternative splicing are two main processes which lead to post transcriptional variations and play important role during eukaryotic gene expression by increasing coding potential. Most of the protein coding transcripts in eukaryotic cells excluding histone are affected by alternative polyadenylation. It has functional roles in tissue specific differentiation and activation of different physiological and disease states. Understanding APA mech-

anism contributes to numerous advances in human health and gains popularity each day. The importance of the APA motivated us to analyze organism-wide APA events and to study tissue-tissue and tissue-disease relationships with respect to gene data showing APA shortening events and tissue based protein to protein interaction networks, in this thesis. In the following section, basic concepts of molecular biology are reviewed to get acquainted with related concepts.

Studying genome sequence is crucial for scientists to figure out how genes direct the growth, development and maintenance of all biological contexts, including prokaryotic and eukaryotic organisms, as well as viruses. It also helps scientists to identify mutations in genes and diseases which makes possible to deliver more effective and personalized treatments. Genes are the sections of deoxyribonucleic acid, DNA, which contains biological and functional instructions. It is responsible of building proteins, developing physical characteristics and providing backup information of every piece of data in cellular level. Recent studies found that specific to human genome, there are 20,687 known protein coding genes which corresponds to 2.94% of genome. Remaining 97% represent gene regulatory regions and nonsense DNA whose functionality has not known yet [30, 31]. DNA is a continuous chain of nucleotide subunits, each composed of a five-carbon sugar, at least one phosphate group, and one of the four nitrogenous bases adenine, cytosine, guanine, and thymine which determines the nucleotide type [32]. Different combinations of nucleotides affect the information for building and sustaining an organism. The structure of DNA is three dimensional double helix with two strands connected by hydrogen bonds and twisted around each other like a spiral. DNA strands are non symmetrical and have two ends, phosphate-bearing (5'), and hydroxyl-bearing (3'). Two complementary strands run in opposite directions and 5' end aligns with 3' end. Upstream and downstream terms are used to identify relative positions of RNA and DNA strands such that while area towards to five prime end is called upstream, area towards to three prime end is called downstream.

As a part of DNA, genes hold the instructions for the synthesis of proteins which organize the cells, transmit messages, manage chemical reactions and responsible from many other vital functionalities of the tissues and organs. Protein synthesis from genes is a complex process and requires lots of modules to work together. It be-

gins with translation procedure which is performed by RNA polymerase enzyme and pre-mature mRNA is constructed from specific transcription unit of DNA. RNA polymerase recognizes starting and finishing point on genome and creates RNA strand by adding RNA nucleotide one at a time. RNA and DNA nucleotides differ in the type of the nitrogenous base they use. RNA uses Uracil instead of Thymine as complementary to Adenine. During transcription, DNA is copied from 3' end to 5' end which yields RNA polymerase to add nucleotides to the 3' end of complementary mRNA strand. The transcribed pre-mature RNA has untranslated regions at both ends as well as introns and exons [33].

### 2.1.1 Post-transcriptional modification

Maturing process has three major modification steps which occur almost simultaneously. The first step is addition of a 7-methyl guanosine cap to the 5'-end to handle recognition and attachment to ribosome. Remaining steps are 3' polyadenylation and splicing, both may lead to producing different proteins from the same transcription unit.

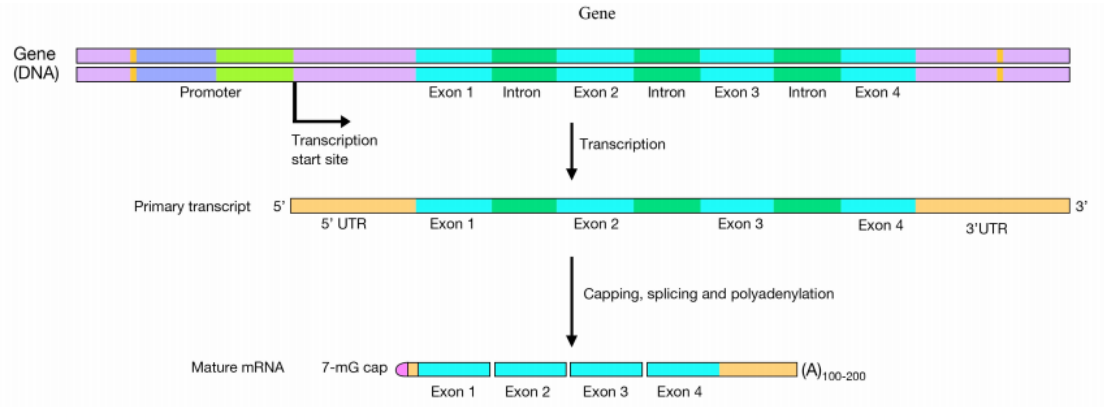


Figure 2.1: Steps in post-transcriptional modification of eukaryotic messenger RNAs. Reprinted from Biochemistry Free & Easy by K. Ahern and I. Rajagopal Press; 3rd Edition edition, February 12 2015, retrieved from https://bio.libretexts.org/ Licensed by CC BY-NC-SA 3.0

### 2.1.2 Splicing and Alternative Splicing

Pre-mRNA has two types of sequences, exons and introns. While introns are the non-coding sections to be removed during maturing process, exons are the sections having code for protein synthesis and need to be translated. Splicing process stands for remaining exons and removing introns to generate the final mature mRNA. There are many variations of splicing events which lead to distinct mRNA isoforms and proteins. Some removes all introns but keeps different combinations of exons. Some excludes part of exons treating them as introns and some keeps part of introns as well as exons. These variations are known as alternative splicing and associated with the tissue type in that the transcription process occurs. Beside protein differentiation, alternative splicing also causes human genetic diseases derived from splicing mutations and need to be studied carefully [34].

### 2.1.3 Polyadenylation and Alternative Polyadenylation(APA)

During maturation of eukaryotic mRNA, 3' end is also modified to enhance the specificity of the recognition and protect strand from ribonuclease digestion. First, the canonical polyadenylation signal sequence, AAUAAA, generally followed by a GU-rich sequence, is marked on the upstream of the actual cleavage site and binded by a multiprotein complex. Then, 3' end is shifted to right position for cleavage. With the help of cleavage factors and polyadenylate polymerase enzyme, mRNA is cleaved between PAS and the GU-rich sequence. The complex is seperated, and the splitted 3' end degrades. Following endonucleolytic cleavage, polyadenylate tail synthesis is started by adding lots of adenine residues onto the upstream cleavage area. As soon as the tail reaches its full length, stop signal is emerged and process is terminated. Size of polyadenylate tail varies between species, e.g. in humans it is 250–300 on average [35].

Apart from its usual place near 3' end, polyadenylation signal sequence can be seen in different parts of mRNA which yields different transcript isoforms from the same gene which is called APA. Signals located in various parts of 3' untranslated region (3'UTR) leads to shortening/lengthening in transcript by creating different poly(A)

positions. This type of APA does not affect protein coding frame but the availability of the binding sites, translation efficiency and stability of mRNA and also may indicate serious health conditions such as rapid cell proliferation and cancer [12, 15]. Signals located in internal exons or introns also lead to APA creating distinct protein isoforms but it is much less common when it is compared to APA events in 3'UTR [36].

### 2.1.4 APADetect Tool and the SLR Concept

Alternative polyadenylation is accepted as a leading regulatory mechanism affecting many cellular operations including development and diversity. Therefore, there are many ongoing researches to detect APA events using microarray technologies such as RNA-Seq and Microarray. Although, there is a rising demand for the analysis, APA event detection continues to be a challenging problem in genomics. As we mentioned before, several polyA regions can be observed during post-transcriptional phase which leads to shortening/lengthening of the 3'UTR sector. The 3'UTR end is also a common target region in microarray experiments, because of its tendency of staying as a tail, i.e., not forming RNA secondary structures. Therefore, most probes in Affymetrix chips are designed to target the 3' UTR section.

In this thesis, SLR (Short isoform to Long isoform expression Ratio, or proximal to distal ratio) values produced by a microarray-based method, APADetect tool, are used to detect 3' UTR isoform variations of genes. APADetect is a cross platform probe level analysis tool to screen and identify potential APA events, showing differential intensities [37, 38]. It uses previously known poly(A) positions [39] to divide probe sets into two differentially expressed groups, proximal and distal, and process hybridization levels of each probe to calculate intensities. It also searches for unknown poly(A) sites by studying probe sets and by discovering groups showing statistically important expression difference.

Probes to the upstream and downstream of a polyA site are partitioned into a proximal group and a distal group, respectively. If a probe overlaps with the poly(A) site, it is ignored. While high intensities in proximal group indicates a shortening event, high intensities in distal group shows a lengthening event. By computing average intensities for each group, proximal to distal ratio is computed and named as the SLR

value.

APADetect applies several filters to eliminate outlier samples and increase accuracy during calculation of SLR values. First, if all probes are either placed completely in the distal or the proximal group, it is not possible to detect and APA event and genes having this type of probe distribution are removed from the experiment set. Second, degraded samples whose distal intensities are significantly higher than proximal intensities are extracted from the sample set. Third, median of absolute differences of each probe for respective group is calculated and deviating probes according to a selected threshold value are discarded. The last filter works similar to the third, but it filters out samples deviating from related control groups.

Studies show that, APA may has a functional role in tissue-specific differentiation such that 3'UTR length difference is observed more in some tissues like ovary, brain, and adrenal than others [40, 17, 18]. Higher SLR value indicates shorter 3'UTR isoform of the transcript, which may show connection with rapid proliferation of cells, higher level of proteins and activation of proto-oncogenes in cancer cells [11, 12, 37]. Altogether, SLR values may help scientist to figure out the roles of APA events in different physiological and disease states and they can be used to examine diverse genes and to find hierarchical similarities between tissues.

## 2.2 Mathematical Background

### 2.2.1 Hierarchical Clustering

Clustering is one of the most important methods to group similar data points by discovering hidden patterns in the data. There are many clustering techniques developed with the current discoveries on data science. Hierarchical clustering is one of the leading and easy to understand technique among them. It requires a distance matrix or raw data as an input and produces a tree like structure, dendrogram, which shows hierarchical relationship between clusters. Initial parameter settings are not required and handled by linear/non-linear regression models. Two top-level methods for finding hierarchical clusters are agglomerative and divisive.

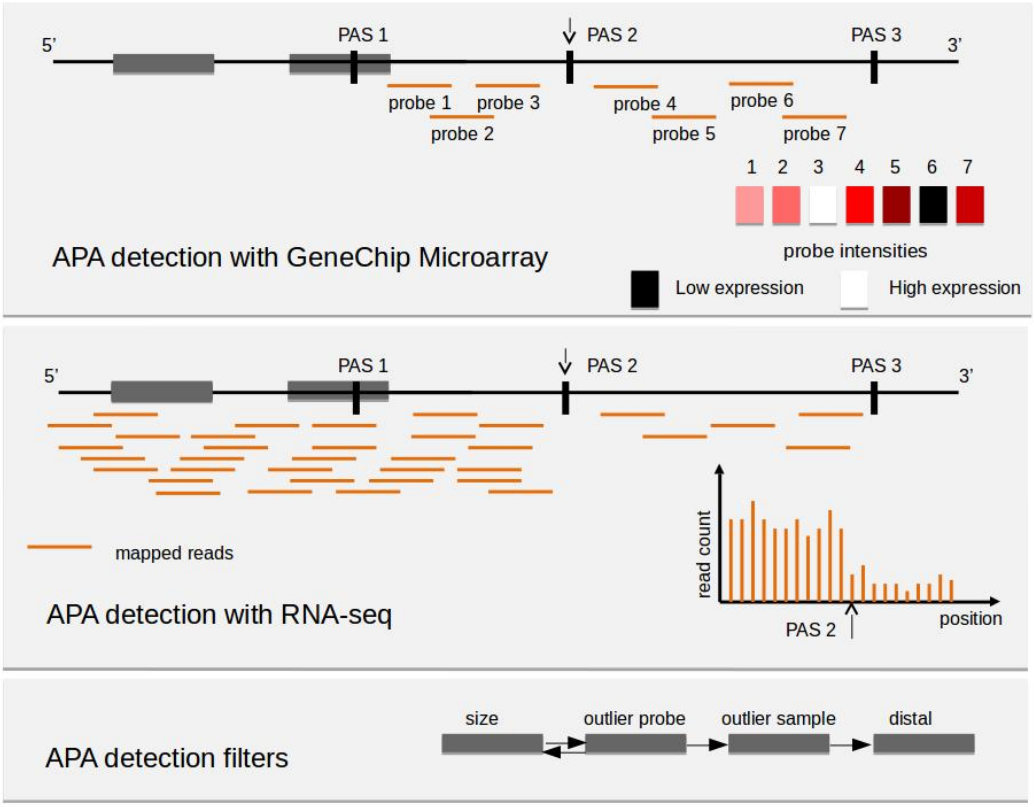


Figure 2.2: Illustration of the APADetect tool on two different platforms. PAS stands for the polyadenylation site and the PAS is used to divide probes into two groups. By determination of proximal/distal groups, four different filters are applied to discard outlier samples. Adapted from 'Alternative Polyadenylation: Another Foe in Cancer', by Erson-Bensan, A.E., & Can, T., 2016, Molecular cancer research : MCR, 14 6, 507-17

#### 2.2.1.1 Divisive Hierarchical Clustering

Divisive Hierarchical Clustering works by adopting a top-down approach. It starts with considering all data points in one big cluster, then splits the most heterogeneous clusters until all data points are in their own cluster. The basis of divisive clustering was introduced as the DIANA (DIvisive ANAlysis Clustering) algorithm [41]

#### 2.2.1.2 Agglomerative Hierarchical Clustering (AHC)

AHC uses a bottom-up approach, which is considering each data point as a separate cluster, then joining similar clusters in a greedy manner by merging them until all the similar clusters are merged together. If the number of clusters is provided, merging process is completed when system reaches desired number of clusters. Different methods are implemented to measure proximity of any two cluster. The most popular and commonly used ones are single linkage, complete linkage, weighted linkage, average linkage, and centroid linkage. Depending on the selected linkage type, distance is calculated between clusters and selected ones are combined into a single cluster. Processed clusters are removed from the set and a new comprised cluster is added.

In general, agglomerative procedure is more open to false decisions made in early stages which makes the divisive procedure more reliable. It is also known that, divisive clustering is powerful for finding large clusters while agglomerative clustering is powerful for identifying small clusters. However, if we compare clustering possibilities of the two approaches, we see that divisive clustering is more computationally expensive than the agglomerative one. While all possible mergers of two samples requires $n(n-1)/2$ combinations for agglomerative clustering, the splitting procedure of $n$ objects requires $2^{n-1}-1$ combinations for divisive clustering. Although DIANA reduces complexity by considering only a subset of all the divisions, it still requires much more computations (O($2^n$) ~ O($n^5$)) than the agglomerative approach. Exponentially growing combinations makes divisive clustering non-preferable and therefore, it is largely ignored in the literature.

#### 2.2.1.3 UPGMA

Unweighted pair group method with arithmetic mean, which is also known as average linkage, is a distance based method to construct a rooted phylogenetic tree [42]. It creates unweighted dendrograms, which means the distances from the root to terminal nodes are all equal in the tree. It requires the rates of evolution among distinct family lines to be almost equal. Because providing such evolutionary relationships is challenging and may still violate the criteria, this method is not commonly used.

We used open source SciPy [43], which is a python library used for scientific and technical computing. Cluster hierarchy package's average linkage method was used to build the tissue hierarchy dendrogram. The naive algorithm has time complexity O($n^3$). SciPy implementation is based on nearest neighbors chain which reduces complexity to O($n^2$) and uses O($n^2$) memory [44].

**UPGMA algorithm steps**

Let i and j be two distinct nodes, d(i,j) is the distance between nodes.
n is the number of nodes(OTU) to be processed.
Terminate only one cluster remains.

1. Find smallest value in matrix D for the pair of distinct nodes and create a new internal node by connecting currently processed terminal nodes to that node.

2. Calculate branch length from the pair members to the new internal node.

$$\delta(a,u) = \delta(b,u) = \frac{D(a,b)}{2}$$

3. Calculate the average distance from the remaining nodes in tree to the new internal node where A and B are clusters and x and y are nodes in clusters.

$$d(A,B) = \frac{1}{\mid A \mid \cdot \mid B \mid} \sum_{x \in A} \sum_{y \in B} d(x,y)$$

4. Update distance matrix D by combining processed terminal nodes in new internal node and by setting distances calculated in previous steps. Please note that matrix size is reduced by one column and row because of the joining process. Repeat algorithm starting with step 2.

#### 2.2.1.4 Neighbor Joining

Neighbor Joining is a distance based method firstly proposed by Naruya Saitou and Masatoshi Nei in 1987 [45]. It can be adopted to different needs but it is especially used for describing the evolutionary history trees based on DNA or protein sequence data. Construction of a tree requires divergence of species or sequences which is provided by a distance matrix. In this thesis, the tree was constructed based on how many

genes differ in terms of their SLR values between tissues. Neighbour joining assumes that genes with lower difference must be closer than other genes in the matrix. It iteratively clusters proteins, builds a new subtree and creates a new distance matrix from gene clusters. It makes a correction of the initial distance matrix and deals with the juxtaposition problem of long and short branches on the same phylogenetic tree. The principle of this method is to minimize the total branch length at each phase of clustering, beginning with a star like tree [45].

**NJ algorithm steps**

Let i and j be two distinct nodes, d(i,j) is the distance between nodes.
Let f and g be two distinct nodes, L(f,g) is the branch length between nodes.
n is the number of nodes(OTU) to be processed.
Terminate when n=3.

1. Create initial star tree.

2. Calculate modified distinct matrix Q from distance matrix D.

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

3. Find smallest value in matrix Q for the pair of distinct nodes and create a new internal node by connecting currently processed terminal nodes to that node, which is also connected to center of the star shape.

4. Calculate the distance from the pair members to the new internal node

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

$$\delta(g,u) = d(f,g) - \delta(f,u)$$

5. Calculate the distance from the remaining nodes in tree to the new internal node

$$d(u,k) = \frac{1}{2}\Big[d(f,k) + d(g,k) - d(f,g)\Big]$$

6. Update distance matrix D by combining processed terminal nodes in new internal node and setting distances calculated in previous steps. Please note that matrix size is reduced by one column and row because of joining process. Repeat algorithm starting with step 2.

According to the algorithm described above, the total time complexity and space complexity of running neighbor joining on a set having n objects are O($n^3$) and O($n^2$) respectively. It is possible to improve performance using different acceleration techniques and heuristics have been proposed to decrease complexity.

There are many libraries and tools available implementing neighbor joining. We used the open source scikit-bio which implements several structures and methods to work with biological data in Python. It provides well documented, high quality and up-to-date stability commitments with its API.

# CHAPTER 3

# SLR BASED HIERARCHICAL CLUSTERING

## 3.1 Brenda Tissue Ontology

Brenda Tissue Ontology (BTO) is a comprehensive structured encyclopedia which connects the biochemical and molecular biological enzyme data of BRENDA (BRaunschweig ENzyme DAtabase) with a hierarchical and standardized collection of tissue-specific terms [46, 47]. It is one of the first ontologies maintaining wide range of tissue related information, which is gathered from literature references and annotated manually by experts. It provides a tree-like subgraph, highlighting placement of tissue on the hierarchy when a tissue specific term is searched. Several information such as BTO id, definition, references, and enzyme source is also linked. Enzyme-organism specific tissue content is growing each day with the data integrated from external sources as well as laboratories studying cell line databases. Brenda Tissue Ontology, widely used in biochemical applications and scientific community, has currently more than 75000 tissue entries updated twice a year[46].

## 3.2 The Gene Expression Dataset and SLR values

GSE7307 (GEO, www.ncbi.nlm.nih.gov/geo/) [48] was used in this thesis, which has been submitted by R. Roth et al. in 2007. It contains gene expression data of normal and diseased human tissues obtained using the Affymetrix U133 plus 2.0 array. It has 677 processed samples, representing over 90 distinct tissues. Disease status, disease type, cell line and gender data are also provided for each sample. The sample IDs of the dataset are given in Appendix A.

### 3.3 Preparation

We used SLR values of genes to examine diverse genes and hierarchical similarity between tissues. The APADetect tool [37, 38], which detects and quantifies alternative polyadenylation (APA) events by analyzing raw intensities of the probes, was used to calculate the SLR values. In the tool, firstly, proximal and distal probe sets are constructed. Then, average probe intensities are identified for every proximal and distal probe and processed gene by gene. After calculating optimum intensities for each sample, proximal to distal ratio is computed and named as SLR (Short/Long Ratio).

We worked on a matrix, which has SLR values of 3304 genes across 677 tissue samples. Because of the problem with redundant assignment of gene symbols in the literature, entrez gene ids were assigned for each gene as well as symbols, which were used as primary identifiers for genes. An R object, org.Hs.egALIAS, from the bioconductor annotation packages [49] was used to provide mappings between common gene symbol identifiers and entrez gene identifiers. After an elimination of unknown named genes and the ones having no entrez gene identifiers, 2380 unique genes were listed in total. SLR values of each gene for each sample were processed according to its magnitude and quantity. Some genes appear more than once with the same probeset id but different polyA site id, which means that more than one alternative polyadenylation site exists for that gene. Therefore, more than one SLR value can be associated with one gene. Because larger/shorter SLR value implies proximal/distal 3' UTR isoform of the transcript, maximum shortening/lengthening output, furthest from 1, was taken into account while doing calculations. A difference threshold of 0.9 was selected to find genes having most differentiated 3' UTR isoforms. Genes having SLR values between 0-0.1 and 1.9- of the corresponding genes were identified for each 677 tissue samples. Consequently, we had 1425 unique genes with high proximal APA after SLR based elimination. Next, tissues were grouped according to sample types. In this dataset, there are six types, Activated, Control, Resting, Treated, Disease and Normal, each of which a tissue profile. Some tissues were experimented on for both Normal and Disease types. Since there were many samples belonging to the same tissue, gene SLR data of samples were merged into one unique tissue sample for each sample types. In total, we obtained the SLR data of 129 unique tissues for

the dataset GSE7307. The dataset before and after this process is given in Table 3.1.

| | Total Sample Count | Merged Sample Count | Gene Count |
|---|---|---|---|
| Activated | 9 | 6 | 570 |
| Control | 18 | 5 | 713 |
| Resting | 9 | 6 | 550 |
| Treated | 18 | 6 | 815 |
| Disease | 119 | 16 | 855 |
| Normal | 504 | 106 | 1325 |
| All | 677 | 145 | 1425 |

Table 3.1: Finalized tissue sample and gene count for each sample type

Additionally, to perform tissue-wide analysis of the human genes, we constructed the actual tissue hierarchy using the Brenda tissue ontology system. Tissue hierarchies were organized according to 2-5 level counts. Animal and other source classification categories were selected to inspect 129 source tissues in our dataset. Because some tissues are labeled in more than one parent tissue, we added additional entries to the hierarchy for those tissues. In the end, the hierarchy contains 21 main tissues with related sub tissues. Table 3.2 shows detailed information.

For each merged sample, genes were sorted according to SLR values and top 20,50,100 genes were selected respectively. The distance matrix, which contains pairwise distances between two sets, was constructed for each sample type to see dissimilarity of samples. The size of the matrix depends on the number of top genes that are analyzed, in our case 20x20, 50x50 and 100x100 matrices were analyzed for each sample class. Rows and columns were tissue samples in each class. Since diagonal entries give the difference between same tissue samples, all of them should be 0 and all the off-diagonal entries should be a positive number. The difference is calculated by counting genes which are not in the intersection of the two considered tissue samples.

$MatrixM = (x_{ij})$ with $1 \leq i, j \leq N$
$A = \{tissueSamples\}$
$tissueTypes = \{Disease, Normal, Activated, Control, Treated\}$
$N = \{20, 50, 100\}$

| Tissue Name | Sub-tissue Count | Tissue Name | Sub-tissue Count |
|---|---|---|---|
| Integument | 14 | Connective Tissues | 6 |
| Gland | 26 | Ganglion | 16 |
| Urogenital System | 27 | Viscus | 19 |
| Muscular System | 8 | Trunk | 9 |
| Cardiovascular System | 9 | Embryonic Structure | 4 |
| Nervous System | 45 | Softbody Part | 2 |
| Hematopoietic System | 7 | Sense Organ | 3 |
| Respiratory System | 3 | Adult Stem Cell | 2 |
| Skeletal System | 3 | Organism Form | 1 |
| Immune System | 4 | Other Source | 4 |
| Head | 46 | | |

Table 3.2: Main tissues with related sub tissue counts

$x_{ii}$ = 0 for all $1 \leq i \leq N, x_{ij} > 0$ if $i \neq j, x_{ij} = x_{ji}, x_{ij} \leq N$ if $i \neq j$
$x_{ij}$ = $Count(\{gene : gene \in A_i \wedge gene \notin A_j \cup gene \notin A_i \wedge gene \in A_j\})$

### 3.4 Dendrogram Creation and Structural Comparison

In this work, hierarchical clustering was used to create dendrogram structures of tissues in the human body, and the accuracy of the clustering is assessed using a standard hierarchy of tissues. The agglomerative clustering technique, which follows a bottom-up approach, was selected. We created dendrograms for each tissue type using neighbor joining and UPGMA techniques. The same workflow was repeated for top most 20, 50, and 100 differentiated genes.

Although there are more accurate character based methods using an optimization criterion such as parsimony, maximum likelihood or compatibility to analyze dendrograms [50, 51], we could not apply them because of our distance based dataset. Character based methods need phylogenetic characters like gene sequence alignment but we were dealing with data having numerical attributes.

Phylo.io was used to visualize hierarchical trees from their Newick [52, 53] tree format. It analyzes different levels and compares the internal structure of two inferred trees for the same set of OTUs[54]. Optimized version of Jaccard index is used to make metric comparison and identify the overlapped structure. Robinson-Foulds(RF)[55], euclidean and SPR metrics are adopted to measure distance between unrooted binary trees. RF distance is also known as the symmetric difference which searches for partitions existing only in one of the trees and calculates the sum of them. Euclidean distance simply measures line distance between nodes in trees being compared. Subtree pruning and regrafting (SPR) distance, which is NP-hard, counts the minimum number of moves needed to obtain one tree from another. Table 3.3 shows the result of tree distances for all, normal, and disease tissue types.

As expected, dendrograms, created using different clustering algorithms, have different structures if we calculate the distance of the overall schema. While NJ preserves actual branch rates, UPGMA assigns equal rates among branches resulting in a discrepancy. The distances get larger as the node count in the tree increases. In addition to that, trees built with top 100 highest SLR values are more differentiated from each other than the ones built with top 20/50 highest SLR values. This shows that, branch rate calculation affects the dataset with size 100 most and leads to data loss. However, the amount of differentiation is mostly negligible. If the internal structures are compared, we see that although there are differences among internal branches, hierarchically closer tissues are often located close on the tree for both UPGMA and NJ structures. The most remarkable tissue groups, whose sub tissues are clustered together were brain, ganglion, nervous system, and muscular systems. Respiratory system tissues except Lung were also located close. Deltoid Muscle, Skelatal Muscle, Thalamus Laterai Nuclei, Testes and Prostate tissues were the most differentiated tissues and located in farthermost branch. If disease and normal samples of same tissues are compared, we again see that tissue groups were neatly separated from each other. Also, we observe that normal samples of most tissues shared the same parent with their disease samples, which demonstrates high similarity. However; Skin, Accumbens and Putamen tissues show correlation with the samples of different tissues.

In order to inspect the effects of proximal APA tissue-wide, we studied various dendrograms with respect to the literature curated tissue hierarchy and presented accuracy

| | Robinson-Foulds | Euclidean | SPR | Node Count |
|---|---|---|---|---|
| Top20 SLR NJ All<br>Top20 SLR UPGMA All | 120/0.63 | 10.807 | 32 | 145 |
| Top50 SLR NJ All<br>Top50 SLR UPGMA All | 119/0.62 | 19.724 | 31 | 145 |
| Top100 SLR NJ All<br>Top100 SLR UPGMA All | 124/0.65 | 35.600 | 32 | 145 |
| Top20 SLR NJ Random<br>Top20 SLR UPGMA Random | 112/0.59 | 11.204 | 30 | 145 |
| Top50 SLR NJ Random<br>Top50 SLR UPGMA Random | 114/0.60 | 21.758 | 31 | 145 |
| Top100 SLR NJ Random<br>Top100 SLR UPGMA Random | 138/0.73 | 21.821 | 35 | 145 |
| Top20 SLR NJ Normal<br>Top20 SLR UPGMA Normal | 123/0.71 | 9.200 | 32 | 106 |
| Top50 SLR NJ Normal<br>Top50 SLR UPGMA Normal | 119/0.65 | 16.337 | 30 | 106 |
| Top100 SLR NJ Normal<br>Top100 SLR UPGMA Normal | 131/0.74 | 29.310 | 36 | 106 |
| Top20 SLR NJ Disease<br>Top20 SLR UPGMA Disease | 11/0.42 | 6.251 | 4 | 16 |
| Top50 SLR NJ Disease<br>Top50 SLR UPGMA Disease | 9/0.35 | 13.595 | 3 | 16 |
| Top100 SLR NJ Disease<br>Top100 SLR UPGMA Disease | 17/0.65 | 17.247 | 4 | 16 |

Table 3.3: Structural comparison of UPGMA and NJ trees

results for each experiment.

## 3.5 Precision Analysis of Dendrogram Structures

Precision of dendrograms was computed with Algorithm 1 and Algorithm 2 which were run on NJ and UPGMA trees respectively. Both algorithms have two phases executed sequentially and give ancestor based accuracy. They both identify near and distant nodes for each processed node and try to find out if detected near nodes are actually in the same family according to the actual biological tissue hierarchy. Because UPGMA and NJ have different branch rate approaches, they handle node identification and threshold selection separately. Algorithms were executed for each agglomerative tree which was created with top most 20, 50, and 100 differentiated genes for each type. Below, we provide results for four dataset types. The first dataset contains gene data for normal, i.e. healthy, tissues. The second one has data for diseased tissues. The third one contains gene data for diseased tissues and respective normal tissues, and the last one has all the tissues for activated, resting, control, and treated tissues as well as normal and diseased ones.

### 3.5.1 Ancestor Based Accuracy of NJ Based Trees

#### 3.5.1.1 Algorithm

Algorithm 1 presents ancestor based accuracies for dendrograms constructed with the Neighbor Joining method. In the first phase, an $N \times N$ matrix is generated having branch distances of related pair of tissues in the entries. $N$ is the number of unique tissues. The edges of trees represent the proximity or genetic distance between two tissues. A smaller distance indicates that tissues having similar proteins according to SLR values are more closely related. In the second phase, ancestors of the sample tissues are determined using the biological tissue hierarchy for each tissue pair. Ancestor data is a list and might contain multiple parent tissues for each sub tissue as a result of our structure. Median, average, and optimal distance values are calculated for each row and assigned as the threshold. These values divide row data into two halves, possibly near and possibly distant tissues. The ones having lower distance than the threshold are gathered and if currently processed tissue and its possibly near tissues have common ancestors, they are labeled as closely related. The contribution

amount to the final sum values depend highly on the size of the ancestor list.

**Algorithm 1** The NJ AncestorBasedAccuracy Algorithm

**Input:** Agglomerative Tree T, Unique Tissue List M,Biological Tissue Hierarchy H
**Output:** Accuracy for given tree

```
1:  for tissueCur in M do                                              ▷ Phase 1
2:      currentNode ← findNode(T,tissueCur)
3:      for tissueComp in M do
4:          comparisonNode ← findNode(T, tissueComp)
5:          d ← calculateDifference(currentNode,comparisonNode)
6:          add d to nodeDistancesMatrix
7:      end for
8:  end for
9:  for tissueCur in M do                                              ▷ Phase 2
10:     get node ancestors A for tissueCur(H)
11:     calculate threshold for tissueCur
12:     for tissueComp in M do
13:         get node ancestors B for tissueComp(H)
14:         intersect ← findIntersection(A,B)
15:         if nodeDistancesMatrix[tissueCur][tissueComp] < threshold then
16:             calculate SumN, countN using intersect
17:         end if
18:     end for
19:     calculate average near node count for tissueComp and add to nList
20: end for
21: return average(nList)
```

#### 3.5.1.2 Results

Results are provided in Table 3.4. Each pair of the columns represent accuracy and average near node count with respect to the selected threshold value. Median distance threshold gives middle value of the sorted distance values, and half of the tissues in the

network are accepted as near node. Because the threshold is so high, we got the lowest accuracy in this case. Average distance threshold is calculated by the arithmetic mean of the distance value list. In this case, possibly, the near node count is calculated for each tissue specifically, which divides near and distant nodes better. Best accuracy results were obtained with the average threshold. In the optimum threshold case, we determined the optimum nearest tissue count to be processed according to tree structure and the total tissue count for each network, and computed accuracy results accordingly. This method gave lower results than the average threshold because, in some cases, distant nodes behaved like near nodes and were included in the calculation. Likewise, near nodes were neglected because of the limited threshold.

| | Median Distance | Avg NNC | Average Distance | Avg NNC | Optimum Distance | Avg NNC |
|---|---|---|---|---|---|---|
| All Top20 | 0.3905 | 71 | 0.60 | 24 | 0.53 | 15 |
| All Top50 | 0.3616 | 71 | 0.74 | 9 | 0.53 | 15 |
| All Top100 | 0.3879 | 71 | 0.81 | 7 | 0.53 | 15 |
| Normal Top20 | 0.4051 | 52 | 0.64 | 15.2 | 0.53 | 10 |
| Normal Top50 | 0.3884 | 52 | 0.85 | 5.89 | 0.55 | 10 |
| Normal Top100 | 0.4150 | 52 | 0.89 | 6.18 | 0.57 | 10 |
| Disease Top20 | 0.7757 | 7 | 0.93 | 5.63 | 0.87 | 4 |
| Disease Top50 | 0.7361 | 7 | 0.94 | 5.44 | 0.81 | 4 |
| Disease Top100 | 0.7757 | 7 | 1 | 4.38 | 0.82 | 4 |
| Normal-Disease Top20 | 0.5096 | 14 | 0.94 | 10.8 | 0.91 | 7 |
| Normal-Disease Top50 | 0.5043 | 14 | 0.93 | 9.29 | 0.85 | 7 |
| Normal-Disease Top100 | 0.4896 | 14 | 0.98 | 7.84 | 0.87 | 7 |

Table 3.4: Summary of ancestor based accuracies of NJ dendrograms

### 3.5.2 Ancestor Based Accuracy of UPGMA Based Trees

#### 3.5.2.1 Algorithm

Algorithm 2 presents a method for computing ancestor based accuracy for dendrograms constructed with the UPGMA method. In the first phase, the maximum distance threshold list which keeps maximum inconsistency coefficient for each non-singleton cluster is created, and flat partitions are formed by given linkage matrix and distance based threshold. The matrix provides encoded hierarchical clustering data and the threshold determines the cluster size. If the threshold is small, just the closest neighbours form a cluster, which increases total cluster count in the network. On the other hand, if the threshold is very high, many distant nodes can be added to the cluster, which decreases the total cluster count. Therefore, it is vital to choose the optimum threshold during clustering. We analyzed maximum, minimum, average and median of maximum distance threshold list to find an optimal threshold. We also provided additional threshold values according to tree structure and total tissue count for each network. Finalizing cluster sets, we assigned each tissue to a proper cluster. In the second phase, ancestors of the sample tissues are determined using biological tissue hierarchy for each tissue pair. Ancestor data is a list and might contain multiple parent tissues for each sub tissue as a result of our structure. Tissues belonging to same cluster and having common ancestors are labelled as closely related and ancestor based accuracy for each tissue is calculated accordingly. The average of calculated accuracies give the network accuracy.

#### 3.5.2.2 Results

Results are provided in Table 3.5. Each pair of the columns represent accuracy and average cluster count with respect to the selected threshold value. Among all, average and optimum thresholds were chosen to represent the accuracy of a network. Average threshold gives middle value of the sorted maximum inconsistency coefficients and clusters were formed to include 4-7 tissues each. For the optimum case, cluster count was determined based on tree structure and total tissue count for each network, and accuracy were calculated accordingly. We decreased cluster count for

**Algorithm 2** The UPGMA AncestorBasedAccuracy Algorithm

**Input:** Tissue Distance Matrix Z, Unique Tissue List M,Biological Tissue Hierarchy H, Threshold T

**Output:** Accuracy for given tree

```
1: findMaximumDistanceForEachCluster(Z)
2: clusterIds ← getClusterIds(Z,T)
3: for clusterId in 1, . . . , max(clusterIds) + 1 do                ▷ Phase 1
4:     find tissues with same clusterId and insert into clusterSet[clusterId]
5: end for
6: for tissueCur in M do                                             ▷ Phase 2
7:     get node ancestors A for tissueCur(H)
8:     for tissueComp in M do
9:         get node ancestors B for tissueComp[(H)
10:        intersect ← findIntersection(A,B)
11:        if tissueCur and tissueComp in same clusterSet then
12:            calculate SumN, countN using intersect
13:        end if
14:    end for
15:    calculate average near node count for tissueComp and add to nList
16: end for
17: return average(nList)
```

both normal/healthy tissue dataset and all tissue datasets to enhance divergence. Clusters had 9-10 tissues for this case. We slightly increased cluster count for diseased tissue dataset and diseased-normal together tissue datasets to prevent distant tissues from being member of the same cluster, and compared results with the ones obtained from NJ experiments. Clusters had 4-5 tissues for this case. While average threshold based analysis gave higher accuracy results for healthy and all tissue datasets, optimum threshold based analysis gave better results for diseased and normal-diseased datasets. Because accuracy depends highly on cluster size, this behavior was expected.

| | Average Threshold | Cluster Count | Optimum Threshold | Cluster Count |
|---|---|---|---|---|
| All Top20 | 0.64 | 29 | 0.56 | 15 |
| All Top50 | 0.67 | 29 | 0.54 | 15 |
| All Top100 | 0.69 | 28 | 0.56 | 15 |
| Normal Top20 | 0.63 | 22 | 0.54 | 10 |
| Normal Top50 | 0.67 | 25 | 0.52 | 10 |
| Normal Top100 | 0.69 | 22 | 0.57 | 10 |
| Disease Top20 | 0.91 | 4 | 0.91 | 4 |
| Disease Top50 | 0.83 | 3 | 0.90 | 4 |
| Disease Top100 | 0.83 | 3 | 0.83 | 4 |
| Normal-Disease Top20 | 0.88 | 4 | 0.95 | 7 |
| Normal-Disease Top50 | 0.88 | 5 | 0.95 | 7 |
| Normal-Disease Top100 | 0.74 | 3 | 1 | 7 |

Table 3.5: Summary of ancestor based accuracies of UPGMA dendrograms

Altogether, if we compare UPGMA results of optimum threshold and NJ results of optimum distance, we observe that UPGMA gives slightly better accuracy, but overall, results were close to each other. Applying the optimum threshold, we obtained similarity more than %50 for normal and all datasets and more than %90 for diseased and normal-diseased datasets. Lowering cluster size or near neighbor count improved accuracy even more. Therefore, we can say that hierarchically similar tissues can be

determined by just looking at the gene data showing higher SLR values. There might be several underlying reasons behind accuracy loss. First of all, some tissues tend to exhibit more APA events than other hierarchically closer tissues, which yields divergence. Secondly, different samples of same tissues such as diseased, activated, healthy etc. may show different APA behaviour from each other, but show high APA similarity with the hierarchically distant ones. To be able to detect these cases, we have to deep dive into the data.

Best results were obtained for the dataset having gene data of diseased tissues and respective healthy tissues. If we analyze the dendrograms, we clearly see that for vast majority of tissues, diseased-healthy samples are highly correlated. It means that the genes showing shorter 3' UTR isoform are common in diseased and healthy samples of same tissues. Skin, putamen and accumbens tissues make an exception, and their samples were located far away from their related pair in the dendrogram and can be focused more in future work.

# CHAPTER 4

# SLR BASED FEATURE LEARNING ANALYSIS

In this chapter, we aim to perform organism-wide analysis of proximal APA events using a multi-layer tissue hierarchy and a gene interaction network. Although we have created dendrograms based on tissue-tissue distances and examined hierarchically closer tissues for different cases, we did not consider gene interaction networks across tissues before. By extracting SLR based features for each gene, gene-tissue correlation and tissue behaviour in case of proximal APA can be studied.

## 4.1 Feature Engineering and Related Work

Feature engineering, transforming raw data into features, has been popular among researchers who study variety of domains ranging from social networks to biological networks. It is one of the most important phases of a data mining workflow, aiming to build an efficient model and to work on classification and prediction of data, because meaningful features and attributes are key factors to improve performance and accuracy.

Protein-protein interaction networks are fundamental as data sources to learn neural embedding based low-dimensional space of features, and they are studied extensively in life sciences. They are used for network alignments, predicting functional labels of proteins, discovering novel interactions between genes, drug discovery, and more [56, 57, 58, 59, 60]. A significant body of genomics research focuses on down-stream learning tasks on specific tissues and novel genes, but relations between tissues and genes are largely ignored. They assume that same proteins in distinct tissues work similar and cellular function is constant across organs and tissues [56, 61, 62]. How-

ever, gene behaviors and functionality may be specific to a particular tissue and tissue hierarchy should be taken into account while extracting rich feature representations for proteins [26]. Among many state-of-the-art approaches, we focused on an algorithm for hierarchy aware multi-layer networks [26]. The algorithm builds on prior unsupervised feature learning methods working on neural architectures [63, 64] and uses random walks and continuous bag of words model to learn features. Random walk efficiently explores node neighbors in a d-dimensional feature space [63]. CBOW produces a distributed representation of nodes and predicts the current node by looking at the surrounding nodes in a window having adjustable size [64]. Ohmnet constructs a tissue hierarchy and layers, representing the PPI network for each distinct tissue in the hierarchy. A biological system is represented as a bunch of proteins and interactions in a PPI network considering gene names as nodes and physical/-functional interactions as edges [65]. The strength of weighted edges are processed to discover functional units[66]. Processing on the multi layer network, neighborhoods for nodes are generated using a random walk approach[63]. Walk length and number of walks parameters are used to determine a sampling strategy. Biased parameters $p$ and $q$ are used to guide the walk by regulating probabilities, which affect neighborhood strategy[63]. After constructing network neighborhoods, initial feature set is generated for each node in every layer by evaluating node-node interactions. This set is then updated iteratively such that gene, i.e. protein, nodes having similar set of neighbors in hierarchically close layers are embedded closely together[26]. In the end, the system learns a d-dimensional feature vector for every protein node in every tissue. The authors of Ohmnet use feature vectors for protein functionality prediction and, for cellular function transferring to unannotated tissues[26]. However, in this thesis, we are only interested in feature embeddings and protein similarity to infer whether protein nodes in a proximal APA cluster are parallel with initial tissue hierarchy. Therefore, we did not perform any classification methods.

## 4.2 Dataset

To perform protein feature extraction, we need a PPI network for each tissue as well as a hierarchical representation of the tissues. We used the supplementary data pro-

vided by Zitnik et al.[26] In their work, the Human PPI network was retrieved from Menche et al.[67] who merged lots of databases having different types of physical interactions. The resulting network gives an interactome having 13,460 proteins connected by 141,296 physical interactions[67]. Interaction type based protein and PPI count is given in table 4.1.

| Interaction Type | Protein Count | Interaction Count |
|---|---|---|
| Regulatory | 564 | 1335 |
| Binary | 8120 | 28653 |
| Literature | 11798 | 88349 |
| Metabolic enzyme-coupled | 921 | 5325 |
| Protein complexes | 2069 | 31276 |
| Kinase network | 1843 | 6066 |
| Signaling | 6339 | 32706 |

Table 4.1: Obtained PPI interaction count from well known studies and databases [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].

Zitnik et al.[26] then expanded the network by combining data sources from three other significant databases[68, 69, 70]. The final global network contained 21,557 proteins interconnected by 342,353 physical interactions[26]. They handled gene to tissue mapping using gene pair annotation data by the method provided by Greene et al. [71]. Greene et al. [71] determined tissue specific proteins by examining gene pair relationship with the tissue such that each gene pair is labeled as co-expressed if at least one of the gene in the pair is tissue specific. Using co-expressed genes and related tissue data, Zitnik et al.[26] presented 107 tissue specific human PPI subnetworks.

## 4.3 Preparation

We selected 48 tissues among 107 to analyze proteins having SLR values higher than a specified threshold value. They have 5000 to 57000 interactions. We used Brenda Tissue Ontology [46, 47] to build a tissue hierarchy. The resulting network has 27

internal tissue nodes and 48 leaf tissue nodes. Several training experiments with different gene datasets were executed. During training, biased parameters $p$ and $q$ were both set to 1. Walk length, number of walk and window size parameters were kept as 10, 5, and 10 respectively. In the first experiment, all proteins and PPI interactions were used to train the system. Running the Ohmnet model, we obtained 117,069 feature vectors with 128 features each. In the second experiment, PPI interaction network of tissues were limited such that we eliminated interactions if neither of the proteins showed high proximal APA. Same gene dataset with high proximal APA, provided in Table 3.1, was used to label proteins in the PPI network. Once elimination was done, tissue networks had 4000 to 26000 interactions. After training, we obtained 104,414 feature vectors with the same amount of features. Because we are interested in proteins with high 3'UTR shortening, we only kept corresponding feature sets, resulting in a total of 11,676 protein feature vectors for 48 unique tissues.

### 4.4 Methods and Experiments

To observe the pattern of protein and tissue similarity, we performed principal component analysis and t-distributed stochastic neighbor embedding methods [72, 73]. Both techniques model high dimensional data to low dimensional space by preserving the neighborhood structure and variance as much as possible. They reduce the complexity of data to accelerate the learning algorithm and visualize high-dimensional datasets. Although PCA gives a mathematical solution, t-SNE uses a probabilistic approach while embedding the feature set into lower dimensions. Because probabilistic approach may require heavy computational power if the number of dimensions is very high, using t-SNE after applying other techniques might be preferred. In our experiments, we dealt with 128-dimensional feature subspace, therefore; we did not require sequential execution of dimensionality reduction methods using t-SNE.

The first experiment gave us all protein embeddings by processing PPI interactions in each layer and similar proteins in hierarchically closer layers. We used output of this experiment as a guidance for hierarchically closer tissues in the normal condition. In the second experiment, we obtained protein embeddings with high proximal APA. Then, we compared the resulting networks of the two experiments to analyze tissue

behaviours in case of proximal APA presence. We also investigated whether genes having higher SLR values showed irregularity or if they clustered together.

For both experiments, PCA and t-SNE methods were executed and the resulting charts were analyzed [74]. PCA was performed on training data with two components and an auto solver, which determines the most efficient linear dimensionality reduction method using a training matrix and the component count. TSNE is run with several parameters. The number of dimensions of the embedded space was chosen as two. Because we were processing a relatively large dataset, perplexity and maximum iteration number for the optimization were set to 50 and 3000, respectively. The Euclidean metric was used to compute the distance between items in a feature array. Finally, the random state was given as 32 to the seed cost function.

When the data is crowded, it might be very difficult to visualize similarity between objects. Therefore, it is a common practice to color objects. Because tissues in our work are represented in multiple hierarchies, first we reduced the hierarchy level to two such that all proteins connect to mid-level tissues and all mid-level tissues connect to an anterior division of the animal body according to Brenda Tissue Ontology. Through 2-level hierarchy of tissues, colorization was performed. Figure 4.1 shows 1-level and 2-level color mapping.

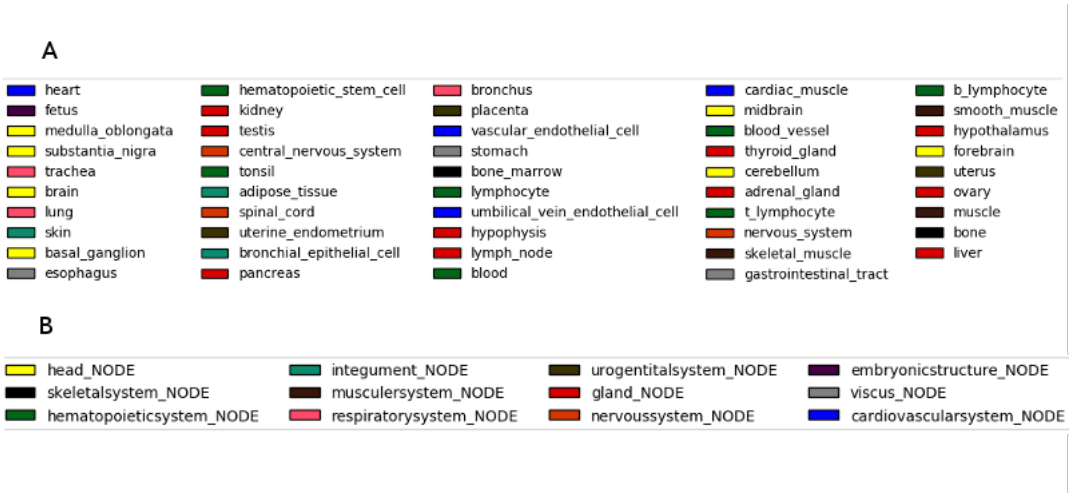


Figure 4.1: 2-level color mapping with distinct monochromatic colors(A). 1-level color mapping with distinct main colors(B).

Each anterior division of the animal body was assigned a main color and related connected tissues were assigned distinct monochromatic colors. During 2D visualization, proteins were also colored according to color information of mid-level tissues they are connected to.

### 4.4.1 Feature Embeddings of Proteins with High SLR Values

PCA and t-SNE methods were first run on the limited set containing proteins with higher SLR values, i.e. higher proximal APA. 2-level tissue hierarchy was used to color and distinguish proteins in same mid-level tissue group as well as in the main tissue group. While in Figure 4.2 proteins were mapped using PCA, in Figure 4.3 proteins were mapped using T-SNE.

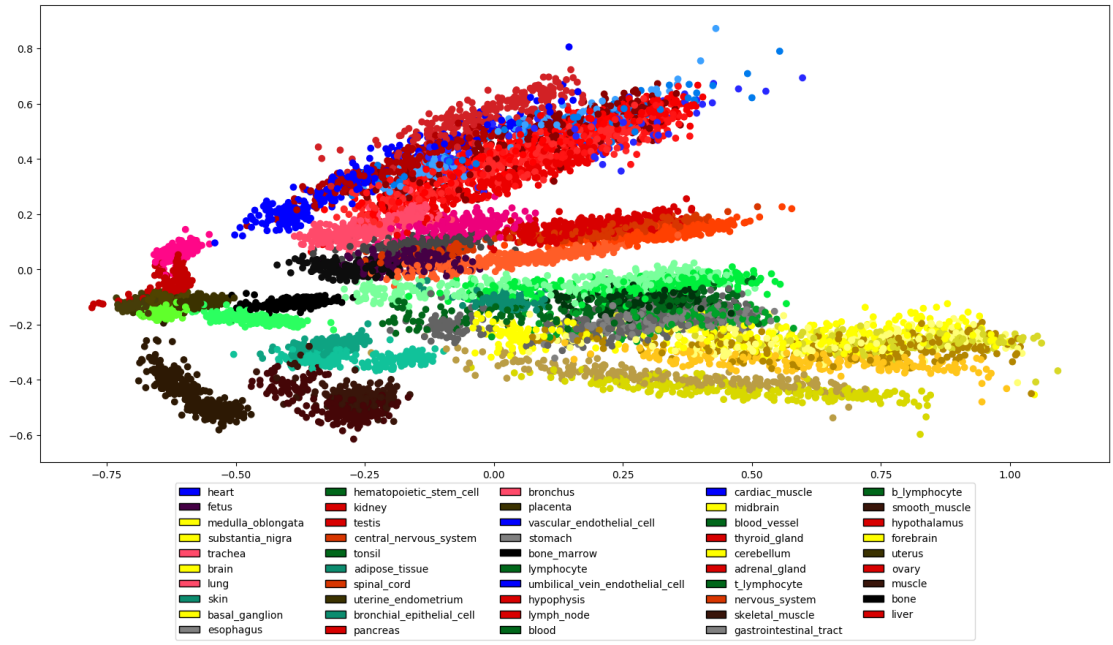


Figure 4.2: PCA representation of limited protein embeddings on 2-levels tissue hierarchy.

According to the figures, we clearly see that proteins in the same tissue are closely distributed and clustered together. Tissues connected to same anterior divisions of animal body are also mapped into closer regions than others. In the T-SNE chart, tissues form separate, well-defined circular shapes. In the PCA chart, tissues are mapped

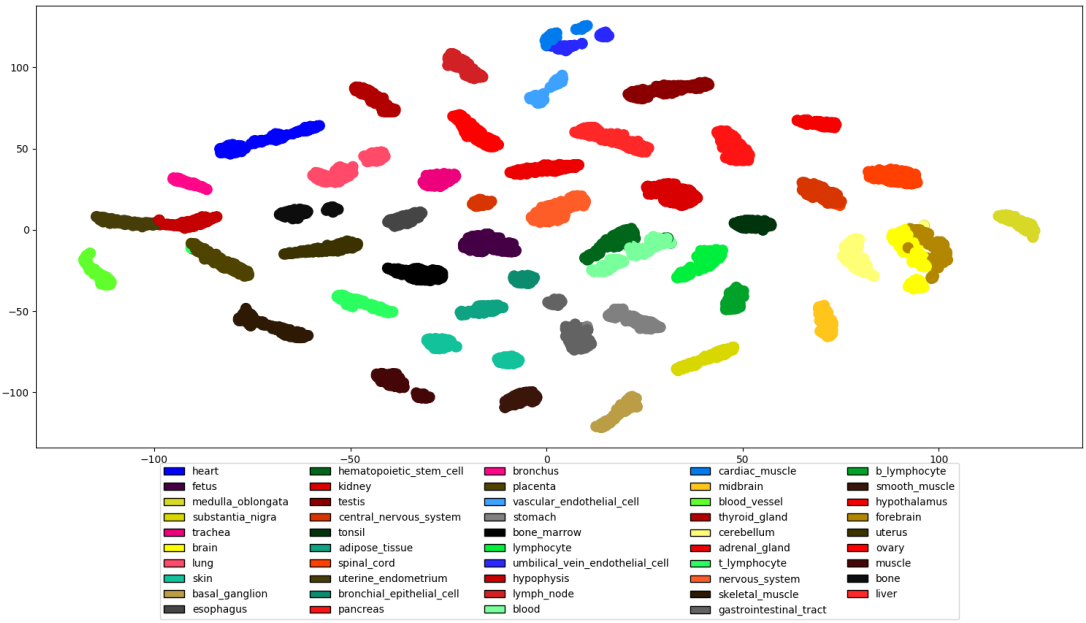


Figure 4.3: T-SNE representation of limited protein embeddings on 2-levels tissue hierarchy.

more linearly but still cluster closely. Because at some points, proteins in different tissues are mapped to the same regions, clear detection of diverged tissues is not possible with the PCA approach. However, clustering is more precise in T-SNE and this allowed us to make inferences. Heart tissue is separated from other cardiovascular system tissues and hypophysis is mapped together with bronchus, blood vessel, and urogenital system tissues rather than gland tissues. Blood vessel and b_lymphocyte are also not clustered together with other proteins in the hematopoietic system and show similarity with the nodes in the urogenital system and bone marrow.

1-level tissue hierarchy was used to color and distinguish proteins in the same main tissue group. Protein mappings stayed the same but coloring became simpler such that mid-level tissues were assigned a main color according to anterior division of animal body they are connected to. While Figure 3.4 shows proteins mapped using PCA, Figure 3.5 shows proteins mapped using T-SNE. In those charts, we cannot observe mid-level tissue neighborhood but we can infer about the similarity of main tissues.

Head and muscular system tissue groups seem more isolated than other tissues. While

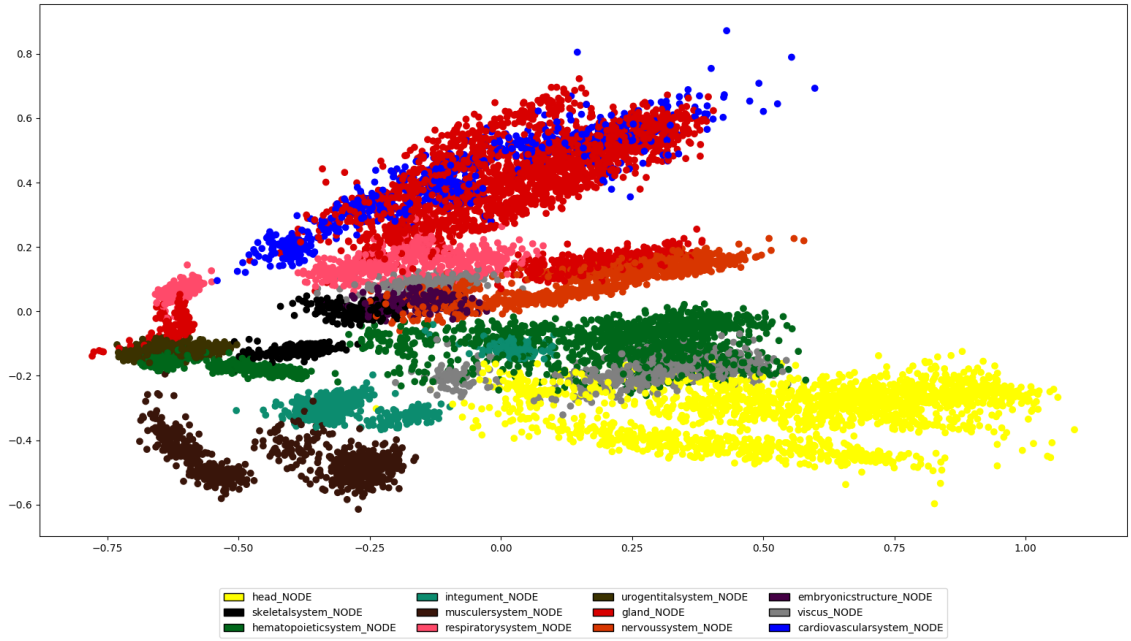


Figure 4.4: PCA representation of limited protein embeddings on 1-level tissue hierarchy.

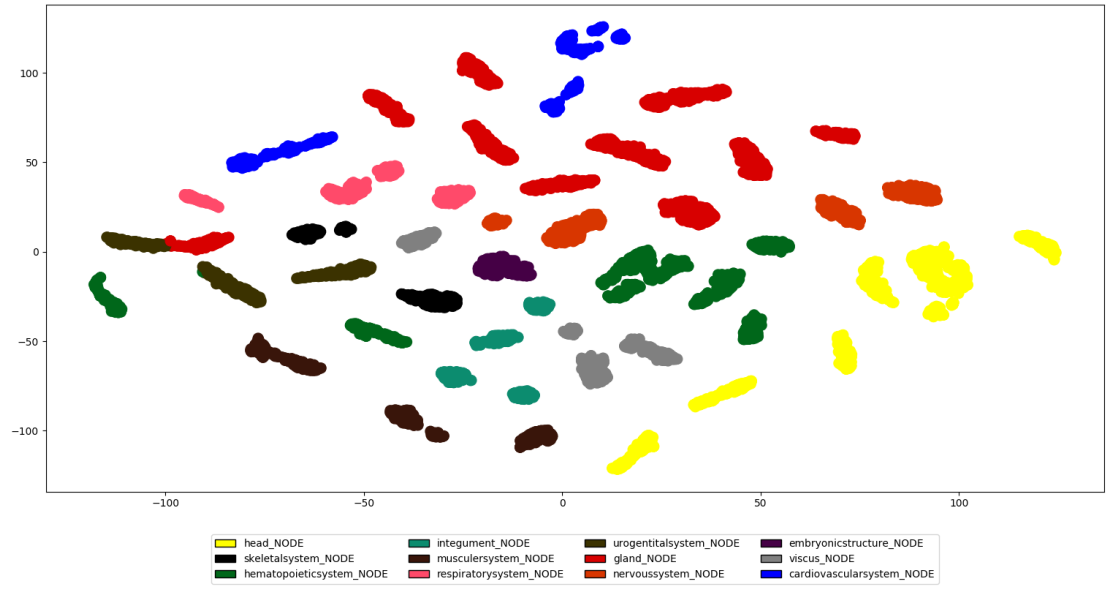


Figure 4.5: T-SNE representation of limited protein embeddings on 1-level tissue hierarchy.

proteins in the head group indicate partial closeness with the viscus group, the muscular group is closest to the integument group. Other tissue groups partly intersect with

each other, which means they might share mutual features. Among them, gland and cardiovascular tissues seem more similar to each other since they have many proteins placed together on PCA's 2-dimensional coordinate system. Their circular shaped clusters are both located at the top-center of the T-SNE chart. Part of the gland nodes are also mapped distinctly or share similar points with respiratory, nervous, and urogenital systems. While one big cluster of the viscus tissue showed similarity with the head, integument, and hematopoietic systems, other smaller clusters appeared to be surrounded by respiratory, skeletal, urogenital, and embryonic structure systems.

### 4.4.2 Feature Embeddings of All Proteins

In order to examine whether genes having shortened 3'UTR show irregularity or better clustering, protein embeddings of the whole gene set were also visualized with the T-SNE and PCA approaches using same colorization technique described before. Since, the dataset is much more crowded than before, it produces high density but difficult to read charts. Therefore, we just used 1-level tissue hierarchy to color and distinguish proteins in same main tissue groups. Figures 4.6 and 4.7 present all the protein embeddings mapped using PCA and T-SNE, respectively.

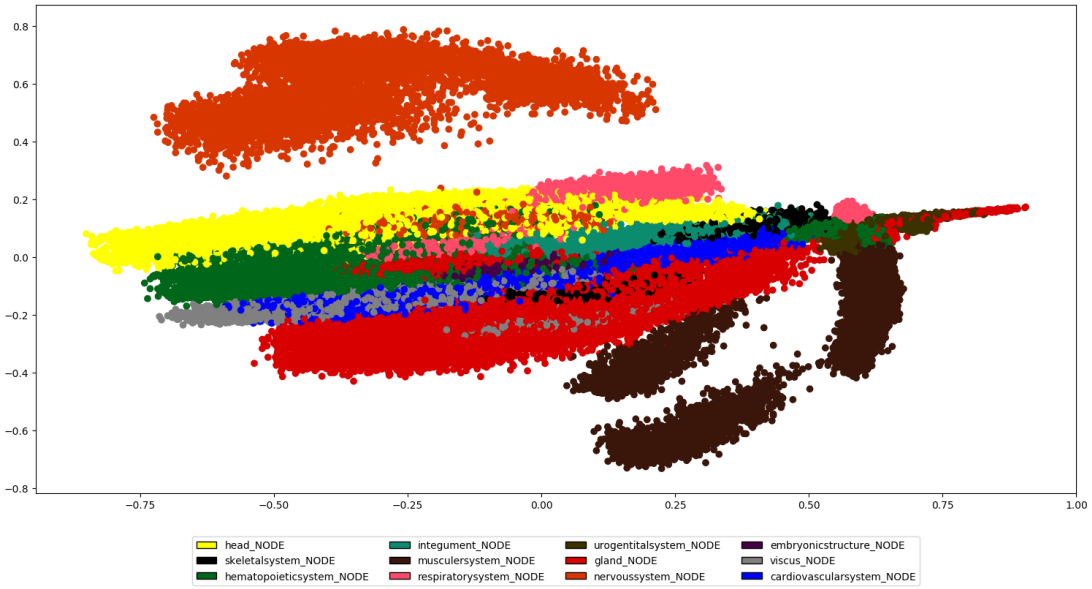


Figure 4.6: PCA representation of all protein embeddings on 1-level tissue hierarchy.

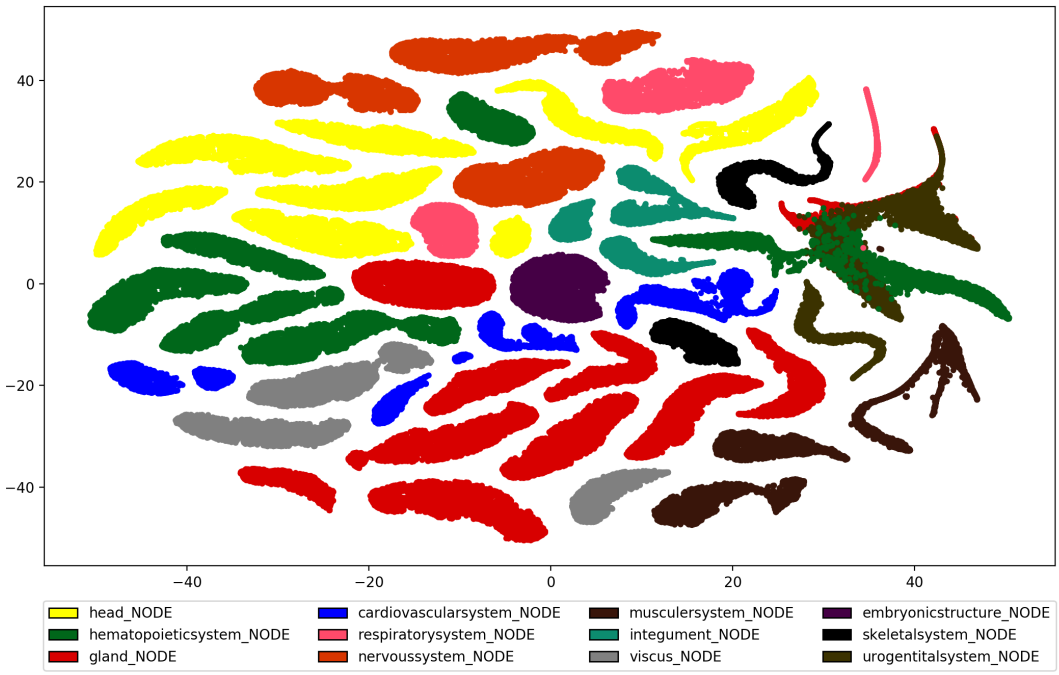


Figure 4.7: T-SNE representation of all protein embeddings on 1-level tissue hierarchy.

According to the figures, we still see closely distributed and clustered mid-level tissues but, among the main tissue groups, some mid-level tissues show divergence. Although all mid-level tissues in embryonic structure, integument, urogenital, and muscular systems cluster in nearby places, some remaining mid-level tissues seem to cluster farther than their tissue group.

As a result of our experiment on whole protein data embeddings with tissue hierarchy, we were able to see that the head tissue was not separated from other tissue groups anymore, as it showed similarity with respiratory, hematopoietic, integument, and partly skeletal and nervous systems. The muscular system is still isolated but maps slightly closer to the gland group. Gland and cardiovascular systems are again similar to each other but, this time, they are placed in a more central position and partially intersect other tissues. One big cluster group of gland is partially close to viscus and skeletal systems. On the other hand, a small cluster of gland is placed together with respiratory, hematopoietic, integument and urogenital systems. Viscus groups also converge in two different places. While one cluster shows similarity with cardiovascular and hematopoietic systems, other cluster is placed near the muscular

system. Tissues in the respiratory system show different mapping behaviour, which yields multiple neighbours mainly including head, gland, nervous, hematopoietic and skeletal systems.

When we compare the experiment done by using embeddings of proteins having high SLR values with the experiment that have all the protein embeddings, we clearly see that mid-level tissue clustering is stable. In both experiments, proteins showing similar interactions according to a particular tissue and hierarchy are clustered together, and tissue regions of the distinct colors are well-separated from each other. Cluster divergence between main-level tissues was reduced with SLR proteins and all tissues in respiratory, head and nervous system were mapped together. Accepting similar main tissues as hierarchically closer in experiment one, we can argue that training the system with only interactions of genes having higher proximal APA changes the mapping of some tissue groups. Moreover, different tissue groups having more uniquely co-expressed proteins with shortening events may show more similarity than other hierarchically closer tissue groups. The unexpected similarity between these hierarchically unrelated tissues may reflect common functional and regulatory roles or disease trends resulting from proximal APA which can be further investigated by wet-lab experiments.

Following inferences can be made for individual main tissue groups from the experiment with SLR protein embeddings compared to the results of the experiment using all the protein embeddings. Although gland and nervous systems are located far from each other in when all embeddings are used, they showed high similarity and mapped together in SLR only embeddings. The head tissue is separated from the respiratory and nervous system, and is associated with the viscus and shows similarity with the hematopoietic system. The viscus is distant from the gland and cardiovascular systems, but it is closer to the head and the integument, and retains the similarity with the hematopoietic system. The cardiovascular system is mapped in a relatively central position and has many neighbor tissues surrounding it when all embeddings are used. However, it is quite different from these tissues (except for the gland) when SLR only embedding is used. Moreover, the respiratory system is located similarly with the gland and cardiovascular systems and it is different from the head and hematopoietic systems. The muscular system is clustered separately in both cases. Although viscus

and gland groups are its immediate neighbours when all embeddings are used, they are replaced with the integument in the case of SLR only embeddings. The nervous system is mapped mostly isolated from other tissue groups but gets sporadically closer to the head, hematopoietic, integument, and respiratory systems when all embeddings are used. However, it dissociates with the head and integument completely, stays as a neighbor of the hematopoietic and respiratory systems and gets closer to the gland, viscus and embryonic structure when SLR embeddings are used. This shows that the utilization of APA events in modeling tissue-level dynamics of genes expression is crucial and complementary to the existing transcriptomic information.

# CHAPTER 5

# GENE ANALYSIS

Studies report that, proximal APA may lead to rapid proliferation of various cells and can be associated with activation of proto-oncogenes in cancer cells [11, 12, 37]. It has already left marks for breast and lung cancers as well as some of heart, endocrine and hematology diseases [14, 15, 16]. Therefore, examining gene behaviours across different tissues for different conditions may help scientists to work on medical prognosis, diagnostics, and treatment options of the APA related diseases and characterize the diversity of polyadenylation. This guides us to one of the motivations of our work. Analyzing genes that show significant 3' UTR shortening for disease and corresponding normal samples, we intended to detect mutual/differentiated genes across different tissues and reveal common biological pathways if there are any.

## 5.1 Preparation

In this work, 1425 different genes having SLR values higher than 1.9 were studied across 677 different samples. The dataset contains gene-SLR values for 129 unique tissues. Sampling was made for six different conditions, which are activated, control, diseased, normal, resting, and treated conditions. For this part of the experiment, we only focused on diseased and normal samples. Among all, just 15 tissues have both diseased and normal samples and inferences were made accordingly. Table 5.1 shows sample counts for normal and diseased states. Each sample stores the list of genes activated on that tissue with the corresponding SLR values. GSM ids of samples are given in Appendix A and NCBI GEO [48] can be examined for further information.

| | Normal | Diseased |
|---|---|---|
| Accumbens | 14 | 4 |
| Breast | 2 | 5 |
| Caudate | 4 | 2 |
| Gloubus Pallidum External | 2 | 4 |
| Gloubus Pallidum Internal | 3 | 4 |
| Myometrium | 22 | 32 |
| Ovary | 5 | 18 |
| Prostate | 7 | 18 |
| Putamen | 13 | 5 |
| Skin | 7 | 3 |
| Substantia Nigra Pars Compacta | 5 | 3 |
| Substantia Nigra Reticulata | 4 | 3 |
| Synovial Membrane | 6 | 5 |
| Thalamus Lateral Nuclei | 2 | 2 |
| Thalamus Subthalamic Nucleus | 2 | 2 |

Table 5.1: Normal and diseased sample counts for tissues used in this part of the study.

## 5.2 Experiments and Results

Since multiple samples may belong to the same tissue, we merged samples of each unique tissue into two different samples, one for the diseased and the other for the normal state in order to conduct accurate experiments. SLR-gene analysis was made considering two parameters. The first parameter is the SLR value of the gene and the second parameter is the number of occurrences of genes showing 3' UTR shortening. Gene behaviour across diseased, normal, and entire samples of tissues were analyzed and genes meeting the criteria were determined. Human Protein Atlas [75, 75, 76] was used to examine tissue and pathology atlas of each notable gene and to identify genes as potential bio-markers in various diseases.

### 5.2.1 Analysis Based on SLR Value

In this case, all provided diseased and normal samples of the same tissue were merged separately according to SLR values of genes. Because one gene may be represented with different SLR values in multiple samples, the highest SLR value obtained for that gene was selected during the merge. Resulting unique tissue samples, which contain sorted genes by SLR values, were generated for diseased and normal samples. Processing all diseased and normal samples, mutually occurring and distinct genes were detected for all tissues. Appendix B.1 give the corresponding gene counts for all the tissues.

By analyzing each tissue separately, we observed that some genes have higher SLR values for both diseased and normal samples. These genes may be more prune to the APA shortening event in any condition. However, we mainly aim for genes showing different characteristics at the diseased state. Therefore, we initially focused on mutually occurring genes whose SLR values on the diseased tissue is higher than the normal tissue. Due to the dense gene dataset, we selected the most differentiated ones. Appendix B.2 gives the names of these genes. It is noteworthy that some genes were found in the output of multiple tissues and CLU, IGL@, TNFSF10, DAZAP2, CALM1, MMP7, STMN1 and COL16A1 are the most common of them. Among these genes, CLU, TNFS10, CALM1, MMP7 and STMN1 are classified as cancer related genes in the Human Protein Atlas [75, 75, 76]. All but COL16A1 is associated with various diseases.

We then analyzed distinct genes whose SLR values are high on either diseased tissues or normal tissues. We excluded genes having SLR values lower than 3 for simplicity. As we anticipated, behaviour of tissues varies. Although genes in Accumbens, Putamen and Skin have high SLR values for normal tissues, genes in Myometrium, Ovary and Breast have high SLR values for diseased tissues. After studying tissues on the tissue hierarchy and the dendrograms, which were created before, we saw that Accumbens and Putamen are nervous system's sub-tissues. They are located very close, sharing the same parent, on Normal-Disease dendrogram. Skin is integument's sub-tissue and located slightly farther from Accumbens and Putamen. Myometrium, Ovary and Breast are both reproductive system tissues and located in the same sub-

tree on Normal-Disease dendrogram. Distinct gene counts, when threshold was selected three and four respectively, for all experimented tissues are given in Table 5.2.

If only distinct genes in diseased samples are studied, Breast, Myometrium, Ovary and Synovial Membrane tissues are the only ones which have gene(s) with SLR values higher than four. Top 10 distinct genes meeting the criteria for each tissues are given in Table 5.3. In Human Protein Atlas, CD44, NAMPT, MAF, PTGS1 and TET2 are categorized as cancer related genes and HNRNPA1, TET2, C22ORF25, USP9X, COL4A1, MAF, BGN and SFTPB are categorized as disease related genes[75, 75, 76].

| | Distinct Gene Count | | | |
|---|---|---|---|---|
| | Normal | | Diseased | |
| | t=3 | t=4 | t=3 | t=4 |
| Accumbens | 30 | 6 | - | - |
| Breast | 1 | - | 8 | 1 |
| Caudate | - | - | - | - |
| Gloubus Pallidum External | - | - | 1 | - |
| Gloubus Pallidum Internal | 1 | - | - | - |
| Myometrium | 1 | - | 5 | 3 |
| Ovary | 1 | - | 15 | 6 |
| Prostate | 1 | - | 3 | - |
| Putamen | 34 | 8 | - | - |
| Skin | 24 | 9 | 2 | - |
| Substantia Nigra Pars Compacta | 1 | - | 1 | - |
| Substantia Nigra Reticulata | 2 | - | - | - |
| Synovial Membrane | 5 | 1 | 2 | 1 |
| Thalamus Lateral Nuclei | 1 | - | 1 | - |
| Thalamus Subthalamic Nucleus | 1 | - | - | - |

Table 5.2: Filtered distinct gene counts according to SLR values for all experimented tissues.

| | Distict Genes | |
|---|---|---|
| | t=3 | t=4 |
| Breast | LRFN1 | |
| | RPS6KB1,MED14,SLC16A3, TSPAN1,USP9X,HNRNPA1,CD44 | |
| Gloubus Pallidum External | SFTPB | - |
| Myometrium | GLRX3,TACC1,COL4A1 | |
| | ZNF614,TC2N | |
| Ovary | CMBL,MAF,CAPZB,BGN,NDRG2,TET2 | |
| | HMOX2,BHLHE41,GRK6,PPFIA2 | |
| Prostate | PTGS1,NAMPT,SLC46A3 | - |
| Skin | C22orf25,BHLHE41 | - |
| Substantia Nigra Pars Compacta | COL16A1 | - |
| Synovial Membrane | EMR2 | |
| | QPCT | |
| Thalamus Lateral Nuclei | SFTPB | - |

Table 5.3: Notable genes processed by SLR values and appearing only in diseased samples. Lower SLR values than threshold=3 were excluded from the set.

### 5.2.2 Analysis Based on Gene Frequency

In this case, all provided diseased and normal samples of the same tissue were merged separately according to gene occurrence. Traversing all samples of each tissue, the total count of each gene was calculated and stored for diseased and normal samples. Mutually occurring and distinct genes were determined and ordered by total count. We first identified genes mutually occurring for diseased and normal samples. Processing each tissue one by one, we measured difference of genes existence between normal and diseased tissues for all genes. The higher the difference, the better detection for genes activated on the diseased state. Distinct gene count data was same as in the previous section. The count of mutually occurring genes, whose occurrence

is superior in either normal or diseased condition, and the count of distinct genes in complete set are both reported in Appendix B.3.

We then experimented on completely distinct genes which show significant 3' UTR shortening either for diseased or normal samples. To reduce clutter, we eliminated distinct genes with total count lower than 3, and selected more frequent ones. Filtered distinct gene count according to frequency for all experimented tissues are given in Table 5.4. As in the analysis based on SLR value, the tissues that have the highest number of distinct genes for normal samples are: Accumbens, Putamen and Skin, and for diseased samples: Myometrium, Ovary and Breast. There were also some additional tissues which were not prominent in the analysis based on gene SLR value but gene occurrence. Substantia Nigra Pars Compacta and Substantia Nigra Reticulata have both more distinct genes than the threshold for normal samples. Prostate is the third among fifteen tissues according to gene occurrence showing 3' UTR shortening in the diseased case. Synovial Membrane tissue is the one having relatively high amount of distinct genes for both diseased and normal cases. Evaluating tissues on the dendrogram structures and the tissue hierarchy, we saw that Accumbens, Putamen, Substantia Nigra Pars Compacta and Substantia Nigra Reticulata are both nervous system tissues. They are located in pairs (Accumbens-Putamen and Substantia Nigra Pars Compacta-Substantia Nigra Reticulata) on Normal-Disease dendrogram. Skin, which is an integument tissue, was found in a slightly farther branch. Myometrium, Ovary, Breast and Prostate are both reproductive system tissues and Synovial Membrane is a connective tissue. All those tissues are clustered together in the same sub tree on Normal-Disease dendrogram.

If only distinct genes in diseased samples are taken into account; Breast, Myometrium, Ovary, Prostate and Synovial Membrane tissues are the prominent ones which have many genes occurring more than 4. Top 10 significant distinct genes meeting the criteria for each tissue are given in Table 5.5. According to tissue and pathology atlas of each gene, DNAJC21, SP110, SARDH and PBRM1 are reported as disease related genes, HSPH1,CD44, PBRM1,PTGS1 and NAMPT are reported as cancer related genes and TET2, TCF3, SETBP1,AARS2 and FOXP1 are reported as both [75, 75, 76].

| | Distinct Gene Count | | | |
|---|---|---|---|---|
| | Normal | | Diseased | |
| | t=3 | t=4 | t=3 | t=4 |
| Accumbens | 132 | 94 | - | - |
| Breast | - | - | 12 | 5 |
| Caudate | 2 | 1 | - | - |
| Gloubus Pallidum External | - | - | 3 | 1 |
| Gloubus Pallidum Internal | 1 | - | 2 | 1 |
| Myometrium | 3 | 1 | 34 | 25 |
| Ovary | 1 | - | 67 | 48 |
| Prostate | 1 | - | 22 | 7 |
| Putamen | 128 | 96 | 1 | - |
| Skin | 57 | 21 | - | - |
| Substantia Nigra Pars Compacta | 7 | 3 | - | - |
| Substantia Nigra Reticulata | 6 | 2 | 1 | - |
| Synovial Membrane | 12 | 4 | 6 | 4 |
| Thalamus Lateral Nuclei | - | - | - | - |
| Thalamus Subthalamic Nucleus | - | - | - | - |

Table 5.4: Filtered distinct gene counts according to frequency for all experimented tissues.

### 5.2.3 Analysis Based on All Data

In this case, all provided diseased and normal samples of tissues were merged together first according to SLR values of genes and second according to gene occurrence. The aim of this analysis was to find candidate genes which may likely to be correlated with critical regulatory functions genome-wide. Resulting genes were added into Table 5.6. The first set shows distinct gene list which is available in just disease sample group but not in any other group. The second set displays distinct gene list that is available in just normal sample group and total occurrence count across all tissues is more than 20. The third set presents mutually occurring genes across all tissues but

| | Distinct Genes | |
|---|---|---|
| | t=3 | t=4 |
| Breast | MED14,CD44,TAGAP, TDRD9,WIZ | |
| | PBRM1,MGAT4A,ADAL, SLC16A3,CORO2A | |
| Gloubus Pallidum External | STAC,FOXP1,ACACB | STAC |
| Gloubus Pallidum Internal | HSPH1,NAP1L1 | HSPH1 |
| Prostate | RPS6KB1,C11orf24,NOP58,NAMPT SYF2,SLC46A3,PTGS1 | |
| | HIPK1,SP110,ENOX2 | |
| Myometrium | ARL1,GLRX3,TROVE2,SETBP1,TLE4,SARDH NCAPH2,TC2N,TBC1D9,CRYZ | |
| Ovary | CAPZB,B3GNT2,TCF3, TET2,RPL19,DNAJC21 GRK6,ZNF326,HSPH1,LMAN2 | |
| Putamen | ARPP19 | - |
| Substantia Nigra Reticulata | PTPRM | - |
| Synovial Membrane | MAP3K7IP3,AARS2, NPAT,SLAMF8 | |
| | LCP2,COBL | |

Table 5.5: Notable genes processed by frequency and appearing only in the diseased samples. Lower SLR values than 3 are excluded from the set.

just the ones whose frequency is much more higher in the diseased samples. Diseased and cancer related genes for each set were also added to the table [75, 75, 76].

### 5.3 Discussions

All in all, both analyses demonstrate that variable sample count for tissues does not have a significant impact on resulted distinct gene counts. Even though distinct gene count and frequency are linearly dependent, data set limitation with proper threshold value annihilate dependency and reveal likely to be important data. For instance, My-

| | Genes | Disease Related | Cancer Related |
|---|---|---|---|
| Set 1 | ZNF614,ATP6V1A,HSD11B2, SUMO3,DYRK4,MAGOHB, IL13RA1,FAM114A1,GLG1, C10orf104,SLCO4A1,DDR2, ENTPD6 | HSD11B2, DDR2 | DDR2 |
| Set 2 | SLC25A17,ESR1,PALMD,RABIF, NR3C1,TTC37,KLK13, REM1,HSPA4,SCRN3,PPIC, MYO1B,OBSL1,FAHD1,DDX18, DLGAP2,TMEM92,CREB1,LPL, IL1R2,ABCE1,DDX28,TMEM18 | ESR1,NR3C1, TTC37,OBSL1, CREB1,LPL | ESR1, KLK13,HSPA4, CREB1,IL1R2 |
| Set 3 | TMEM182,TMEM38B,HIPK1,TYW3, SCFD2,ZNF345,PLOD2,MMP10, NTRK3,MRPL28,EMR2,NKD2, CNTN6,TPT1,IGBP1,ALDOB, ING1,SUB1,TMEM9B,AP3D1, SEC11A,THBS1, ARPP19 | TMEM38B, PLOD2,NTRK3, IGBP1,ALDOB, ING1,AP3D1 | MMP10,NTRK3, ALDOB,ING1, THBS1 |

Table 5.6: Genes with prominent SLR behaviour genome-wide

ometrium was the tissue with the highest normal sample count but it was containing much less distinct genes than lots of other tissues with fewer normal samples. Similarly, Ovary and Prostate have the same amount of diseased samples but the Ovary has much more distinct genes than the Prostate in both of the analyses.

Another thing to note is that hierarchically closer tissues show similar behaviour in case of diseased or normal conditions. Although nervous system tissues are more prone to 3' UTR shortening in the normal state, reproductive system tissues are likely to show proximal APA in the diseased state. It is anticipated that most of notable genes which were presented for each analysis were not common because source data set and criteria were different for each case. Investigating the distinct genes which frequently occur in diseased samples with higher SLR values, we found some notewor-

thy genes which are TET2 and GRK6 for Ovary, GLRX3 and TC2N for Myometrium and MED14 and CD44 for Breast. All are protein coding genes. Although cancer tissue analysis is still pending for some; TET2, TC2N and CD44 genes are classified as disease and cancer related genes and prognostic markers in various cancer types according to The Human Protein Atlas [75, 75, 76].

# CHAPTER 6

# DISCUSSIONS AND CONCLUSIONS

In this thesis, we have aimed to comprehend the significance and consequence of alternative polyadenylation events in differentiation of functional regulatory mechanisms which may lead to various diseases as well as accumulation of abnormal cells and malignant (cancerous) growths. Although there are numerous state-of-the-art prior work and ongoing research, there are still many unanswered questions. Different from related researches, we mainly focus on genome-wide analysis of proximal APA on human transcripts. We analyzed the effect of proximal APA on both tissues and genes. Significant 3' UTR shortening events are observed and reported across variety of human tissues. However, length variation and degree may depend on tissue type and hierarchical similarity of tissues. Besides, tissue specificity also affects 3' UTR isoform expression of activated genes, which means the same gene may be activated on many tissues but show proximal APA in just one tissue.

We performed three different analyses in this thesis to examine human tissues and genes in the presence of 3' UTR shortening events. We aimed to investigate tissues becoming distant from their hierarchically closer neighbors and also to detect tissue groups whose members are clustering together and showing similarity with the closely related tissues according to the biological tissue hierarchy. In addition to that, we tried to identify several novel genes that are related with various diseases.

We worked on a primary dataset which was profiled for gene expression using the Affymetrix U133 plus 2.0 array. It has 677 human tissue samples, each keeps active gene-SLR value assignments. Samples may be profiled for different characteristics, but we mainly focused on Diseased and Normal ones. Higher SLR shows higher proximal APA for a gene. Since each tissue has several samples, we combined them

according to tissue characteristics and obtained one sample having active genes with higher SLR values for each tissue.

The first analysis investigates whether hierarchically closer tissues show similar proximal APA behaviour both on Diseased and Normal conditions or not. We used 129 merged tissue samples and constructed various dendrogram structures from distance matrices. Distance matrices were created processing top 20-50-100 genes with maximum SLR values for each tissue. NJ and UPGMA hierarchical clustering algorithms were used in reconstructing dendrograms. We presented two algorithms to find ancestor based accuracies. They simply cluster tissue nodes in an input tree according to some criteria and measure closeness of the tree with the actual tissue hierarchy. Using optimum threshold for criterion, over $\%93$ accuracy was obtained for tree constructed by diseased and correspondent normal samples. It was over $\%90$ for diseased samples and over $\%50$ for normal samples. Most ideal results were achieved with average distance threshold for NJ and optimum threshold for UPGMA. It supports the idea that top genes showing significant 3'UTR shortening both in hierarchically closer tissues and disease-normal samples of the same tissue are mostly similar. Their presence and length variation may change but it is not considered within this analysis. We saw that hierarchically closer tissues mostly cluster together in the dendrograms and the tissue family can be constructed by just looking at top activated genes by SLR values. Most diseased-normal couples of tissue samples also show high correlation but Putamen and Skin tissues diverged from normal samples in the disease state.

The second analysis creates two different multi-layer tissue networks, considering proteins with SLR or not, and investigates if genes activated in closer layers have similar feature set or not. It also examines tissue behaviour for two networks and reveals main tissue groups which differentiated from closer neighbors when the dataset is filtered according to SLR values of the genes. Unlike the previous analysis, PPI network of each tissue was used to feed the system and to create feature embeddings of each gene activated in the tissue. Processed tissue set was slightly different from previous analysis, therefore, a new tissue hierarchy up to 5 levels was created and given to the system as a multiscale tissue hierarchy. Ohmnet model, built on state of the art Word2Vec, was adapted to learn a d-dimensional feature vector for every gene node in every tissue [26, 63, 64]. Two different embeddings were constructed, one

with all PPI network and other with filtered PPI network according to proteins having high SLR values. Since the size of feature vector is 128, PCA and t-SNE was used to model data into low dimensional space. We then visualized each node, i.e. gene, with color mapping based on tissue to analyze similarity of tissues. We first demonstrated that genes in the same sub-tissues were mapped together and clearly differentiated from other tissues. Second, most of the sub-tissues were also clustered together with other sub-tissues belong to same main tissue. This shows that hierarchically closer tissues are highly correlated with each other in both experiments. Comparing two experiments, we can say that although sub-tissues continue to be well separated from others, some of hierarchically closer main tissues, such as nervous systems and head, were mapped better in the experiment with SLR values. Even though we can not directly discuss hierarchical relation of main tissues, all derived from 'whole body' tissue according to Brenda, we can accept the resulting network with all PPI data as a reference and make following inference. Distinct tissue groups, having similar active gene data set filtered by SLR value, may indicate closeness. Because those tissues were shown to share more associated co-expressed genes and interactions according to our model, common functional and regulatory disease trends among them may further investigated and validated by water based experiments.

The third analysis is on tissue specific proximal pol(A) site differentiation. It also seeks notable genes by processing normal-diseased tissue sample pairs and tries to find out if those genes are classified as markers of various disease and cancer types by examining their protein class. The analysis is conducted on the same tissue dataset as pre-processed in the first analysis but this time, focuses on only diseased and normal samples. SLR value-gene analysis were made according to 3'UTR length variation and occurrence of genes. First mutually occurring then distinct genes of the disease-normal pair of each tissue were studied. Resulting genes were filtered according to threshold. Among all tissues, Breast, Ovary and Myometrium are the top three whose distinct genes are more prune to 3'UTR shortening in disease condition. They are all reproductive system tissues and located closely in the dendrograms created before. Skin, Accumbens and Putamen are the prominent tissues in normal condition. Accumbens and Putamen are nervous system tissues and siblings in the dendrogram. Skin is an integument tissue and positioned a little farther from them. Prostate, a

reproductive system tissue, is also important such that it gave one of the higher results in case of gene occurrence in diseased case. According to the results, it can be noted that hierarchically closer tissues tend to show similar shortening behaviour. Nervous and reproductive system tissues are the most promising group for further experiments because they provided most differentiated proximal APA data. Apart from these, we also shared some noteworthy genes which are proven to be associated with various diseases and cancer categories for further biological research. Detailed data can be found on Chapter 6.

## 6.1 Future Work

There are several directions for future work. In our research, SLR data, which was extracted by performing micro array gene expression profiles, was used to examine genes showing alternative polyadenylation. With the increasing number of expressed sequence data in public databases and recently developed methods specialized for comprehensive APA profiling, such as 3'-enriched RNA-seq and PAS-Seq, it becomes easier to reveal more genes showing APA behaviour in mammalian cells. Analyses within the scope of this thesis can be enhanced with more comprehensive set of genes and accuracy results can be examined to see if there is any improvement.

In addition to that, we analyzed organism wide APA events in our work and studied effects of shortening events in terms of tissue hierarchy. Because there are many tissues to handle, our analysis sometimes yielded general inferences. Further experiments may concentrate on more specific tissues and genes which show excessive shortening events and evaluate gene behaviours specific to those tissues.

## REFERENCES


[1] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC®: Transcriptional regulation, from patterns to profiles," 2003.

[2] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob, "The IntAct molecular interaction database in 2010," *Nucleic Acids Research*, 2009.

[3] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2012 Update," *Nucleic Acids Research*, 2012.

[4] C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers, "The BioGRID Interaction Database: 2011 update," *Nucleic Acids Research*, 2011.

[5] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. I. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database - 2009 update," *Nucleic Acids Research*, 2009.

[6] M. Kanehisa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, 2005.

[7] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. O. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences*, 2007.

[8] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stümpflen, and H. W. Mewes, "CORUM: The comprehensive resource of mammalian protein complexes," *Nucleic Acids Research*, 2008.

[9] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan, "PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Research*, 2012.

[10] A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker, "A directed protein interaction network for investigating intracellular signal transduction," *Science Signaling*, 2011.

[11] A. Sarma, R. Sandberg, C. B. Burge, J. R. Neilson, and P. A. Sharp, "Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites," *Science*, 2008.

[12] C. Mayr and D. P. Bartel, "Supplementary Information - Widespread Shortening of 3UTRs by Alternative Cleavage and Polyadenylation," *Cell*, 2009.

[13] Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, and B. Tian, "Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development," *Proceedings of the National Academy of Sciences*, 2009.

[14] A. Lembo, F. Di Cunto, and P. Provero, "Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer," *PLoS ONE*, 2012.

[15] A. E. Erson-Bensan and T. Can, "Alternative Polyadenylation: Another Foe in Cancer," *Molecular Cancer Research*, 2016.

[16] J.-W. Chang, H.-S. Yeh, and J. Yong, "Alternative Polyadenylation in Human Diseases," *Endocrinology and Metabolism*, 2017.

[17] H. Zhang, J. Y. Lee, and B. Tian, "Biased alternative polyadenylation in human tissues.," *Genome biology*, 2005.

[18] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr, "Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression," *Genes and Development*, 2013.

[19] P. J. Shepard, E. A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi, "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq," *RNA*, 2011.

[20] A. Derti, P. Garrett-Engele, K. D. MacIsaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak, "A quantitative atlas of polyadenylation in five mammals," *Genome Research*, 2012.

[21] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu, "Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing," *Genome Research*, 2011.

[22] M. H. Kim, B. H. You, and J. W. Nam, "Global estimation of the 3' untranslated region landscape using RNA sequencing," *Methods*, 2015.

[23] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel, "Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs," *Nature*, 2011.

[24] Y. Li, Y. Sun, Y. Fu, M. Li, G. Huang, C. Zhang, J. Liang, S. Huang, G. Shen, S. Yuan, L. Chen, S. Chen, and A. Xu, "Dynamic landscape of tandem 3' UTRs during zebrafish development," *Genome Research*, 2012.

[25] C. C. MacDonald and K. W. McMahon, "Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond," 2010.

[26] M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," in *Bioinformatics*, 2017.

[27] N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? A proposed definition and overview of the field.," *Methods of information in medicine*, 2001.

[28] I. San Segundo-Val and C. S. Sanz-Lozano, *Introduction to the Gene Expression Analysis*, pp. 29–43. New York, NY: Springer New York, 2016.

[29] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells," *PLoS ONE*, 2014.

[30] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, p. 57, sep 2012.

[31] C. Bach and P. Patra, "Human genome regulation," *Bioengineered*, 2016.

[32] W. P. Albert B., Johnson A., Lewis J., Raff M., Roberts K., *Moleculer Biology of the Cell, 4th Edition*. 2002.

[33] A. A. Bicknell, C. Cenik, H. N. Chua, F. P. Roth, and M. J. Moore, "Introns in UTRs: Why we should stop ignoring them," *BioEssays*, 2012.

[34] G. S. Wang and T. A. Cooper, "Splicing in disease: Disruption of the splicing code and the decoding machinery," 2007.

[35] J. E. Darnell, L. Philipson, R. Wall, and M. Adesnik, "Polyadenylic acid sequences: Role in conversion of nuclear RNA into messenger RNA," *Science*, 1971.

[36] K. Glover-Cutter, S. Kim, J. Espinosa, and D. L. Bentley, "RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes," *Nature Structural and Molecular Biology*, 2008.

[37] B. H. Akman, T. Can, and A. Elif Erson-Bensan, "Estrogen-induced upregulation and 3'-UTR shortening of CDC6," *Nucleic Acids Research*, 2012.

[38] Y. Ilguner, *Prediction of polyadenylation sites by probe level analysis of microarray data*. PhD thesis, Middle East Technical University, 2013.

[39] J. Y. Lee, I. Yeh, J. Y. Park, and B. Tian, "PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes," *Nucleic Acids Research*, 2007.

[40] P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai, "Widespread and extensive lengthening of 39 UTRs in the mammalian brain," *Genome Research*, 2013.

[41] J. E. Gentle, L. Kaufman, and P. J. Rousseuw, "Finding Groups in Data: An Introduction to Cluster Analysis.," *Biometrics*, 2006.

[42] R. Sokal, C. Michener, and U. of Kansas, *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin, University of Kansas, 1958.

[43] E. Jones, T. Oliphant, and P. Peterson, "Scipy: Open source scientific tools for python," 2001.

[44] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," 09 2011.

[45] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, 1987.

[46] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg, "The BRENDA Tissue Ontology (BTO): The first all-integrating ontology of all organisms for enzyme sources," *Nucleic Acids Research*, 2011.

[47] A. Chang, I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. W. Sensen, and D. Schomburg, "BRENDA in 2015: Exciting developments in its 25th year of existence," *Nucleic Acids Research*, 2015.

[48] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, pp. 207–210, 01 2002.

[49] M. Carlson, "org.Hs.eg.db: Genome wide annotation for Human. R package version 3.7.0.," 2018.

[50] J. Felsenstein, "Phylogenies from Molecular Sequences: Inference and Reliability," *Annual Review of Genetics*, 1988.

[51] R. Scott, B. MacPherson, and R. Gras, "Ecosim, an enhanced artificial ecosystem: addressing deeper behavioral, ecological, and evolutionary questions," in *Intelligent Systems, Control and Automation: Science and Engineering*, 2019.

[52] G. Cardona, F. Rosselló, and G. Valiente, "Extended Newick: It is time for a standard representation of phylogenetic networks," *BMC Bioinformatics*, 2008.

[53] G. Olsen, "Gary Olsen's interpretation of the "Newick's 8:45" tree format standard," 1990.

[54] O. Robinson, D. Dylus, and C. Dessimoz, "Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web," *Molecular Biology and Evolution*, 2016.

[55] D. F. Robinson and L. R. Foulds, "Comparison of weighted labelled trees," in *Combinatorial Mathematics VI* (A. F. Horadam and W. D. Wallis, eds.), (Berlin, Heidelberg), pp. 119–126, Springer Berlin Heidelberg, 1979.

[56] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Šmuc, Y. A. Kourmpetis, A. D. Van Dijk, C. J. Ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg, "A large-scale evaluation of computational protein function prediction," *Nature Methods*, 2013.

[57] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, 2003.

[58] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," *CoRR*, vol. abs/1011.4071, 2010.

[59] A. E. Aladağ and C. Erten, "SPINAL: Scalable protein interaction network alignment," *Bioinformatics*, 2013.

[60] M. Gustafsson, C. E. Nestor, H. Zhang, A. L. Barabási, S. Baranzini, S. Brunak, K. F. Chung, H. J. Federoff, A. C. Gavin, R. R. Meehan, P. Picotti, M. À. Pujana, N. Rajewsky, K. G. Smith, P. J. Sterk, P. Villoslada, and M. Benson, "Modules, networks and systems medicine for understanding disease and aiding diagnosis," 2014.

[61] M. Žitnik and B. Zupan, "Data fusion by matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[62] M. Kramer, J. Dutkowski, M. Yu, V. Bafna, and T. Ideker, "Inferring gene ontologies from pairwise similarity data," in *Bioinformatics*, 2014.

[63] A. Grover and J. Leskovec, "node2vec," 2016.

[64] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," pp. 1–12, 01 2013.

[65] D. Vella, S. Marini, F. Vitali, D. Di Silvestre, G. Mauri, and R. Bellazzi, "MTGO: PPI Network Analysis Via Topological and Functional Module Identification," *Scientific Reports*, 2018.

[66] Q. Zhong, S. J. Pevzner, T. Hao, Y. Wang, R. Mosca, J. Menche, M. Taipale, M. Tasan, C. Fan, X. Yang, P. Haley, R. R. Murray, F. Mer, F. Gebreab, S. Tam, A. MacWilliams, A. Dricot, P. Reichert, B. Santhanam, L. Ghamsari, M. A. Calderwood, T. Rolland, B. Charloteaux, S. Lindquist, A.-L. Barabasi, D. E. Hill, P. Aloy, M. E. Cusick, Y. Xia, F. P. Roth, and M. Vidal, "An inter-species protein-protein interaction network across vast evolutionary distance," *Molecular Systems Biology*, 2016.

[67] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A. L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, 2015.

[68] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, and H. Hermjakob, "The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases.," *Nucleic acids research*, 2014.

[69] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal, "A proteome-scale map of the human interactome network.," *Cell*, 2014.

[70] A. Chatr-Aryamontri, B. J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2015 update," *Nucleic Acids Research*, 2015.

[71] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, D. I. Chas-

man, G. A. Fitzgerald, K. Dolinski, T. Grosser, and O. G. Troyanskaya, "Understanding multicellular function and disease with human tissue-specific networks," *Nature Genetics*, 2015.

[72] I. T. Jolliffe, *Principal Component Analysis. Second Edition*. 2002.

[73] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 2008.

[74] F. Pedregosa, V. Michel, O. Grisel OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, A. Duchesnay, and F. Duchesnay EDOUARDDUCHESNAY, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," *Journal of Machine Learning Research*, 2011.

[75] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. M. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. Von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. Von Heijne, J. Nielsen, and F. Pontén, "Tissue-based map of the human proteome," *Science*, 2015.

[76] P. J. Thul, L. Akesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, A. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. Von Feilitzen, K. S. Lilley, M. Uhlén, and E. Lundberg, "A subcellular map of the human proteome," *Science*, 2017.

## APPENDIX A

## GSE7307 SAMPLE IDS

GSM175786, GSM175787, GSM175788, GSM175789, GSM175790, GSM175791, GSM175792, GSM175793, GSM175794, GSM175795, GSM175796, GSM175797, GSM175798, GSM175799, GSM175800, GSM175801, GSM175802, GSM175803, GSM175804, GSM175805, GSM175806, GSM175807, GSM175808, GSM175809, GSM175810, GSM175811, GSM175812, GSM175813, GSM175814, GSM175815, GSM175816, GSM175817, GSM175818, GSM175819, GSM175820, GSM175821, GSM175822, GSM175823, GSM175824, GSM175825, GSM175826, GSM175827, GSM175828, GSM175829, GSM175830, GSM175831, GSM175832, GSM175833, GSM175834, GSM175835, GSM175836, GSM175837, GSM175838, GSM175839, GSM175840, GSM175841, GSM175842, GSM175843, GSM175844, GSM175845, GSM175846, GSM175847, GSM175848, GSM175849, GSM175850, GSM175851, GSM175852, GSM175853, GSM175854, GSM175855, GSM175856, GSM175857, GSM175858, GSM175859, GSM175860, GSM175861, GSM175862, GSM175863, GSM175864, GSM175865, GSM175866, GSM175867, GSM175868, GSM175869, GSM175870, GSM175871, GSM175872, GSM175873, GSM175874, GSM175875, GSM175876, GSM175877, GSM175878, GSM175879, GSM175880, GSM175881, GSM175882, GSM175883, GSM175884, GSM175885, GSM175886, GSM175887, GSM175888, GSM175889, GSM175890, GSM175891, GSM175892, GSM175893, GSM175894, GSM175895, GSM175896, GSM175897, GSM175898, GSM175899, GSM175900, GSM175901, GSM175902, GSM175903, GSM175904, GSM175905, GSM175906, GSM175907, GSM175908, GSM175909, GSM175910, GSM175911, GSM175912, GSM175913, GSM175914, GSM175915, GSM175916, GSM175917, GSM175918, GSM175919, GSM175920, GSM175921, GSM175922, GSM175923, GSM175924, GSM175925, GSM175926, GSM175927, GSM175928, GSM175929,

GSM175930, GSM175931, GSM175932, GSM175933, GSM175934, GSM175935, GSM175936, GSM175937, GSM175938, GSM175939, GSM175940, GSM175941, GSM175942, GSM175943, GSM175944, GSM175945, GSM175946, GSM175947, GSM175948, GSM175949, GSM175950, GSM175951, GSM175952, GSM175953, GSM175954, GSM175955, GSM175956, GSM175957, GSM175958, GSM175959, GSM175960, GSM175961, GSM175962, GSM175963, GSM175964, GSM175965, GSM175966, GSM175967, GSM175968, GSM175969, GSM175970, GSM175971, GSM175972, GSM175973, GSM175974, GSM175975, GSM175976, GSM175977, GSM175978, GSM175979, GSM175980, GSM175981, GSM175982, GSM175983, GSM175984, GSM175985, GSM175987, GSM175988, GSM175989, GSM175990, GSM175991, GSM175992, GSM175993, GSM175994, GSM175995, GSM175996, GSM175997, GSM175998, GSM175999, GSM176000, GSM176001, GSM176002, GSM176003, GSM176004, GSM176005, GSM176006, GSM176007, GSM176008, GSM176009, GSM176010, GSM176011, GSM176012, GSM176013, GSM176014, GSM176015, GSM176016, GSM176017, GSM176018, GSM176019, GSM176020, GSM176021, GSM176022, GSM176023, GSM176024, GSM176025, GSM176026, GSM176027, GSM176028, GSM176029, GSM176030, GSM176031, GSM176032, GSM176033, GSM176034, GSM176035, GSM176036, GSM176037, GSM176038, GSM176039, GSM176040, GSM176041, GSM176042, GSM176043, GSM176044, GSM176045, GSM176046, GSM176047, GSM176048, GSM176049, GSM176050, GSM176051, GSM176052, GSM176053, GSM176054, GSM176055, GSM176056, GSM176057, GSM176058, GSM176059, GSM176060, GSM176061, GSM176062, GSM176063, GSM176064, GSM176065, GSM176066, GSM176067, GSM176068, GSM176069, GSM176070, GSM176071, GSM176072, GSM176073, GSM176074, GSM176075, GSM176076, GSM176077, GSM176078, GSM176079, GSM176080, GSM176081, GSM176082, GSM176083, GSM176084, GSM176085, GSM176086, GSM176087, GSM176088, GSM176089, GSM176090, GSM176091, GSM176092, GSM176093, GSM176094, GSM176095, GSM176096, GSM176097, GSM176098, GSM176099, GSM176100, GSM176101, GSM176102, GSM176103, GSM176104, GSM176105, GSM176106, GSM176107, GSM176108, GSM176109, GSM176110, GSM176111, GSM176112, GSM176113, GSM176114, GSM176115, GSM176116, GSM176117, GSM176118, GSM176119, GSM176120, GSM176121, GSM176122, GSM176123, GSM176124, GSM176125, GSM176126, GSM176127, GSM176128,

GSM176129, GSM176130, GSM176131, GSM176132, GSM176133, GSM176134, GSM176135, GSM176136, GSM176137, GSM176138, GSM176139, GSM176140, GSM176141, GSM176142, GSM176143, GSM176144, GSM176145, GSM176146, GSM176147, GSM176148, GSM176149, GSM176150, GSM176151, GSM176152, GSM176153, GSM176154, GSM176155, GSM176156, GSM176157, GSM176158, GSM176159, GSM176160, GSM176161, GSM176162, GSM176163, GSM176164, GSM176165, GSM176166, GSM176167, GSM176168, GSM176169, GSM176170, GSM176171, GSM176172, GSM176173, GSM176174, GSM176175, GSM176176, GSM176177, GSM176178, GSM176179, GSM176180, GSM176181, GSM176182, GSM176183, GSM176184, GSM176185, GSM176186, GSM176205, GSM176206, GSM176207, GSM176208, GSM176209, GSM176210, GSM176211, GSM176212, GSM176213, GSM176214, GSM176215, GSM176216, GSM176217, GSM176218, GSM176219, GSM176220, GSM176221, GSM176222, GSM176223, GSM176224, GSM176225, GSM176226, GSM176227, GSM176228, GSM176229, GSM176230, GSM176231, GSM176232, GSM176233, GSM176234, GSM176235, GSM176236, GSM176237, GSM176238, GSM176239, GSM176240, GSM176241, GSM176242, GSM176243, GSM176244, GSM176245, GSM176246, GSM176247, GSM176248, GSM176249, GSM176250, GSM176251, GSM176252, GSM176253, GSM176254, GSM176255, GSM176256, GSM176257, GSM176258, GSM176259, GSM176260, GSM176261, GSM176262, GSM176263, GSM176264, GSM176265, GSM176266, GSM176267, GSM176268, GSM176269, GSM176270, GSM176271, GSM176272, GSM176273, GSM176274, GSM176275, GSM176276, GSM176277, GSM176278, GSM176279, GSM176280, GSM176281, GSM176282, GSM176283, GSM176284, GSM176285, GSM176286, GSM176287, GSM176288, GSM176289, GSM176290, GSM176291, GSM176292, GSM176293, GSM176294, GSM176295, GSM176296, GSM176297, GSM176298, GSM176299, GSM176300, GSM176301, GSM176302, GSM176303, GSM176304, GSM176305, GSM176306, GSM176307, GSM176308, GSM176309, GSM176310, GSM176311, GSM176312, GSM176313, GSM176314, GSM176315, GSM176316, GSM176317, GSM176318, GSM176319, GSM176320, GSM176321, GSM176322, GSM176323, GSM176324, GSM176325, GSM176326, GSM176327, GSM176328, GSM176329, GSM176330, GSM176331, GSM176332, GSM176333, GSM176334, GSM176335, GSM176336, GSM176337, GSM176338, GSM176339, GSM176340, GSM176341, GSM176342, GSM176343, GSM176344,

GSM176345, GSM176346, GSM176347, GSM176348, GSM176349, GSM176350, GSM176351, GSM176352, GSM176353, GSM176354, GSM176355, GSM176356, GSM176357, GSM176358, GSM176359, GSM176360, GSM176361, GSM176362, GSM176363, GSM176364, GSM176365, GSM176366, GSM176367, GSM176368, GSM176369, GSM176370, GSM176371, GSM176372, GSM176373, GSM176374, GSM176375, GSM176376, GSM176377, GSM176378, GSM176379, GSM176380, GSM176381, GSM176382, GSM176383, GSM176384, GSM176385, GSM176386, GSM176387, GSM176388, GSM176389, GSM176390, GSM176391, GSM176392, GSM176393, GSM176394, GSM176395, GSM176396, GSM176397, GSM176398, GSM176399, GSM176400, GSM176401, GSM176402, GSM176403, GSM176404, GSM176405, GSM176406, GSM176407, GSM176408, GSM176409, GSM176410, GSM176411, GSM176412, GSM176413, GSM176414, GSM176415, GSM176416, GSM176417, GSM176418, GSM176419, GSM176420, GSM176421, GSM176422, GSM176423, GSM176424, GSM176425, GSM176426, GSM176427, GSM176428, GSM176429, GSM176430, GSM176431, GSM176432, GSM176433, GSM176434, GSM176435, GSM176436, GSM176437, GSM176438, GSM176439, GSM176440, GSM176441, GSM176442, GSM176443, GSM176444, GSM176445, GSM176446, GSM176447, GSM176448, GSM176449, GSM176450, GSM176451, GSM176452, GSM176453, GSM176454, GSM176455, GSM176456, GSM176457, GSM176458, GSM176459, GSM176460, GSM176461, GSM176462, GSM176463, GSM176464, GSM176465, GSM176466, GSM176467, GSM176468, GSM176469, GSM176470, GSM176471, GSM176472, GSM176473, GSM176474, GSM176475, GSM176476, GSM176477, GSM176478, GSM176479, GSM176480, GSM176481

# APPENDIX B

# SUPPLEMENTARY TABLES

## B.1 Distinct/Mutually Occurring Gene Count of Complete Set Based on SLR values

| | Mutually Occurring Gene Count | | Distinct Gene Count | |
|---|---|---|---|---|
| | Normal>Disease | Disease>Normal | Normal | Diseased |
| Accumbens | 123 | 15 | 291 | 7 |
| Breast | 118 | 170 | 33 | 137 |
| Caudate | 72 | 42 | 52 | 16 |
| Gloubus Pallidum External | 36 | 57 | 16 | 33 |
| Gloubus Pallidum Internal | 59 | 42 | 50 | 22 |
| Myometrium | 241 | 269 | 85 | 116 |
| Ovary | 131 | 228 | 48 | 188 |
| Prostate | 46 | 223 | 17 | 131 |
| Putamen | 141 | 10 | 305 | 3 |
| Skin | 148 | 61 | 255 | 35 |
| Substantia Nigra Pars Compacta | 72 | 42 | 50 | 26 |
| Substantia Nigra Reticulata | 57 | 43 | 40 | 23 |
| Synovial Membrane | 151 | 154 | 80 | 97 |
| Thalamus Lateral Nuclei | 55 | 39 | 40 | 22 |
| Thalamus Subthalamic Nucleus | 44 | 38 | 29 | 19 |

Table B.1: Distinct/Mutually Occurring Gene Count of Complete Set Based on SLR values.

## B.2 Mutual genes having higher SLR values in diseased samples

| | Gene Names |
|---|---|
| Accumbens | CLU |
| Breast | IGL@, PABPC3, TNFSF10, MMP7, RHOB, DAZAP2, STMN1, RPL13, SLITRK6, MRP63 |
| Caudate | CLU, FOXP1 |
| Gloubus Pallidum External | TNFSF10, CALM1, CLU, KIAA1245, ARPC2, DAZAP2, APC, RPL13, ANP32A, CUGBP2 |
| Gloubus Pallidum Internal | CLU, KIAA12, CALM1 |
| Myometrium | RHOB, PAM, PNRC1, TNFSF10, LOC284454, IGL@, VASP, GSTM5, MMP7, STMN1 |
| Ovary | IGL@, FDX1, C4A, PAM, COL16A1, DAZAP2, ANKRD12, MMP7, CLU, RGS10 |
| Prostate | IGL@, TNFSF10, PNRC1, PAM, MMP7, AMOT, CORO1C, KLK4, COL16A1, ANKRD12 |
| Putamen | CLU, CALM1, RTN1 |
| Skin | IGL@, STMN1, RGS10, GSTM5, COL16A1, HADHA, CORO1C, DAZAP2, AK2, HNRNPA1 |
| Substantia Nigra Pars Compacta | CLU, APP |
| Substantia Nigra Reticulata | CLU, AQP4, SNX3, COL16A1, CMBL |
| Synovial Membrane | IGL@, KIAA1245, SLC16A3, STMN1, RIBC1, FLII, ARPC2, PLB1, APIP, CORO1C |
| Thalamus Lateral Nuclei | TNFSF10, CALM1, APC, KIAA1245, IL21R, DAZAP2, AQP4 |
| Thalamus Subthalamic Nucleus | TNFSF10, CLU, CALM1, APC, RTN1,AQP4, SNX3, PABPC3, MYST2 |

Table B.2: Mutual genes which having higher SLR values in diseased samples.

## B.3 Distinct/Mutually Occurring Gene Count of Complete Set Based on SLR frequency

| | Common Gene Count | | Diverged Gene Count | |
|---|---|---|---|---|
| | Normal>Disease | Disease>Normal | Normal | Diseased |
| Accumbens | 133 | 1 | 291 | 7 |
| Breast | 5 | 241 | 33 | 137 |
| Caudate | 93 | 2 | 52 | 16 |
| Gloubus Pallidum External | 7 | 75 | 16 | 33 |
| Gloubus Pallidum Internal | 12 | 66 | 50 | 22 |
| Myometrium | 61 | 391 | 85 | 116 |
| Ovary | 9 | 297 | 48 | 188 |
| Prostate | 5 | 244 | 17 | 131 |
| Putamen | 142 | 5 | 305 | 3 |
| Skin | 158 | 18 | 255 | 35 |
| Substantia Nigra Pars Compacta | 93 | 5 | 50 | 26 |
| Substantia Nigra Reticulata | 73 | 7 | 40 | 23 |
| Synovial Membrane | 195 | 41 | 80 | 97 |
| Thalamus Lateral Nuclei | 9 | 5 | 40 | 22 |
| Thalamus Subthalamic Nucleus | 16 | 5 | 29 | 19 |

Table B.3: Distinct/Mutually Occurring Gene Count of Complete Set Based on SLR frequency.