Summer 7-31-2018

Identifying Cancer Subtypes Using Unsupervised Deep Learning

Tejaswini Mallavarapu
KSU

Identifying Cancer Subtypes Using Unsupervised Deep Learning

A Thesis Presented to
The Faculty of the Computer Science Department

by

Tejaswini Mallavarapu

In Partial Fulfillment
of Requirements for the Degree
Master of Science, Computer Science

Kennesaw State University
July 2018

Identifying Cancer Subtypes Using Unsupervised Deep Learning

Approved:

_____

Dr. Mingon Kang - Advisor

_____

Dr. Dan Chia-Tien Lo– Department Chair

_____

Dr. Jon Preston - Dean

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Kennesaw State University, I agree that the university library shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish, this thesis may be granted by the professor under whose direction it was written, or, in his absence, by the dean of the appropriate school when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from or publication of, this thesis which involves potential financial gain will not be allowed without written permission.

_____

Tejaswini Mallavarapu

Notice To Borrowers

Unpublished theses deposited in the Library of Kennesaw State University must be used only in accordance with the stipulations prescribed by the author in the preceding statement.

The author of this thesis is:

Tejaswini Mallavarapu

1100 S Marietta PKWY,
Marietta, GA 30060

The director of this thesis is:

Dr. Mingon Kang

1100 S Marietta PKWY,
Marietta, GA 30060

Users of this thesis not regularly enrolled as students at Kennesaw State University are required to attest acceptance of the preceding stipulations by signing below. Libraries borrowing this thesis for the use of their patrons are required to see that each user records here the information requested.

ACKNOWLEDGEMENTS

With all respect, I would first like to thank Almighty God for giving me knowledge, strength and support I need. Second, I would like to express my very great appreciation to my advisor, Dr. Mingon Kang, for his valuable guidance, sheer support and constant encouragement throughout this entire research. Third, I would like to express my gratitude to my beloved husband, Mr. Sumanth Pudota and our little daughter Nehansika who have been my pillars, my joy and guiding light. Their love, patience and encouragement always upheld me in completing my thesis. Furthermore, I would like to thank my parents, in-laws, my brother and other family members for the motivation, inspiration and moral support.

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Cancer is a heterogeneous disease which has many subtypes that can be distinguished at molecular, histopathological, and clinical stage. Accurate diagnosis of the specific subtypes of cancer is vital to identify distinct disease states and opens up the possibility for effective personalized therapies that yield the greatest response. Many unsupervised machine learning techniques are applied to the genomic data of the tumor samples and the patient clusters are found to be of interest if they can be associated with a clinical outcome variable such as the survival of patients. In this thesis, we introduce two new clustering algorithms for cancer subtype identification, which fuses the information of gene expression and pathway database to group samples into biologically meaningful clusters. We call our first approach as R-PathCluster which is based on Restricted Boltzmann Machine. In this method, we used pathway markers as input dataset instead of gene expression data to identify unknown subtypes in Glioblastoma Multiforme (GBM). We developed SPACL model, sparse pathway based clustering for the identification of cancer subtypes in which multilayered deep belief network is used. In this model we used pathway data to extract the complex nonlinear relationships in identifying clusters. We assessed the performance of two models with several traditional clustering methods and found that our models out performed in clustering short term and long term survivals by lowest p-value in log rank test and Kaplan-Meier survival analysis. Our models provide solution to comprehensively detect subtypes and interpret in biological sense as these use pathway data.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Cancer Introduction

Cancer is a complex genetic disease characterized by uncontrolled, uncoordinated, and undesirable growth of abnormal malignant cells [1]. Its development and progression are generally connected to a sequence of changes in the activity of cell cycle regulators like proteins and enzymes. They can originate from any parts of the body and has the ability to spread throughout the body. In a healthy human body, most body cells follow a regular path; they partition, proliferate, and programmatically die [2]. The cell division rate of normal cells varies at different phases of each person's life. For instance, in younger person, healthy cells partition faster to enable him to grow properly, while in adult most of the cells divide to substitute aging or damaged cells in an adult. Cancer cells, on the other hand, grow abnormally and divide for their whole lives, replicating into more and more harmful cells, which consequently form lumps or mass of tissue called tumors [3]. This unrestricted cellular growth eventually leads to a transformation of normal cells into tumor cells by infiltrating normal tissues and organs.

The uncontrolled growth is caused by changes in DNA called genetic mutation. The DNA is a package of large number of individual genes, each of which contains a set of instructions telling the cell what functions to perform, as well as how to grow and divide. In normal cells, DNA repair genes looks for errors in DNA and attempts to repair itself.

If it does not succeed, the cell dies. However, in cancerous cell, mutations in DNA repair genes can cause the cell to impair its normal function. As a result the cell does not die as it is supposed to but replicates with similar damaged DNA [4]. Although inherited damaged DNA play a major role in about 5 to 10 percent of all cancers [5], majority of the genetic mutations are non-hereditary.

The mutations in DNA arise from multiple genetic and environmental factors. Genetic factors include faulty DNA copying introduced by DNA polymerases during cell division, Single Nucleotide Variants(SNV), small insertions and deletions (indels), larger copy number aberrations [6], gene expression changes, and epigenetic changes, including histone modifications and DNA methylation. Environmental factors include exposure to tobacco smoke, sunlight, etc. [7]. Exposure to Ultraviolet (UV) radiation and ionizing radiation (X-rays and atomic particles) damages DNA and transform normal cells into rapidly proliferating cancer cells. The ability of IR to cause leukemia cancer was significantly shown by the increased rates of leukemia among survivors of the nuclear bombs dropped in World War II [8][9]. Several studies have proved that long term exposure to sun damages DNA and cause melanoma, a type of skin cancer because of UV radiation[10].

Despite intensive efforts in research and treatment, cancer is still considered one of the leading causes of morbidity and mortality worldwide. According to the American Cancer Society, cancer is the second most common cause of death in the US, which accounts for nearly one of every four deaths [11]. Incidence of this illness is rising every year.

According to North American Association of Central Cancer Registries and National Center for Health Statistics, there were 1,735,350 new cases of cancer and 609,640 cancer deaths are projected to occur in 2018 in United States [12].

Early detection of tumors may increase patient's survival chances. Therefore, it is important to develop new diagnostic techniques and medical treatments which helps people to fight against this insidious disease. A major obstacle to design effective cancer therapies is the accurate stratification of patients. For example, at the time of diagnosis, it is critical to distinguish which patients possess aggressive tumors and which of them progress slowly. Aggressive treatment methods practiced on the latter worsen the quality of the patient life.

Cancer cells changes continuously while replicating because of aberrations in genes that regulate cell division. As the cancer advances over time, more genes will be aberrated leading to different types of daughter cancer cells. Although all cancer cells progress from single parent cell, the lump of cells that make up a cancer are not same, hence defined as heterogeneous. For example, when a breast cancer tumor is about one centimeter in size, millions of different cells that make up the lump are identified. Each cancer subtype has its own genetic identity, and gene expression pattern [13]. Thus two people with same age, height, weight, ethnicity, and similar medical histories, probably have two different breast cancer subtypes. It is therefore necessary to identify cancer subtypes in order to capture the characteristic essence of individual elements of this heterogeneous disease. Moreover, these cancer subtypes respond well to different treatment therapies or even

combination therapy. For example, one subtype of breast cancer known as estrogen receptor (ER) positive respond well to hormone therapy, whereas the human epidermal growth factor receptor 2 (HER2) positive subtype respond well to chemotherapy but HER2 negative subtype does not respond to chemotherapy. [14]. Thus, identifying clinically relevant subtypes of cancer plays an important role in designing, developing and improving more personalized and effective prognosis/treatment.

The genomic alterations in cancer cells have long been studied using low-throughput approaches, such as targeted gene sequencing, cytogenetic techniques [15], systematic mutagenesis [16], and genetic linkage analysis [17]. However, these traditional experimental approaches are tedious, time-consuming, and expensive [6]. Recent advancements in Next-Generation Sequencing (NGS) and high-throughput DNA sequencing techniques have revolutionized cancer genomics, and collaborative projects such as The Cancer Genome Atlas (TCGA) [18] and the International Cancer Genome Consortium (ICGC) [19] publicly released DNA sequences from thousands of tumors. These technologies along with emergence of hundreds of molecular markers have provided us with remarkable opportunity to study the molecular signatures of human cancers in a much more refined manner. In addition, large-scale cancer genomic studies have revealed wider genetic diversity in the same type of cancer as each subtype is characterized with unique behaviors in clinical and molecular profiles, such as survival rates, gene signature and copy number aberrations. Therefore, adequate methods are urgently needed to discover cancer subtypes to predict the biologic behavior and develop most effective therapeutic approaches based on high-throughput and high-dimensional data.

Cancer subtype discovery is to find out previously unknown subtypes. Several machine learning techniques have provided efficient solutions for cancer subtype discovery than those of histology based methods. In histopathology methods, cancer is detected by categorizing the image biopsy into cancerous or noncancerous by pathologists [20] and hence highly susceptible to human errors and bias. For identifying subtypes of adenocarcinoma, a type of lung cancer, the degree of agreement rate among independent pathologists was only 41% [21]. The fact that histology based methods are time consuming and not reliable [22] attracted machine learning approaches.

Machine learning is a method of data analysis that allows the automation of model building and learning with given input data. As of today, there are a lot of machine learning algorithms available to choose for given problem. Furthermore, microarray data analysis methods are becoming cheaper and generating enormous amount of data. Therefore, a number of machine learning approaches have been explored [23][24]. Combining Gene expression data with clustering based approaches revolutionized cancer subtype identification and provides an important insight in analyzed gene expression data.

In this thesis we proposed two novel pathway-based clustering methods based on unsupervised deep learning for discovering new cancer subtypes. Our methods incorporate prior biological knowledge of pathway data and simultaneously cluster samples into distinct groups. Integrating pathway database to clustering model itself helps us better select representative genes for clustering and thus generate biologically

informative clusters. These methods can more effectively identify clusters that are otherwise obscured by a large number of genes.

In this thesis, we used Glioblastoma Multiforme cancer (GBM) dataset in order to evaluate our models clustering efficiency. GBM is the most common and lethal primary brain cancer, which accounts for about half of all malignant primary brain tumors [25]-[26]. GBM have historically been viewed as a single pathologic entity but mounting evidence suggests that distinct glioblastoma substantially differ at the molecular level from one patient to another. The studies in this thesis therefore aim to identify various GBM subtypes that can lead to targeted therapy.

1.2 Problem Statement

The problem that we have addressed in this work is to identify subtypes of Glioblastoma Multiforme (GBM) by using pathway data in order to biologically interpret the identified clusters for developing personalized medicine. Also to resolve the challenges caused by using high dimensional low sample size (HDLSS) gene expression data.

1.3 Contributions

The main contributions of this thesis are:

- Proposed a new algorithm R-PathCluster, a pathway-based clustering method based on Restricted Boltzmann Machine for subtypes identification.
- Proposed a novel model SPACL based on Deep Belief Networks approach for finding subtypes by extracting complex nonlinear relations in pathways.
- Incorporating hierarchical representation of pathways.
- Finding biological relationship for identified subtypes.

1.4 Thesis Organization

The thesis is organized as follows. This thesis comprises six chapters

Chapter 1 introduces the thesis and states the problem definition.

Chapter 2 provides a brief description of the GBM data and their subtypes.

Chapter 3 describes related work in relation to our contributions.

Chapter 4 discuss about the machine learning techniques used in this thesis.

Chapter 5 introduces our proposed first novel approach R-PathCluster and its results and biological interpretation.

Chapter 6 presents another novel approach SPACL, based on deep neural net and elaborates on empirical results of the experiments that we conducted as well as the comparisons with other methods and discussions.

Finally, Chapter 7, states conclusion on this thesis.

CHAPTER II

BIOLOGICAL BACKGROUND

In this chapter we briefly introduce general information regarding GBM, its epidemiology, background, prevalence, incidence and its subtypes.

2.1 Glioblastoma Multiforme (GBM)

GBM is type of brain cancer which arise from brain supportive tissue called glial cells. These are highly aggressive tumors that rapidly infiltrate adjacent healthy brain tissue making very difficult to treat [27]. Although the etiology of glioblastoma has been thoroughly researched, immediate causes have not been found. In general, the causes of brain tumors are mostly unknown and hereditary factors can only be linked to approximately five percent of patients. Common factors that are thought to contribute to GBM development are genetic, environmental hazard. Due to the dismal prognosis, the World Health Organization (WHO) classifies GBM as a IV grade astrocytoma or high grade glioma (HGG) [28][25]. GBM is characterized by an extensive intratumoral heterogeneity which makes it extremely difficult to understand and treat, hence termed as ‘Multiforme’ [29][30]. The histopathological features of GBM include nuclear atypia, mitotic activity, cellular pleomorphism, microvascular proliferation, vascular thrombosis and necrosis [31][29]. This complexity, and putative cancer stem cell subpopulation combined with an incomplete atlas of epigenetic and genetic lesions, has contributed to make GBM complex disease [27]. These are most prevalent malignant tumors making up 54% of all gliomas and 16% of all primary brain tumors [32] and almost certainly lead to

a subject's death because of poor prognosis. It is estimated that 23,880 adults (13,720 men and 10,160 women) and 3,560 children will be diagnosed with a primary brain tumor and central nervous system (CNS) tumor and 16,830 (9,490 men and 7,340 women) people will be estimated to die in the United States in 2018 [33]. Glioblastoma patients have a median overall survival time of around 14 months [34][35]. With intensive therapy, the median survival rate can be extended up to 14.6, 16.1 or 16.8 months according to Phase three clinical trials published in the New England Journal of Medicine [36][37]. The five-year survival rate which tells what percent of people live at least 5 years after the tumor is found for primary malignant brain tumor is around 34% for men and 36% for women. Glioblastoma is a disease, which does not discriminate. It can occur at any age and to either gender but highest incidence rates can be seen in older males aged 45-65 than females [38]. In children they develop in between 5 – 9 years of age. Only 3-5% of Glioblastoma patients survive more than 3 years and are referred to as long-term survivors. Among all brain tumors, GBM show the greatest numbers of genetic abnormalities and considered as the highest-grade glioma with worst prognosis. The current standard treatment for this kind of cancer is the surgical resection followed by radiation therapy and chemotherapy [39].

2.2 Subtypes

Glioblastoma itself can be categorized based on histopathology into conventional (93%), gliosarcoma (2%), or giant cell glioblastoma (5%) [40]. GBMs can be divided into *de novo* primary GBMs and secondary GBMs [41]. Primary GBM (pGBM) that develops from glial cells is most common type but secondary GBM (sGBM) which progresses

from preexisting lower grade diffuse astrocytomas (grade II) or anaplastic astrocytoma (grade III) is less frequent. The pGBM is commonly seen in elderly patients and has a clinical history of less than 6 months while sGBM usually develops in younger patients [42]. Although both these subtypes don't differ morphologically and clinically but can be distinguished based on genetic alterations and deregulations of molecular pathways[28].

Based on progression and survival outcomes, The Cancer Genome Atlas (TCGA) and other groups have recently classified glioblastoma into gene expression-based subtypes. These include classical, pro-neural, neural, and mesenchymal [43]. Three subtypes were identified by array-based DNA methylation assay platforms in which one subtype formed a tight cluster with a highly characteristic DNA methylation profile "glioma CpG island methylator phenotype" or G-CIMP [44]. The G-CIMP subtype has high similarity for pro-neural subtype. The microRNA profiling study by Kim et al [45] identified five clinically and genetically distinct subclasses of glioblastoma based on neural precursor cell type. These include radial glia, oligoneuronal precursors, neuronal precursors, neuroepithelial/neural crest precursors and astrocyte precursors. Interestingly, when compared to the subclasses identified by Verhaak et al [43], the microRNA-based oligoneural, astrocyte, radial glial and subclasses were enriched in tumors from the proneural, mesenchymal and classical subtypes respectively.

Accurate prediction of cancer subtypes can aid in directing patient's therapies. Genomic techniques provide useful high-throughput tools for diagnosis and treatment of GBM cancer. The huge amount of genomic data and resources that have been generated, have

allowed researchers to use unsupervised machine learning techniques to establish distinct tumor subtypes. Also early diagnosis underlies every therapeutic strategy against GBM by improving the survival rate. We show that our approach lead to clusters of interest.

CHAPTER III

RELATED WORKS

Rapid advancement in high throughput technologies enables the measurement of thousands of gene expression data simultaneously. Hence many studies are applying clustering techniques on high throughput multidimensional genomic data to identify cancer subtypes. Hierarchical clustering algorithm is used on diffuse large B-cell lymphoma(DLBCL) cancer dataset to identify two subtypes that are distinguished from each other by the differential expression of hundreds of different genes, and these genes relate each subgroup to a separate stage of B-cell differentiation and activation [46]. As clustering algorithms like k-means, hierarchical were largely heuristic, model-based clustering methods were proposed to cluster gene expression data. In these methods, samples that are generated by multivariate normal distribution are clustered into best match distributions [47]. HMM-mix model was developed from model based clustering approach and identified cancer subtypes by analyzing comparative genomic hybridization (aCGH), which is data for DNA copy number alterations.

Ensemble clustering techniques are used on gene expression data for stable and better performance than single clustering techniques. This technique provides a clustering ensemble by either using different clustering algorithms like k-means, hierarchical clustering, spectral clustering, etc. or using single clustering algorithm with different parameters and initializations. Bagged clustering procedures are proposed to generate and

aggregate multiple clusterings on gene expression data of leukemia and melanoma cancers and identified 3 and 2 subtypes respectively [48]. Enhanced Maximum Block Improvement (eMBI), a new matrix factorization framework for biclustering identified five cancer subtypes in CRC and four subtypes in lung cancer [49].

More recently, co-clustering or bi-clustering [50][51] methods are also used to identify cancer subtypes by analyzing gene expression data. Clustering only in the sample space may fail to discover the patterns that a set of samples exhibit similar gene expression behaviors only over a subset of genes. Co-clustering simultaneously performs clustering on both samples and genes [52]. These algorithms can produce sets of genes that are co-regulated under sample subset. A network-assisted co-clustering method was developed to identify cancer subtypes (NCIS). In this method gene interaction network is combined with gene expression profiles in order to group genes and samples simultaneously into clusters which are biologically meaningful. This method divides samples into different clinical subtypes by assigning weights to genes depending on their connections in network [53]. Co-clustering ensemble is developed which is similar to clustering ensemble. This algorithm provides a framework by merging multiple base co-clustering results to generate a more stable and robust consensus co-clustering. Spectral co-clustering ensemble was proposed, which uses bipartite graph partition to leverage multiple base co-clustering [54].

PathCluster, a software package was developed to use gene sets as input data to identify cancer subtypes in conjunction with agglomerative hierarchical clustering algorithm. This

method also revealed unknown links between different annotation categories of clusters [55]. K-means algorithm was applied on Triple-Negative Breast Cancer (TNBC) gene expression data and identified seven subtypes and consensus clustering was employed to analyze the robustness of these subtypes [56]. Network based clustering approaches like 'Network-based Affinity Propagation' (NetAP) model identified uterine endometrial carcinoma and adenocarcinoma cancer subtypes [57]. In this method, subtypes were identified by using affinity propagation clustering algorithm on gene similarity matrix of patients which was computed by constructing gene interaction network with genes and patient's tumor profiles.

Multimodal Deep Belief Network (DBN) was proposed to identify cancer subtypes in breast and ovarian cancers [58]. In this model, three Restricted Boltzmann Machines (RBMs) are employed, one each for multi-omics data, such as DNA methylation, miRNA expression, and gene expression. The hidden layers of all the three RBMs were merged by using common hidden layer at the top. This model also identified miRNAs and key genes that possess specific roles in the pathogenesis of cancer subtypes. Subtypes of human colorectal carcinoma was identified by Multimodal Deep Boltzmann Machines (DBM). This method incorporates both clinical data and gene expression on joint RBMs [59]. iCluster algorithm distinguish three distinct GBM tumor subtypes by using joint analysis of DNA methylation, gene expression data, and copy number variation [60].

As the subtypes of cancer differ in network or pathway level, identifying subtypes by conventional clustering approaches based on gene expression is highly inadequate. Hence

we proposed two different pathway based clustering approaches to detect cancer subtypes. These two novel models have probabilistic neural network framework and use pathway data as prior biological knowledge in clustering the gene expression data of samples.

CHAPTER IV

DEEP LEARNING TECNIQUES

Deep learning, which is a branch of machine learning has more complicated algorithms that can model features with high-level abstraction from data. Currently, many deep learning methods have achieved success in computational biology. In this thesis we aimed at developing new approaches based on deep learning methods. Therefore, in this chapter we briefly introduce Restricted Boltzmann Machine (RBM) and Deep Belief Network on which R-PathCluster and novel methods are based on for clustering respectively.

4.1 Restricted Boltzmann Machine

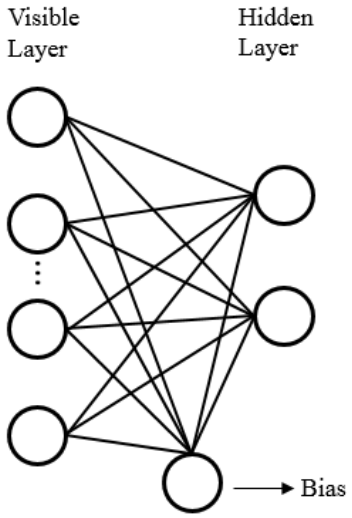

Figure 4.1: Graphical depiction of an RBM

Restricted Boltzmann Machine (RBM) is a probabilistic stochastic graphical model, which was initially introduced by Smolensky et al. in 1986, under the name Harmonium [61]. However, RBM came in to lime light in 2002 after Hinton invented a fast learning

algorithm to train them [62]. RBM has visible (observed) and hidden (latent) layers. It is a model especially made to learn a probability distribution over the inputs. RBM can be understood as a Markov random field (MRF) with latent factors that explains the input visible data using binary latent variables. A node in the hidden layer is a transformation of the visible layer, and each node in the visible layer is a transformation of the nodes in the hidden layer. In RBM, each node has a probability and a state and both are used during training. A graphical depiction of an RBM is shown in Figure 4.1. Because of absence of intra connections with in the layers, the visible units are conditionally independent given the hidden units, and vice versa, so that both conditional distributions are easily tractable. The hidden units represent the posterior distribution of variables in the visible layer.

Since the inter connections between the visible and latent hidden units form a bidirectional bipartite graph, a tractable learning algorithm is existed which trains a non-linear transformation function between these spaces. The non-linear transformation function is represented by the set of edges in the graph as well as two bias terms. The parameters of the transformation are computed by minimizing an energy function for the training set. If the energy increases, the probability that a generated visible value is represented by the data distribution decreases and vice versa. The energy function that is minimized in order to maximize the probability of the data is derived as:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i \in visible} c_i v_i - \sum_{j \in hidden} b_j h_j - \sum_{i,j} v_i h_j W_{ij}\,, \qquad (1)$$

where **v** = (v_1, v_2 … v_L) is the visible vector and **h** = (h_1, h_2 … h_K) is the hidden representation, $\theta = \{b, c, W\}$ are model parameters in which $c \in \mathbb{R}^L$ and $b \in \mathbb{R}^K$ are bias terms for visible and hidden layers respectively, $W \in \mathbb{R}^{LxK}$ is a weight matrix that defines a potential symmetric connections between visible input variables and stochastic binary hidden variables. The joint probability distribution over vectors **v** and **h** can be defined from energy $E(\mathbf{v}, \mathbf{h}; \theta)$ equation as:

$$P(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z(\theta)} e^{-E(\mathbf{v},\mathbf{h};\theta)}, \tag{2}$$

$$Z(\theta) = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h};\theta)}, \tag{3}$$

where $Z(\theta)$ is the normalizing constant or partition function that involves a sum of energies of all possible pairs of visible and hidden vectors. The probability of the visible vector is obtained by marginalizing over the space of hidden vectors as:

$$P(\mathbf{v}; \theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h};\theta)}. \tag{4}$$

From the above equations, the conditional distributions for visible and hidden vectors can be derived as:

$$P(\mathbf{h}|\mathbf{v}; \theta) = \prod_{j=1}^{K} p(h_j|\mathbf{v}), \tag{5}$$

$$P(\mathbf{v}|\mathbf{h}; \theta) = \prod_{i=1}^{L} p(v_i|\mathbf{h}). \tag{6}$$

Further simplified as:

$$p(h_j = 1|\mathbf{v}) = sig\left(\sum_i W_{ij} v_i + b_j\right), \tag{7}$$

$$p(v_i = 1|\mathbf{h}) = sig\left(\sum_j W_{ij}h_j + c_i\right), \quad (8)$$

where *sig(x)* = 1/(1+exp(-x)) is the logistic sigmoid function.

The resulted conditional probability for a node is called its activation probability. Sampling the activation probabilities leads to states for the nodes. For each node the sampling function g(x) is used to obtain the state of the node from its activation. For binary nodes, Bernoulli sampling is used to obtain states:

$$\mathrm{g(x)} = \begin{cases} 1, & if\ x > Unif(0,1) \\ 0, & otherwise \end{cases}, \quad (9)$$

$$h_j' = g(h_j), \quad (10)$$

$$v_i' = g(v_i). \quad (11)$$

For example, the state h_j is set to 0 if the computed activation h_j is lesser than a sample in a uniform distribution computed between 0 and 1. These states are only used for learning process.

4.1.1 Training

RBMs are defined as energy based models. The probability of each configuration for a model is inversely proportional to the scalar energy. Hence, the higher the energy for a given state configuration, the lower the probability that network will be found in that state and vice versa. In other words, RBM maximize the probabilities of the input samples (**v**) or minimize the negative log likelihood with respect to parameters. The derivatives of the log likelihood function with respect to the parameters are:

$$\frac{\partial}{\partial W_{ij}} logp(\mathbf{v}) = \underbrace{\mathbb{E}_{P_{data}}[v_i h_j]}_{\textbf{positive phase}} - \underbrace{\mathbb{E}_{P_{model}}[\hat{v}_i h_j]}_{\textbf{negative phase}}, \quad (12)$$

$$\frac{\partial}{\partial b_j} logp(\mathbf{v}) = \mathbb{E}_{P_{data}}[h_j] - \mathbb{E}_{P_{model}}[h_j], \quad (13)$$

$$\frac{\partial}{\partial c_i} logp(\mathbf{v}) = \mathbb{E}_{P_{data}}[v_i] - \mathbb{E}_{P_{model}}[\hat{v}_i], \quad (14)$$

where $\mathbb{E}_{P_{data}}[\cdot]$ represents an expectation with respect to the data distribution $P_{data}(\mathbf{h}, \mathbf{v}; \theta) = P(\mathbf{h}|\mathbf{v}; \theta)P_{data}(\mathbf{v})$, with $P_{data}(\mathbf{v}) = \frac{1}{(\mathrm{N})}\Sigma_{\mathrm{n}}\delta(\mathbf{v} - \mathbf{v}_{\mathrm{n}})$ which represents the empirical distribution, and $\mathbb{E}_{P_{model}}[\cdot]$is an expectation with respect to the distribution defined by the model. $\mathbb{E}_{P_{data}}[\cdot]$ can also be called as the data-dependent expectation, and $\mathbb{E}_{P_{model}}[\cdot]$ as the model's expectation.

In RBM visible states are conditionally independent of hidden states and vice versa. The conditionally independent distributions (**v**|**h**) and p (**h**|**v**) leads to Gibbs Sampling procedure, a Markov Chain Monte-Carlo (MCMC) algorithm. Gibbs sampling works by groups of two or more variables getting sampled and conditioned on all other group of variables. Thus, we sample (**v**|**h**) and p (**h**|**v**) by alternating between $v_i \sim \mathrm{p}(v_i|h_{j-1})$ and $h_j \sim \mathrm{p}(h_j|v_{i-1})$ and running a Markov chain to convergence.

In equation (12), two parts of the gradients are generally referred as the positive and negative phases. It is easy to compute gradients of the positive phase as it is conditioned on the value of a training sample. The gradients of "negative phase" is difficult (intractable) to compute as it requires infinite steps of Gibbs Sampling to reach the

stationary state of the model and to get optimal parameters. Therefore, the approximation method called contrastive divergence (CD) algorithm was proposed in learning procedure in order to efficiently perform gradient descent. The CD algorithm tries to fix the parameters so that the probability distribution represented by the network corresponds to the training data and so that the arrangement expresses the relations between input features well. After learning, the RBM provides a finite representation of the observation's distribution.

4.1.2 Contrastive Divergence

The CD algorithm is based on Gibbs sampling process, but does two specific optimizations. First, instead of starting at a random state of the visible units, it starts at the state of a training vector. Moreover, instead of waiting for the convergence of the Markov chain by taking infinite steps of Gibbs Sampling, it simply takes a limited number of steps (usually just one step) of Gibbs Sampling to approximate it. Figure 4.2 illustrate the Gibbs sampling step. The general idea behind CD is that even just a few steps of the Markov chain will provide a direction for the gradient in the state space for the Markov chain, and provide the training algorithm with the appropriate correction to the gradient. The CD algorithm can be done with a certain number K of steps of Gibbs sampling (called CD-K), typically one full step of Gibbs sampling. Empirical results show that CD algorithm with one full step of Gibbs sampling is an effective and efficient learning algorithm. This has the advantage of not requiring to perform alternating Gibbs Sampling for many iterations. Overall, these improvements are making CD a very fast algorithm. This learning rule of maximizing the log-likelihood over the data distribution is

equivalent to minimizing the Kullback-Leibler (KL) divergence between the data distribution and the equilibrium distribution of the model after Gibbs sampling.

$$\frac{\partial}{\partial W_{ij}}\frac{1}{m}\sum_{l=1}^{m} log\, P\left(\mathbf{v}^{(l)}\right) = \frac{1}{m}\sum_{l=1}^{m} v_i^{(l)}\, P\left(h_j = 1 \middle| \mathbf{v}^{(l)}\right) - \frac{1}{m}\sum_{l=1}^{m} \hat{v}_i^{(l)} P\left(h_j = 1 \middle| \hat{\mathbf{v}}^{(l)}\right). \quad (15)$$

$$\frac{\partial}{\partial b_j}\frac{1}{m}\sum_{l=1}^{m} log\, P\left(\mathbf{v}^{(l)}\right) = \frac{1}{m}\sum_{l=1}^{m} P\left(h_j = 1 \middle| \mathbf{v}^{(l)}\right) - \frac{1}{m}\sum_{l=1}^{m} P\left(h_j = 1 \middle| \hat{\mathbf{v}}^{(l)}\right). \quad (16)$$

$$\frac{\partial}{\partial c_i}\frac{1}{m}\sum_{l=1}^{m} log\, P\left(\mathbf{v}^{(l)}\right) = \frac{1}{m}\sum_{l=1}^{m} v_i^{(l)} - \frac{1}{m}\sum_{l=1}^{m} \hat{v}_i^{(l)}. \quad (17)$$

Notice that $\hat{\mathbf{v}}$ in these equations indicates the reconstructed visible units sampled from the m[th] step of Gibbs Sampling.

The CD-1 learning step for one sample can be summarized as follows:

- Initialize the visible units to a training sample $\mathbf{v}^0$.
- Compute the probabilities of the hidden units($\mathbf{h}^0$) and sample a hidden activation vector from $P(\mathbf{h}|\mathbf{v}^0)$.
- Calculate the outer product of $\mathbf{v}^0$ and $\mathbf{h}^0$ and call this the positive gradient.
- Compute the probabilities of the visible units ($\mathbf{v}^1$) from $P(\mathbf{v}|\mathbf{h}^0)$. and sample a reconstruction ($\mathbf{v}^1$) of the visible units.
- Resample the hidden activations $\mathbf{h}^1$ based on$\mathbf{v}^1$. (Gibbs sampling step).
- Calculate the outer product of $\mathbf{v}^1$ and $\mathbf{h}^1$ and call this the negative gradient.
- Update the weight matrix *W* based on the positive and negative gradients.
- Update the biases for visible and hidden layers.

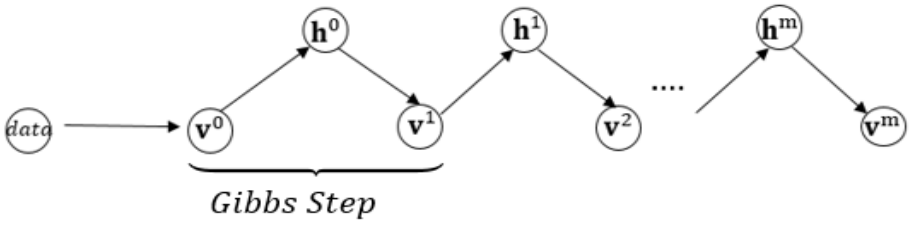

Figure 4.2: Gibbs Sampling

RBM Algorithm:

RBM Update (**v**, *W*, *b*, *c*)
v is a sample from training distribution
W is the weight matrix
b is bias vector for input units
c is bias vector for hidden units
η is learning rate for stochastic gradient descent in contrastive divergence
repeat {over the training dataset}
set $\mathbf{v^0} = \mathbf{v}$
compute the posterior Q = (P ($\mathbf{h}|\ \mathbf{v}^0$))
Sample $\mathbf{h^0}$ from Q
For k = 1 to m do
Sample $\mathbf{v^m}$ from p ($\mathbf{v}|\mathbf{h}^{m-1}$)
Compute the posterior Q = P ($\mathbf{h}\ |\mathbf{v}^m$)
Sample $\mathbf{h^m}$ from Q
end for
Update weights and bias with contrastive divergence
$W = W + \eta\ (\mathbf{h^0}.\ \mathbf{v^0}) - P\ (\mathbf{h^m} = 1|\mathbf{v^m}).\ \mathbf{v^m}$
$b = b + \eta\ (\mathbf{v^0} - \mathbf{v^m})$
$c = c + \eta\ (\mathbf{h^0} - P\ (\mathbf{h^m} = 1|\mathbf{v^m}))$
Until convergence

4.2 Gaussian-Bernoulli RBM

When the visible units are real values and hidden units are stochastic binary values, RBM is called Gaussian-Bernoulli RBM. The energy function of this model becomes

$$E(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{2\sigma^2} \sum_{i \in visible} (c_i - v_i)^2 - \sum_{j \in hidden} b_j h_j - \frac{1}{\sigma} \sum_{i,j} v_i h_j W_{ij}. \quad (18)$$

The probability distribution under this model is:

$$p(h_j = 1 | \mathbf{v}) = sig\left(\frac{1}{\sigma_i} \sum_i W_{ij} v_i + b_j\right), \quad (19)$$

$$p(v_i | \mathbf{h}) = \mathcal{N}\left(\sigma_i \sum_i W_{ij} h_j + c_i, {\sigma_i}^2\right), \quad (20)$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. Each visible unit conditioned on hidden states is modeled by a Gaussian distribution, whose mean is shifted by the weighted combination of the hidden unit activations. The derivative of the log-likelihood with respect to the model parameters takes a very similar form when compared to binary RBMs:

$$\frac{\partial}{\partial W_{ij}} \log p(\mathbf{v}) = \mathbb{E}_{P_{data}}\left[\frac{v_i}{\sigma_i} h_j\right] - \mathbb{E}_{P_{model}}\left[\frac{v_i'}{\sigma_i} h_j\right]. \quad (21)$$

4.3 Deep Belief Network

Deep Belief Networks (DBN) were first proposed in 2006 by Hinton [63]. It is referred as a stochastic generative model with several layers of hidden units. DBN can be used as supervised or unsupervised algorithm. There are connections between the layers, but not between nodes of the same layer.

DBN is a stacked architecture of RBMs. Deep belief networks similar to Restricted Boltzmann Machine, are probabilistic models that use hidden variables to learn features from the data. The structure of DBN can be seen in Figure 4.3. Unlike RBMs that learn features from input data and is limited in what it can represent, DBNs use multiple layers of hidden units, and gives more hierarchical structure and allows them to learn higher level representations.

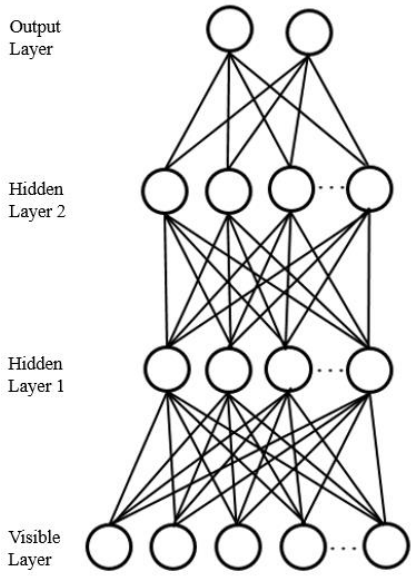

Figure 4.3: Structure of Deep Belief Network

4.3.1 Training

The training of the DBN has two phases: pre-training and fine-tuning. Hinton introduced greedy layer-wise training algorithm in pre-training to initialize parameters before performing any discriminative or generative fine-tuning. In pre-training, the layers of the

DBN are separated pairwise to form two-layered network called RBM. Each RBM is trained independently using contrastive divergence, such that the output or hidden layer of the lower RBM becomes input or visible layer for the next higher-level RBM and so forth such that each stack learns to encode the previous stack. This process can be repeated multiple times until all the layers in the stacked architecture are greedily trained stack by stack to learn complex probabilistic features. The goal of the pre-training process is to perform rough approximations of the model parameters, θ, learned by an RBM that define both p(**v**|**h**; θ) and the prior distribution over hidden vectors, $p(\mathbf{h}|\theta)$. Then the probability of generating a visible vector, v, can be derived as:

$$p(\mathbf{v}) = \sum_{\mathbf{h}} p(\mathbf{h}|\theta)\, p(\mathbf{v}|\mathbf{h};\ \theta). \tag{22}$$

After learning parameters, θ, $p(\mathbf{h}|\theta)$ can be replaced by learning a better model that treats the hidden vectors as the visible layer for another RBM. This replacement improves the variational lower bound on log probability of generating the training data under the composite model and guarantees the increase in performance.

Unsupervised algorithm employs only pre-training phase, while supervised uses both pre-training and fine tuning. In case of supervised method, once the network is pre-trained, the model looks like feed forward network and can be fine-tuned using a backpropagation algorithm such as SGD or CG. This better initialization allows for fast convergence and generally requires less refinements of the fine-tuning step.

In case the input data is real-valued, the first layer is then represented with a Gaussian-Bernoulli RBM and the rest of the layers with Bernoulli-Bernoulli RBMs.

DBN Algorithm (Unsupervised Greedy layer-wise training):

W^l is the weight matrix for layer l
v is the input training distribution for the network
b^l is the visible bias vector for layer l
c^l is the hidden bias vector for layer l
R is the no. of layers to train
for l = 1 to R **do**
    Initialize $b^l = 0$, $c^l = 0$, $W^l = 0$
  While not stopping criterion **do**
        Sample $\mathbf{h^0}$ from **v**
      for i = 1 to l **d**o
          Sample $\mathbf{h}^i$ from p ($\mathbf{h}^i \mid \mathbf{h}^{i-1}$)
    end for
      RBM Update ($\mathbf{h}^l$, W^l, b^l, c^l)
      end while
 end for

CHAPTER V

R-PATHCLUSTER

To solve the problem of finding cancer subtypes which can be biologically interpreted, we developed two different strategies. In the first approach, we propose a shallow net clustering method that use pathway markers which incorporates pathway dataset (i.e. the interactions between genes) as prior knowledge and cluster samples into distinct groups. In the second approach, we developed a deep net clustering approach that incorporates pathway data as hidden layer instead of directly integrating both gene expression and pathway data. Adding pathway knowledge to these methods will help in selecting genes which are representative for clustering and thus generate clusters that are biologically informative. In this chapter, we discuss the first proposed method, its experimental results and biological interpretation in detail. We describe latest model in next chapter.

5.1 Model

A novel pathway-based clustering approach, R-PathCluster was developed to detect cancer subtypes. Architecture of R-PathCluster is illustrated in Figure 5.1. R-PathCluster follows Gaussian RBM algorithm to obtain subtypes, and thus it has two layers: Input layer and Cluster layer. Input layer takes continuous values while the hidden or cluster layer units remain binary. The dataset used is pathway markers instead of gene expression data directly. Pathway based analysis assures robust and reproducible results whereas gene expression profile is not reproducible among datasets. Number of

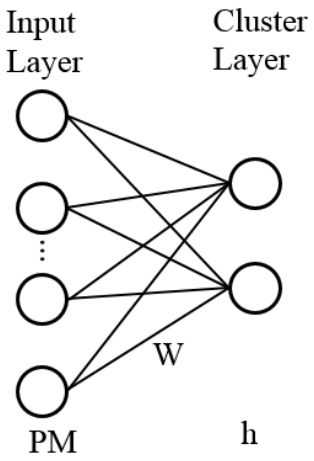

Figure 5.1: Architecture of R-PathCluster

nodes in input layer depends on number of pathway markers and nodes in cluster layer depends on number of distinct subtypes we want to identify in the dataset. An outline of R-PathCluster framework is illustrated in Figure 5.2.

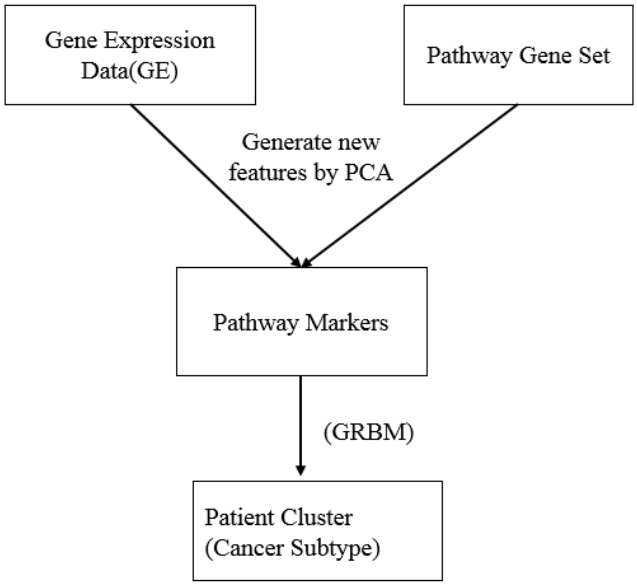

Figure 5.2: Schematic diagram of R-PathCluster

Pathway markers are generated from gene expression and pathway datasets. Let **G** = $\{\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \ldots, \boldsymbol{g}_n\}$ be the gene expression data of s numbers of samples, i.e., $\mathbf{G} \in \Re_{s\times n}$. From Gene Set Enrichment analysis (GSEA) which is an online pathway database, functional annotation of genes can be known. Initially a binary bi-adjacency matrix $\boldsymbol{A} \in \mathbb{B}_{p\times n}$ is created based on the biological relationship between *p* number of pathways and *n* number of genes. If a gene $\boldsymbol{g}_i$ is present in pathway p_j, then the element $\boldsymbol{A}_{ij}$ in the bi-adjacency matrix is indicated by one otherwise zero. Gene pathways are constructed by matching genes in all the pathways and genes from the gene expression data. Then *j*-th pathway includes a set of gene expressions:

$$p_j = \{\boldsymbol{g}_i | \forall i \in \mathbb{Z} : \boldsymbol{A}_{ij} = 1\}. \tag{23}$$

Thus all the pathways in our study will contain genes with gene expression data for all samples leading to multi-dimensional dataset. To deal with this problem, Principle Component Analysis (PCA) was applied to get linear combinations of gene expression data called pathway markers to fit our model. Based on the property of PCA, these pathway markers would not only reduce the dimension of gene expression data but keep most of the gene information.

The gene expression of the gene set on the *j*-th pathway is projected to represent a pathway marker of the pathway:

$$t_j = p_j a_j, \tag{24}$$

where a_j is first principle component obtained by principle component analysis:

$$a_j = argmax \frac{a_j^T p_j^T p_j a_j}{a_j^T a_j}. \tag{25}$$

Finally, using this equation, pathway markers are obtained for a pathway which represent the multiple gene expression data.

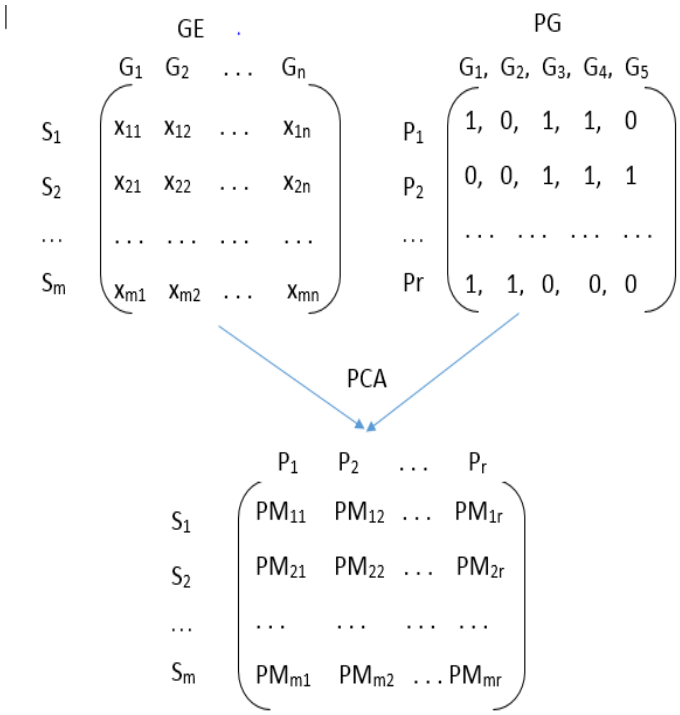

Figure 5.3: Generation of Pathway Markers using PCA

When R-PathCluster was trained with stochastic gradient and Hinton's contrastive divergence approach similar to RBM, it computes the posterior probability for all samples in the cluster layer. The nodes in the cluster layer depends on 'k' number of clusters. Then, an input data of pathway markers is assigned to a cluster that maximizes the posterior probability:

$$argmax\, p(h_j|t_j), \qquad j = 1, \ldots, k, \tag{26}$$

where

$$p(h_j|t_j) = sig(\sum_i W_{ij} a_i + b_j). \tag{27}$$

5.2 Experimental Results

We conducted experiments with different datasets using Gene expression data and pathway data of GBM patients to evaluate our method. Further we evaluated the performance of our model with other existing clustering methods such as hierarchical clustering, K-means, and general RBM models with different input data.

5.2.1 Datasets

Gene expression data of all GBM patients and pathway database are used to generate pathway markers from these datasets. Each dataset is discussed in detail in next sections.

5.2.1.1 Gene Expression

Because of wide range of applications of gene expression data in cancer diagnosis, prognosis, gene treatments, and other domains [64] [65], gene expression data analysis has been gaining lot of attention [66]. With the rapid development of high-throughput biotechnologies, it is easy to collect large amount of gene expression data with very low costs. Gene expression is the process that measures the gene activity level in the given tissue and thus provides information about the complex activities in the corresponding cells. This information is usually obtained by measuring the amount of generated messenger ribonucleic acid(mRNA) during transcription, a process that measures how active or functional the corresponding gene is [67]. As cancer is associated with multiple genetic and regulatory aberrations in the cell, these should reflect in the gene expression data. Biologists using microarrays measure gene expression levels under various specific

experimental conditions to analyze gene functions, regulatory mechanisms and cancer subtypes [68]. Microarrays generally permit to measure the expression levels of tens of thousands of genes simultaneously.

In this thesis, we used gene expression matrix of GBM patients obtained from The Cancer Genome Atlas (TCGA, available at https://cancergenome.nih.gov). The GBM dataset includes gene expression data of 12,042 genes for 522 samples and survival status and survival time. Patients who survived more than 24 months are considered as Long Term Survival (LTS) patients, and patients who are deceased and survived less than 24 months are considered as short term survival (non-LTS) patients. Patients whose survival time is less than 24 months and are still living are considered as censored patients. In our dataset, there are 99 LTS samples, 365 non-LTS samples and 58 censored samples. The gene expression data is normalized between zero and one.

5.2.1.2 Pathway Database

Standard microarray analysis treats every single gene equally, assuming their expression is independent from each other. As mentioned before, cancer is a complex web of interactions among multiple genetic and environmental risk factors. As a result, the phenotype or disease is not the outcome of single factor but multiple factors that influence each other. The assumption of independent gene expression significantly limits its capability of providing the underlying relationship among the listed genes or the biological mechanism that may cause the phenomenon or phenotype being studied. In order to gain better understandings on the mechanisms of gene regulation, researchers tried to integrate more

biologically meaningful information with microarray data. Hence pathway databases came in to lime light as these have more explanatory power than differently expressed genes and also reduce the complexity of analysis [69][70].

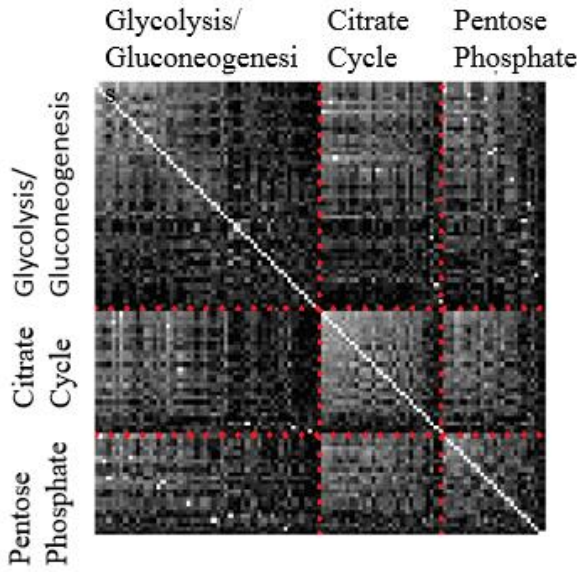

Figure 5.4: The square of Pearson correlations between genes

A biological pathway is defined as a series of molecular actions in a cell that leads to a certain product or a change in the state of the cell. Pathways can also turn genes on and off, or spur a cell to move. When something goes wrong in a pathway, the result can be a disease such as cancer or diabetes. Identifying the interdependencies and relation between the genes, proteins and other molecules and their casual effects in the interrupted mechanism can helps in developing personalized medicine for certain diseases. Each pathway consists of different set of genes that involved in specific outcome. There are many types of biological pathways. Among the most well-known are pathways involved in metabolism, gene regulation and signal transmission. For example, Glycogenolysis is a metabolic pathway that produce glucose from glycogen [71]. Pathway data for different

species is publicly available from databases in Molecular Signatures Database (MSigDB, available at http://software.broadinstitute.org/gsea/msigdb).

For GBM cancer, in our research we considered four pathway databases that include Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Pathway Interaction Database (PID) and, BioCarta. We showed the square of Pearson correlations between genes of three pathways glycolysis/gluconeogenesis, citrate cycle, and pentose phosphate in Figure 5.4, in which we noticed that genes with in the same pathways had higher correlations comparing to the genes involved in different pathways. In the Figure 5.4, the dotted lines in the red indicates the blocks of the correlations between various genes in a pathway.

5.2.1.3 Pathway Markers

Pathway markers were generated by using functional gene sets obtained from the four pathway databases from Molecular Signatures Database (MSigDB). We considered 1,273 pathway gene sets from overall pathway datasets. Each pathway has different set of genes. Gene expression matrix has 12,042 genes data and 522 samples. Among these genes, number of genes that have at least one functional annotation of pathways are 6,079. Then, Principle Component analysis (PCA) was applied on the gene expression data of1,273 functional gene set and by taking first principle component pathway markers are produced. The pathway markers obtained represent co-gene expression levels of the functional gene sets.

5.2.2 Experimental Setting

We followed a typical design of RBM network for R-PathCluster. Sigmoid function is applied for the activation. For the optimal model, we set hyper parameters like learning rate as 0.0003, Cluster layer nodes (k), ranges from 2 to 4. Number of input nodes will be 1,273. We updated the parameters using stochastic gradient descent (SGD)

5.2.3 Results

The R-PathCluster performance was evaluated by comparing with other clustering methods like hierarchical clustering, *k*-means clustering, and RBM models with different input data. In hierarchical clustering and *k*-means clustering, Ward's minimum variance method and Euclidean distance function are used respectively.

In order to compare the performance, the following three different types for input data were considered:

a) all gene expression data,
b) gene expression data of genes that are members of the functional gene sets of 1,273 pathways,
c) pathway markers.

The three types of the input data are annotated by the subscripts '**GE**', '**PG**', and '**PM**' respectively. We applied all clustering methods for these three types of input data. For easy interpretation we annotated *k*- means$_{\mathbf{GE}}$ as *k*-means clustering method with all gene expression data, while *k*- means$_{\mathbf{PG}}$ indicates a *k*-means clustering that takes only genes of 1,273 pathways.

In R-PathCluster method we used only Pathway Markers, while in other two RBM models, RBM_{GE} and RBM_{PG} 12,042 genes from gene expression data and 6,079 numbers of genes from functional gene sets of 1,273 pathways were introduced in the visible layer of these RBM models respectively while the hidden layer corresponds to the clustering.

5.2.3.1 Performance Metrics

The data should be normalized between zero and one for all the models. These clustering methods were repeated several times with multiple number of clusters in order to find the optimal number of clusters of the GBM samples. We carried out experiments for each model with given number of clusters for 10 times and selected the model with highest performance as optimal model. We analyzed the clustering performance with the different cluster numbers between two and four for all models. For clustering methods, performance is evaluated using Silhouette index. Therefore, in this research, we examined the average silhouette score for all the models. The silhouette score ranges from negative one to positive one, where a higher value represents better clustering. The silhouette index of the clustered data shows a degree of purity within a cluster and the quality of separation between clusters. Also in order to verify distinct subtype identification we employed Kaplan-Meier survival analysis. The survival plot of a clustered dataset demonstrates the survival distribution of each subgroup and distinct survival rates in each cluster would indicate tumor subtypes.

Further, we computed p-value using log rank statistical test in order to test significance between clusters. For more than two clusters, we considered the average p-value of pairwise log-rank tests as well as the minimal p-value for the statistical significance between clusters.

Experimental results for all methods with different cluster sizes are shown in Table1. We trained our model with 65,000 epochs until the model is converged with minimum cost score. Figure 5.5 displays learning curve of R-PathCluster for Two clusters. We indicated the highest values in the performance measurement and p-value (test of significance) with different levels of significance 0.05(*) and 0.01(**) in bold font. According to the results in the table, *k*-means$_{\mathbf{PM}}$ has the highest silhouette score of 0.2847 for 2 clusters, whereas our model R-PathCluster showed silhouette score of 0.1895 but interestingly it provides lowest p-value of 0.0017 in the Kaplan-Meier survival analysis, while for *k*-means$_{\mathbf{PM}}$ p-value is 0.0388. Figure 5.6 illustrates the Kaplan-Meier survival analysis for 2 to 4 clusters using R-PathCluster. Also the average survival months difference between the two clusters is largest for R-PathCluster with 4.7 survival months while for other methods the difference was around 3 months. These results show that R-PathCluster contributes a better solution in identifying subtypes that consider the survival rates when compared to other methods.

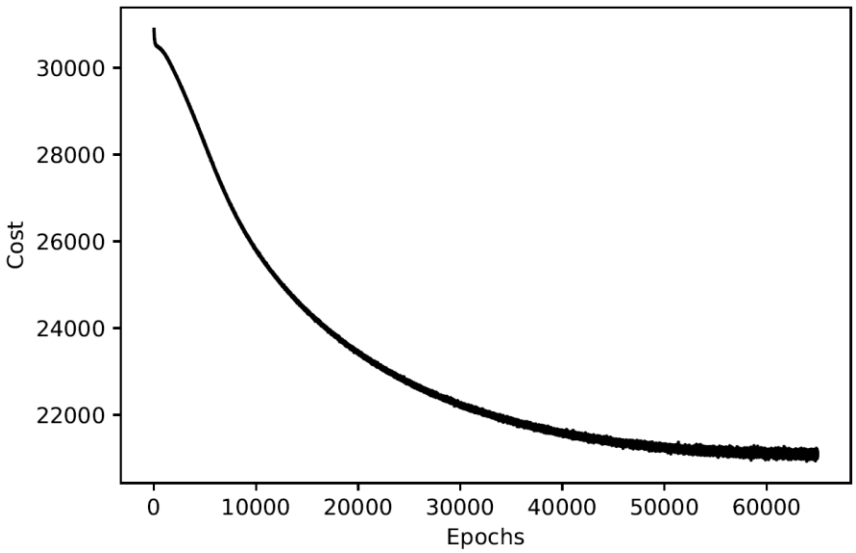

Figure 5.5: Learning curve of R-PathCluster for two clusters

We also conducted experiments for 3 and 4 clusters with all methods but the results are not that significant when compared with 2 clusters. This proves that the samples of GBM can be clustered into two groups as two clusters show higher silhouette scores and lowest p-value.

Table 1: Comparison of results of all methods using R-PathCluster

Tumor Subtypes	Methods	Mean(Survival months)	Silhouette Score	Average p-value	Min(p- values)
2	*k*-means$_{\mathbf{G}}$	17.8, 14.7	0.1049	0.0370*	0.0370*
	k-means$_{\mathbf{PM}}$	19.0, 16	**0.2847**	0.0388*	0.0388*
	Hierarchical$_{\mathbf{G}}$	14.4, 17.2	0.1169	0.0713	0.0713
	Hierarchical$_{\mathbf{PM}}$	15.4, 18	0.1637	0.0455*	0.0455*
	RBM$_{\mathbf{G}}$	16.8, 15.7	0.2568	0.5219	0.5219
	RBM$_{\mathbf{PM}}$	19.8, 16.2	0.214	0.1301	0.1301
	R-PathCluster	13.5, 18.2	0.1895	**0.0017****	0.0017**
3	*k*-means$_{\mathbf{G}}$	17.2, 20.8, 13.7	0.1051	0.0878	0.0052**
	k-means$_{\mathbf{PM}}$	17.4, 13.6, 21.7	0.2018	0.0579	0.0023**
	Hierarchical$_{\mathbf{G}}$	14.4, 16.7, 20.9	0.1213	0.0944	0.0225*
	Hierarchical$_{\mathbf{PM}}$	15.4, 19.7, 17.7	0.1556	0.3963	0.0502
	RBM$_{\mathbf{G}}$	22.3, 20.6, 16.3	0.2284	0.2927	0.2128
	RBM$_{\mathbf{PM}}$	16.4, 21.3, 18.1	0.2493	0.7056	0.5644
	R-PathCluster	12.9, 17.9, 15.2	0.1016	0.1851	0.0170*
4	*k*-means$_{\mathbf{G}}$	23.1, 13.7, 17.3, 15.7	0.1073	0.2286	0.0011**
	k-means$_{\mathbf{PM}}$	16.7, 13.3, 21.1, 18.8	0.1674	0.2139	0.0074**
	Hierarchical$_{\mathbf{G}}$	14.4, 16.7, 26.5, 16.3	0.1219	0.2814	0.0037**
	Hierarchical$_{\mathbf{PM}}$	16.8, 13.9, 19.7, 17.7	0.1419	0.3586	0.0186*
	RBM$_{\mathbf{G}}$	21.6, 16.4, 21.8, 17.6	0.2291	0.7142	0.3892
	RBM$_{\mathbf{PM}}$	28.7, 20.1, 29.8, 16.1	0.1864	0.4444	0.1534
	R-PathCluster	13.1, 13.2, 14.6, 18.1	0.0839	0.4168	0.0054**

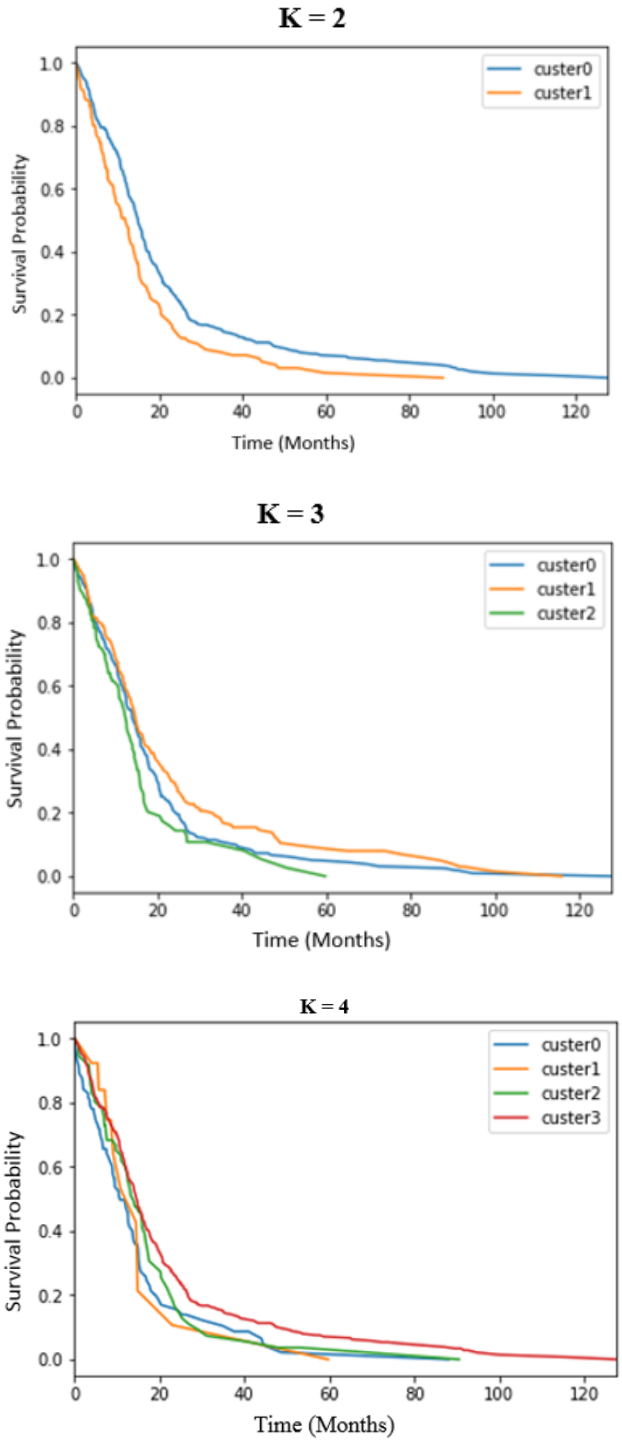

Figure 5.6: Survival plots of R-PathCluster for k=2,3,4

In Figure 5.6, we can interpret when k = 2, the survival probability of the cluster 0 shows significantly greater than the cluster 1.

5.2.3.2 Comparison of Age Distributions

To compare difference of mean ages between clusters we applied one-way ANOVA test. One tailed test is preferred to increase detection power. Figure 5.7 denotes boxplot of age distributions of clusters when k=2. As the ANOVA test p- value is 2.7e-3 which is less than 0.05, we can conclude with 95 % confidence interval that subgroups we identified are significantly different in terms of age.

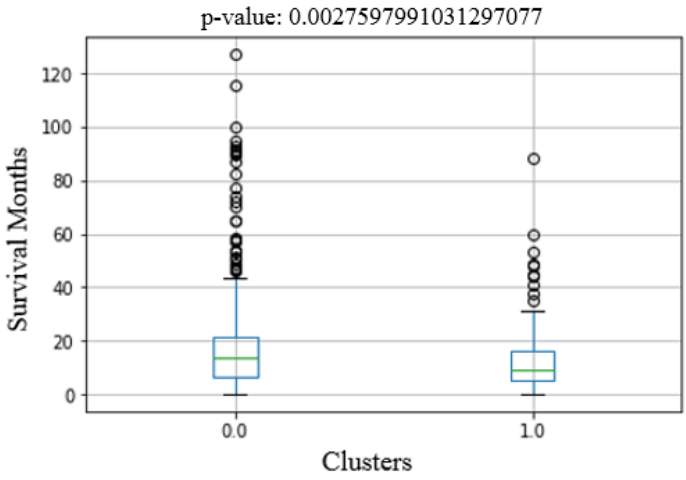

Figure 5.7: ANOVA comparison of age for k = 2

In Figure 5.7, y-axis represents patient survival months; x- axis represent cluster ids. Start edge and end edge of a boxplot shows months range for each cluster and middle line in the box indicates mean value of patients in the cluster.

5.3 Discussion

Although R-PathCluster yield most promising results, we further analyzed its contribution in interpreting the identified clusters biologically. In this section, we demonstrate a biological mechanism inferred by R-PathCluster by examining the coefficient values. First we examined the pathway markers that has the highest coefficient values and top five pathways for each cluster are identified and shown in Table 2. From KM survival analysis, we identified that cluster 0 represents LTS cluster and cluster 1 represents non-LTS cluster.

Table 2: Top five pathways for each cluster using R-PathCluster

Cluster	Pathways	*W* in clust	*W* in clust	Genes
0	cytochrome P450	0.5469	0.4004	91
	ER-Phagosome pathway	0.5468	0.0890	90
	p53-dependent G1 DNA damage response	0.5440	0.1948	29
	PDGFR-alpha signaling pathway	0.5278	0.2455	14
	nitric oxide signaling pathway	0.4975	0.3631	17
1	sphingolipid metabolism	0.1502	0.5757	38
	Regulation of RhoA activity	0.2724	0.5575	118
	PPAR	0.2214	0.5247	11
	IL22 soluble receptor signaling pathway	0.2454	0.5072	11
	wnt signaling pathway	0.1938	0.5031	39

According to the weights in the table, the five pathways, cytochrome P450 pathway, ER-Phagosome pathway, p53-dependent G1 DNA damage response pathway, PDGFR-alpha

signaling pathway, and nitric oxide signaling pathway belong to the LTS cluster (cluster 0), while sphingolipid metabolism pathway, Regulation of RhoA activity pathway, Peroxisome proliferator-activated receptors alpha (PPARA) pathway, IL22 soluble receptor signaling pathway, and wnt signaling pathway belong to the non-LTS cluster (cluster 1). Specifically, sphingolipid metabolism pathway produces sphingosine kinase-1 enzyme which produce sphingosine 1-phosphate (S1P) that activate growth and invasiveness of glioma cells. Higher expression levels of the sphingosine kinase-1 enzyme is correlated with short term survival rate of glioblastoma patients [72]. Over expression of p53 gene is reported in LTS glioma patients than non-LTS because of its lower proliferation rate compared with typical GBM survivors [73]. High expression of Wnt signaling/β-catenin recorded short survival and poor prognosis in GBM patients [74], whereas amplification of PDGFR-alpha did not alter the survival rate of GBM patients [75].

CHAPTER VI
SPACL: Sparse Pathway-based Clustering for Cancer Subtypes

6.1 Model

In this chapter we discuss our proposed Sparse PAthway based CLustering (SPACL) model to find biologically interpretable cancer subtypes by clustering pathway data. The architecture of our model consists of input layer, three hidden layers and final cluster layer. The first layer called as input layer that represents the gene expression data. Second layer is called pathway specific hidden layer that incorporates pathway data. Two hidden layer that represents hierarchical relationship among pathways and first hidden layer and finally a cluster layer that corresponds the subtypes of cancer. In this model we also considered sparsity between and in the layers which gives us good interpretability for our model. The architecture of our model can be seen in Figure 6.1.

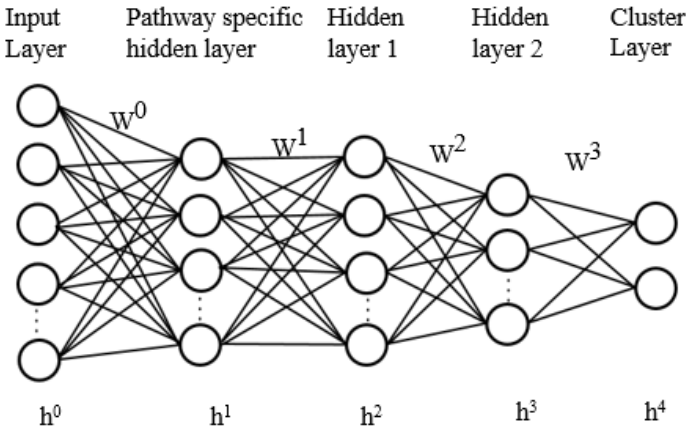

Figure 6.1: Architecture of our SPACL model.

6.1.1 Input Layer

Input layer corresponds to the gene expression of given cancer data. Number of nodes in input layer varies with respect to number of genes in the dataset. Each node corresponds to one gene feature. Assume that L number of genes present in at least one pathway. So the nodes in input layer for our model will be L.

6.1.2 Pathway specific hidden layer

This layer corresponds to biological pathway dataset. Each node in this layer represents an individual pathway. The pathway data is derived from Reactome pathway database. Pathway databases has associations between pathways and genes. Each pathway has set of genes and each gene can be associated with multiple pathways. The connections between input layer and pathway specific hidden layer is interpreted as the biological relationship between the genes and pathways. Thus, the pathway specific hidden layer makes it possible to interpret the model as pathway based analysis. In order to represent the connections between L genes in input layer and P^{R} pathways in pathway specific hidden layer, binary bi-adjacency matrix $\mathbf{A} \in \mathbb{B}^{P^{R} \times L}$ was created. An element $\mathbf{A}_{ij}$ is set to one if gene j belongs to pathway i, otherwise zero. This bi-adjacency matrix is used to model the sparsity between input and pathway specific hidden layer.

6.1.3 Hidden Layer

The hidden layers in general represents complex nonlinear associations embedded in the input data, such that different hidden layers try to capture different levels of complexity. In

our proposed SPACL model, hidden layer encodes activation states of combination of pathways. Sparsity is applied between the pathway specific hidden layer and hidden layer using mask matrix which enables to interpret the relationship. Selection of optimal number of hidden layers and number of nodes in the hidden layer is important as it impacts the performance of the model. In general, adding more hidden units improves the model representational power. However, larger number hidden nodes increase the risk of over fitting as the learned features becomes strongly correlated and run time efficiency decreases because of too many parameters.

6.1.4 Cluster layer

This layer encodes posterior probability for each cluster. Number of nodes in this layer corresponds to number of clusters. By using argmax function, each sample is assigned to a cluster that has highest posterior probability.

6.1.5 Sparsity

In our model sparsity between layers is regulated by using jointDrop which includes dropout and DropConnect. Dropout is generally used to reduce overfitting. The dropout module executes the idea of randomly eliminating hidden neurons according to a dropout ratio φ and rebuilds a new small sparse network with leftover hidden nodes and all visible nodes. When dropout is applied the conditional distributions for input and hidden nodes will be:

$$p(h_j = 1|\mathbf{v}, d) = d_k . sig\left(\sum_i W_{ij} v_i + b_j\right), \quad (28)$$

$$p(v_i = 1|\mathbf{h}, d) = d_k . sig\left(\sum_j W_{ij} h_j + c_i\right), \quad (29)$$

where d is the binary vector $d \in \{0,1\}$. Each random variable d_k takes the value 1 with dropout ratio φ, independent of others. If d_k takes the value 1, the hidden unit h_j is retained, otherwise it is dropped from the model.

DropConnect introduces dynamic sparsity by dropping connections between hidden layers. The sparsity is determined by mask matrix **M** on the connections between hidden layers as:

$$\mathbf{h}^{(l+1)} = f\left((W^l \star \boldsymbol{M}^l)\mathbf{h}^l + b^l\right), \quad (30)$$

where $\star$ represents element wise multiplication and $f(.)$ is non-linear activation function. $\mathbf{h}^l$ is the output of the feature vector of layer *l*, W^l and b^l are fully connected weight matrix and bias of the corresponding layer respectively. The mask matrix **M** is a binary matrix encoding the connection information. This matrix **M** is generated with respect to a sparsity level (*r*)which indicates the proportion of weights to be dropped in each layer. In mask matrix, an element $\mathbf{M}_{ij}$ is set to one if the absolute value of the corresponding weight W_{ij} is greater than some threshold Q, an r-th percentile of absolute values of weights W; otherwise element is set be zero. Sparsity level (r) value ranges between 0 and 100, where 0 indicates fully connected neural network while 100 is for no connections between hidden layers. During training for each layer, the optimal sparsity level (r*) is approximated for each iteration. To obtain optimal sparsity, cost function is computed for each iteration with different sparsity levels and by applying cubic-spline

interpolation to the cost scores, the sparsity level that has minimum cost score can be selected as optimal sparsity. For this we assume cost function is continuous with respect to sparsity. The individual setting of the sparsity on each layer shows different levels of biological associations between genes and pathways.

Thus, in our model, we employ two different types of mask matrices. The mask matrix $\mathbf{M}^{(0)}$ between input layer and pathway specific hidden layer is determined by the binary bi adjacency matrix based on relationship between genes and pathway database where as for other layers it is determined by sparsity level. Thus, mask matrices are formulated as:

$$\mathbf{M}^{(l)} = \begin{cases} 1\left|W^{(l)}\right| \geq Q^{(l)}, & Q^{(l)} = Q^{(l)}Q^{(l)}_{100(1-r)\%}\left(\left|W^{(l)}\right|\right),\ if\ l\ \neq 0 \\ \mathbf{A}, & if\ l = 0 \end{cases}. \qquad (31)$$

i.e., $Q^{(l)}$ is $100(1-r)\%$ -th left percentile of $\left|W^{(l)}\right|$ if $l\ \neq 0$.

In addition, in order to reduce overfitting, we also added L2 regularization or weight decay that penalizes the quadratic values of weights in to objective function and decrease the noise. An additional term is added to the cost function to account for weights that have grown larger. So the cost function, C becomes:

$$C = \sum\left(\mathbf{h}^{(l+1)}\mathrm{W}^{l} - \mathbf{h}^{l}\right)^{2} + \frac{1}{2}\lambda\|\mathrm{W}^{l}\|^{2}, \qquad (32)$$

where λ is the regularization parameter used to control just how quickly the weights decay. Changing the size of λ can shift the priority of the minimization function from the original cost function (better modeling the distribution of the dataset) to ensuring that the weights

stay small. As is implied by the name of this type of regularization, the modified cost function does not take into account the biases.

6.1.6 Training

The main advantage of our model when compared to other conventional models is initializing small sub network instead of whole network which improves performance and reduce computational complexity in case of High Dimension Low Sample size (HDLSS) data.

When the model is constructed at first we initialize the connections between input layer and pathway specific hidden layer with prior biological knowledge of pathways. Binary bi-adjacency matrix, **A** generated from gene expression and pathway database is used in determining the active and inactive connections between input and pathway specific hidden layer. Active connection weights and bias are initialized with random values while inactive connection weights are set to zero. As our model is based on deep belief network, it follows greedy layer wise training algorithm. Hence the first two layers input and pathway specific hidden layer forms two layered network called RBM and trained independently using contrastive divergence until the model is converged. During training, we applied dropout technique so that sub network is selected. Then the output or hidden layer of the this RBM becomes input or visible layer and with hidden layer 1 form new RBM and follows same training procedure. The sparsity is introduced in the nodes by drop out and in weighted connections between layers by mask matrix generated from cubic-spline interpolation to the cost function. This process can be repeated multiple times for remaining layers and

finally can cluster the data based on the probability values at cluster layer. Figure 6.2. depicts the training of our model. The weighted connections between input layer and pathway specific hidden layer are invariant over the entire training and can be seen in Figure 6.2A and applying dropout can be seen in Figure 6.2B. Figure 6.2C explains training of our model with sparsity and following greedy layer wise approach.

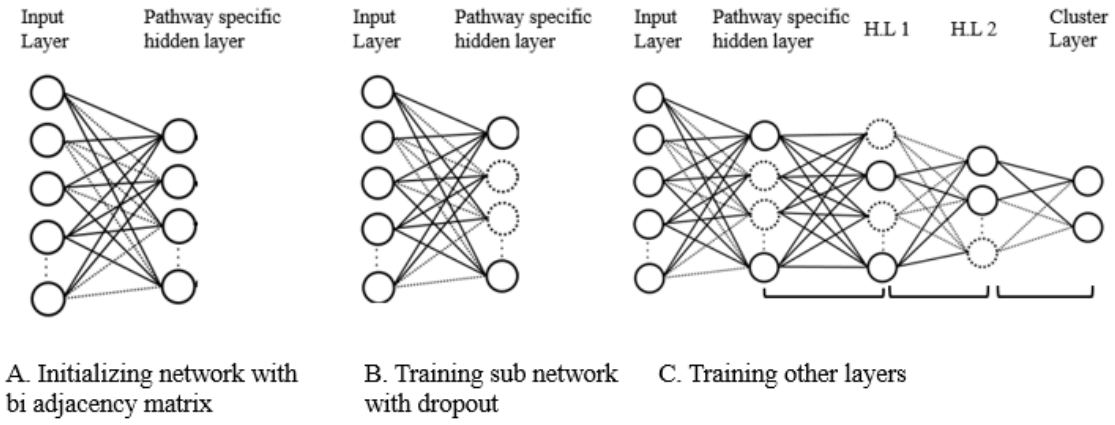

Figure 6.2: Training of SPACL

6.2 Experimental Results

We applied our model on GBM cancer patient's dataset and compared the performance of our model with other clustering methods.

6.2.1 Datasets

For the pathway based analysis, we used pathways from 'Reactome' pathway database which include 674 pathways. For input features, we considered only those genes that belong to at least one pathway in Reactome pathway database. Thus the nodes in the input

layer will be 4,362 genes, that belongs to at least one pathway and number of nodes in pathway specific hidden layer will be 674. Then we constructed mask matrix **M** between input layer and pathway specific hidden layer that has dimensions of (4362 x 674). Note that data should be normalized to mean of zero and a standard deviation of one.

6.2.2 Experimental setting

Our model consists of input layer, pathway specific hidden layer, hidden layer and cluster layer. Final frame work of our model consists of 4,362 input nodes, 674 pathway nodes, 400 hidden layer 1 nodes, 100 hidden layer 2 nodes and 2 cluster layer nodes (varies depending on cluster size). We conducted the experiment for different cluster numbers ranging from two to five. Sigmoid function and mean square error were considered for activation and cost function respectively. For optimal model, we empirically determined hyper parameters from multiple experiments and set learning rate as 0.5 and λ as 1e-4, dropout ratio for all the layers as 0.7.

6.2.3 Results

We evaluated our model by comparing the performances with other clustering methods like hierarchical, k-means, spectral, RBM and DBN models. For all the models, we used 4,362 genes, that belongs to at least one pathway as input.

Depending on the initial values, the clustering methods may produce different results. Since the RBM and DBN models follows the stochastic gradient descent approach, the initial values play an important role in the performance. Hence the optimal models of the

clustering methods were obtained from the best of 10 replications with random initial values.

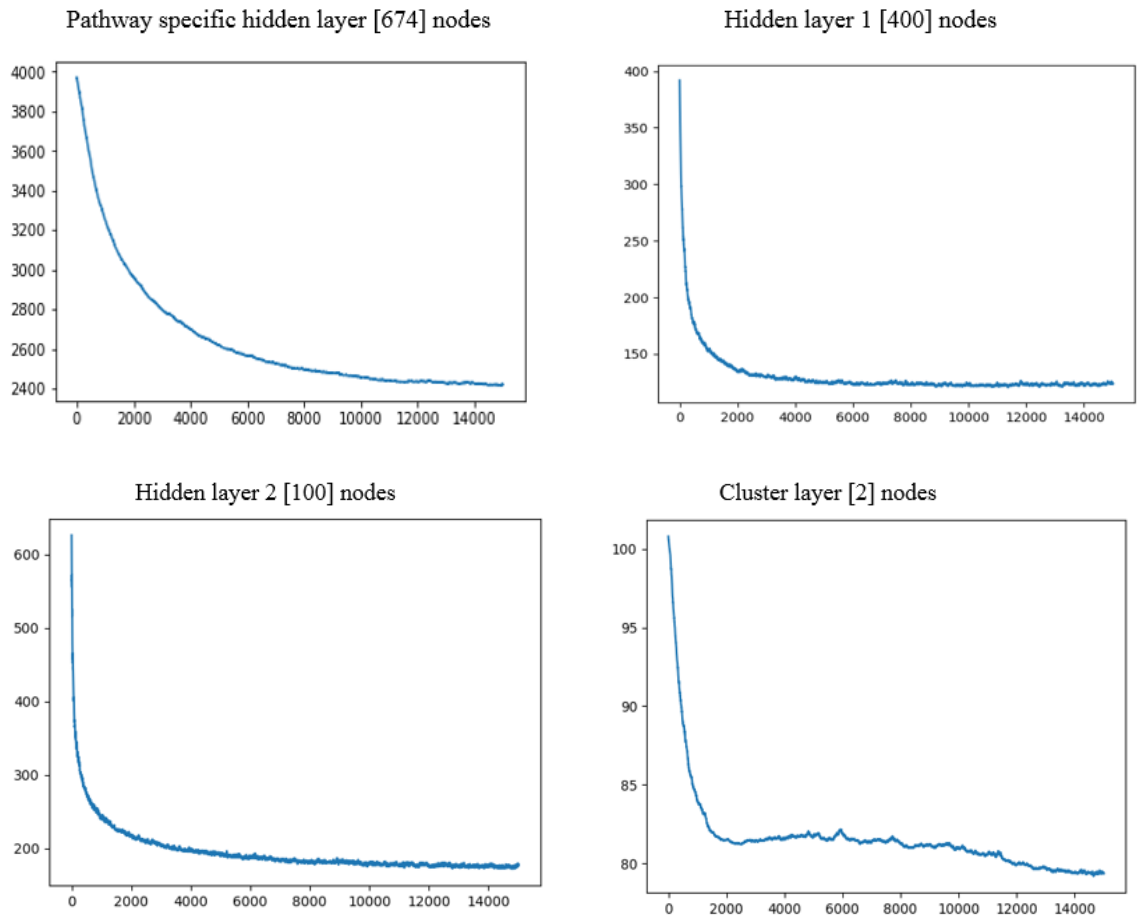

Figure 6.3: Learning curves for all layers

We repeated the experiment with various number of clusters in order to find the optimal number of clusters of GBM samples. In this study, we evaluated the clustering performance with the cluster number between two to four by silhouette index and Kaplan-Meier survival analysis.

The average silhouette score and p-value of all methods are shown in Table 2. We trained our model with 15,000 epochs for each layer. Learning curves for each layer can be seen

in Figure 6.3 for Two clusters. According to the results in the table, our model outperforms other models with the highest silhouette score of 0.18658 and lowest p-value of 0.0026 in the Kaplan-Meier survival analysis. The Kaplan-Meier survival analysis for 2 to 4 clusters using our model is depicted in Figure 6.4. Also we performed ANOVA test to compare the age distribution between clusters and that results for two clusters can be seen in Figure 6.5. When k = 2, the difference of the average survival months between the two clusters is largest (7.49) for our proposed SPACL model while for other methods the difference was less than 2 months. From these results it is clearly evident that our model provides a better solution for identifying subtypes that consider the survival rates.

We also conducted experiments for three and four clusters with all methods but the results are not that significant when compared with two clusters. Figure 6.6 illustrates the silhouette plot for 2 to 4 clusters. This proves that the samples of GBM can be clustered into two groups as two clusters show higher silhouette scores and lowest p-value.

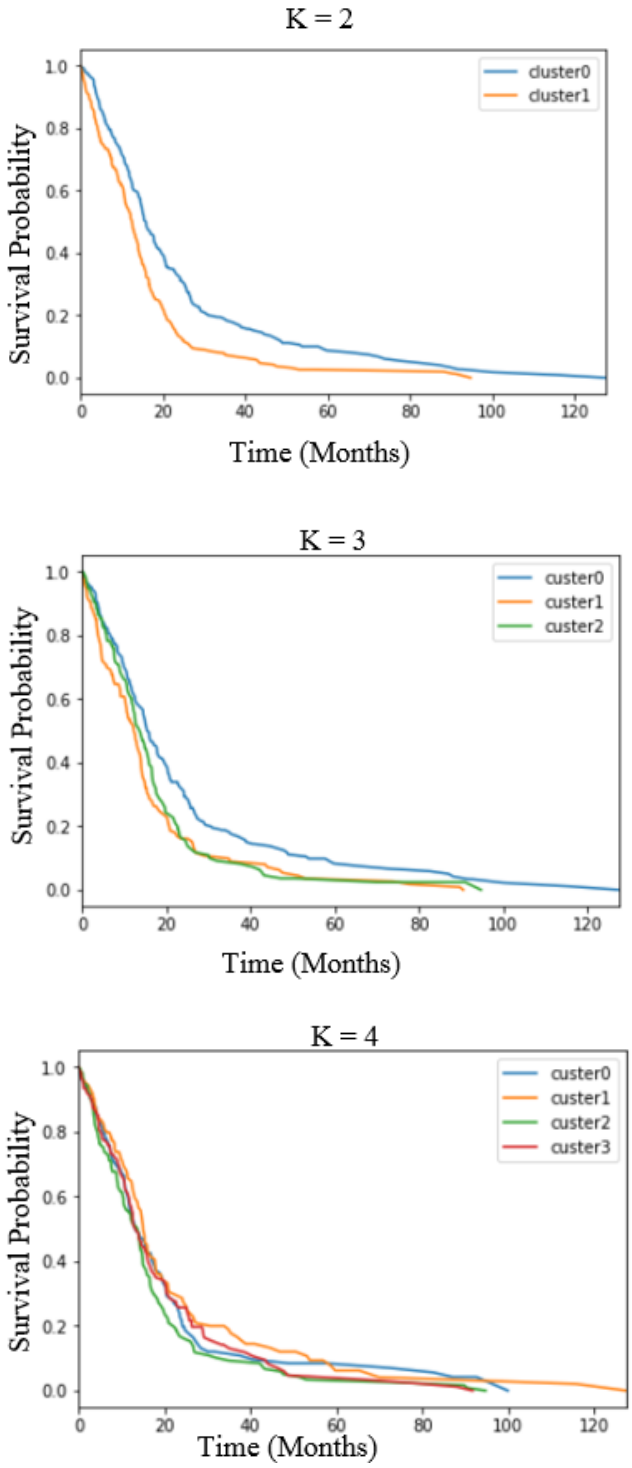

Figure 6.4: Survival plots of SPACL for K=2,3,4.

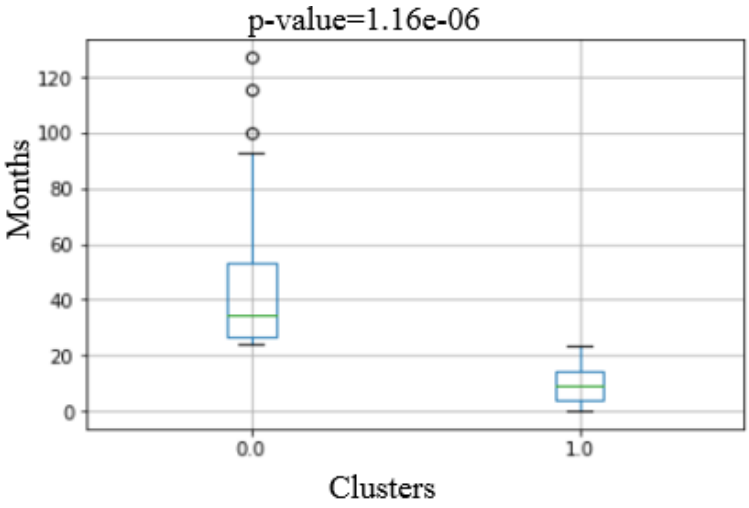

Figure 6.5: ANOVA comparison of age for k = 2

Table 3: Comparison results of various methods using SPACL

Clusters	Methods	Silhouette	p-value	Min(p-values)
2	k-means	0.09400	0.022539*	0.022539
	Hierarchical	0.08734	0.009182*	0.009182*
	Spectral	0.00267	0.381899	0.381899
	RBM	-0.00181	0.549138	0.549138
	DBN	0.00568	0.329547	0.329547
	SPACL	**0.18658**	**0.000002****	0.000002**
3	K-means	0.08785	0.05755	0.004974*
	Hierarchical	0.09652	0.035002*	0.006188*
	Spectral	-0.02309	0.358068	0.156878
	RBM	0.00153	0.487635	0.294562
	DBN	0.00736	0.367452	0.125843
	SPACL	0.10415	0.067926	0.000277**
4	K-means	0.08899	0.174404	0.001720*
	Hierarchical	0.09894	0.129076	0.005093*
	Spectral	-0.00972	0.381114	0.195027
	RBM	0.00896	0.468135	0.293485
	DBN	0.01736	0.185236	0.007958*
	SPACL	0.11528	0.229467	0.016398*

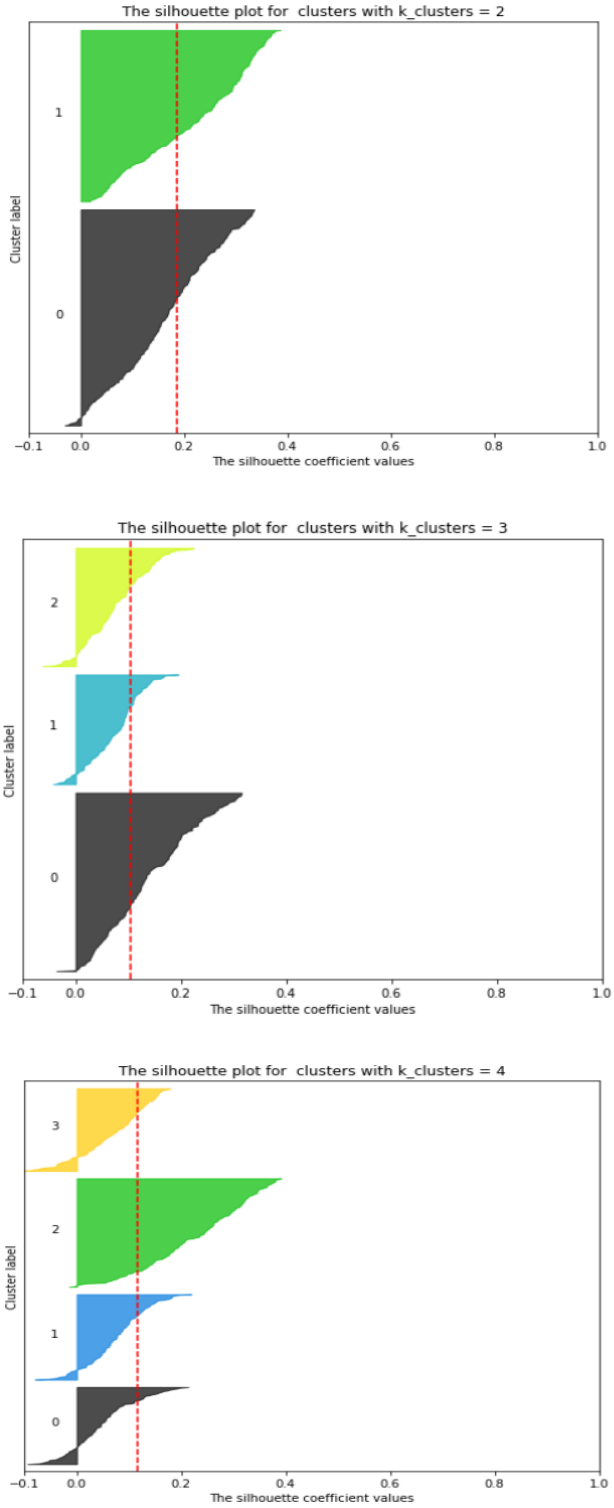

Figure 6.6: Silhouette Analysis for proposed SPACL model for k = 2,3,4

6.3 Discussion

We analyzed the experimental results biologically which is a noticeable contribution of our novel model. We examined the coefficient values between last hidden layer and cluster layer. Figure 6.7 depicts the heat map of absolute weight values after sorting, where the weighted connections which are dropped are colored in weight. This image discloses the distinct patterns of weights for two neurons in cluster layer.

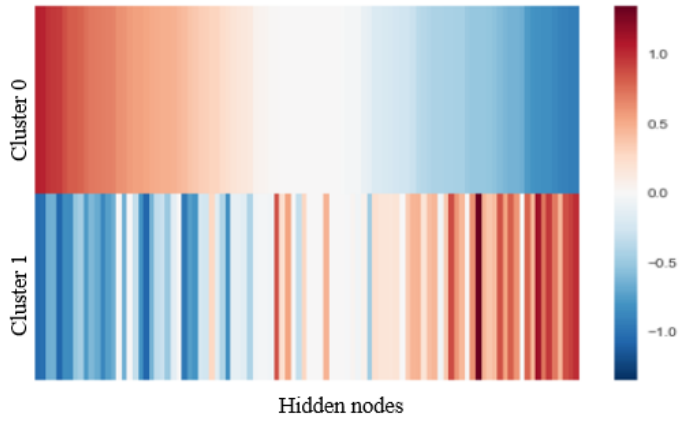

Figure 6.7: Weights between the hidden layer and cluster layer

Further, we analyzed the posterior probabilities of output nodes, which are exhibited in Figure 6.8. This figure reveals that the samples belong to cluster 0 has higher posterior probabilities for output node 0 when compared to node 1 and vice versa for samples belong to cluster 1. The top five pathways for each cluster are shown in Table 4. Pathways are identified based on the larger weighted connections between hidden layer and pathway specific hidden layer. From our model we find the pathways like metabolism of proteins, Regulation of kit signaling, p53 dependent G1 dna damage response, signaling to p38 via RIT and RIN and Hemostasis contributed to cluster 0

(LTS), whereas PI3K AKT activation, signaling by wnt, signaling by constitutively active EGFR, signaling to RAS, TGF beta receptor signaling activates SMADS pathways involved in cluster 1 (non-LTS).

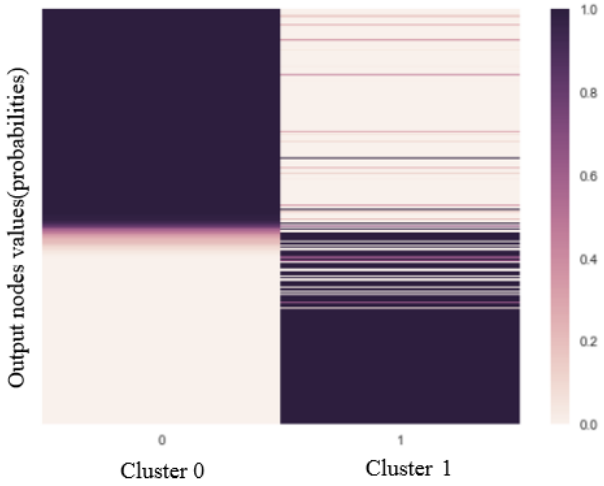

Figure 6.8: The output node values for two clusters

Table 4: Top five pathways for each cluster using SPACL

Cluster	Pathways	Pathway size	Genes
0	Metabolism of proteins	518	335
	Regulation of kit signaling	17	14
	p53 dependent G1 dna damage response	57	51
	signaling to p38 via RIT and RIN	15	13
	Hemostasis	466	397
1	PI3K AKT activation	38	29
	signaling by wnt, ,	65	60
	signaling by constitutively active EGFR,	18	16
	signaling to RAS	27	24
	TGF beta receptor signaling activates SMADS	26	22

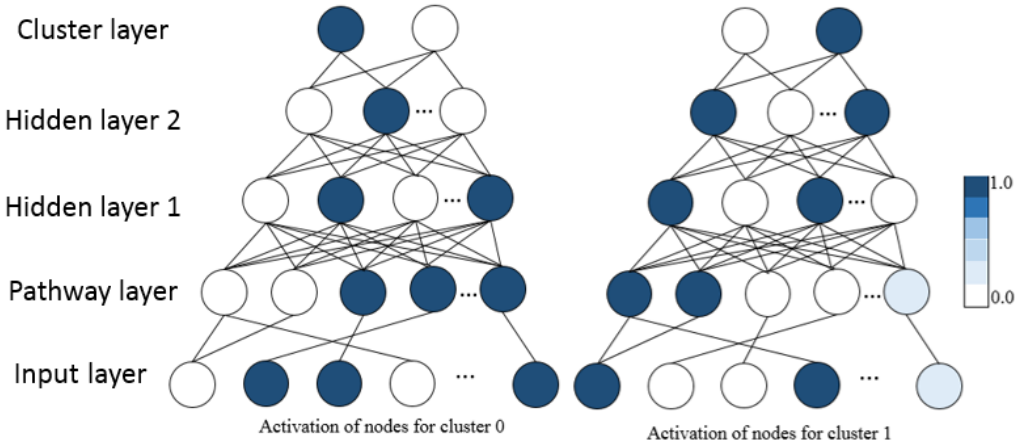

Figure 6.9: Hierarchical representation of pathways

Biological literature shows that all these pathways are significant in providing personalized medicine by targeting these pathways and corresponding genes. Figure 6.9 illustrates training of our model by activating the pathways and hidden nodes by corresponding genes. Metabolism of proteins and KIT signaling pathway activate hidden nodes one and three in hidden layer one. Further these two nodes will activate hidden node one in hidden layer two and in turn activates the output node 0 in cluster layer. This represents that the activation levels of these two pathways will be in higher percentage in long term survival patients compared to short term survival patient's type.

CHAPTER VII

CONCLUSION

Cancer is caused by multiple genomic alterations or dysfunction in molecular systems. It is a complex and heterogeneous disease with multiple subtypes and lack of knowledge on subtypes hinders developing effective targeted therapies and bringing in the personalized medicine objective. With the rapid advancement in micro array and next-generation sequencing (NGS) technologies, it is possible to analyze a large cohort of patients and record patient genomic alterations and expression dysregulations. This accelerates many opportunities in redefining cancer subtypes.

The widely adapted approach in finding subtypes is applying unsupervised techniques on genomic data of patients. The clusters are deemed interesting if they are found to be associated with a clinical variable of interest. In this work, we aim at developing methods that can identify cancer subtypes that can be biologically interpreted from high-throughput gene expression data. We proposed two different new clustering approaches that incorporates the pathway information to detect biologically informative sample subtypes. In the first approach, we developed R-PathCluster model in which pathway markers are used as input data instead of gene expression data and showed promising clustering performance. In the second approach we proposed SPACL model that takes leverage of prior biological knowledge of pathway database. We added pathway data as separate layer to compute hierarchical complex nonlinear representations between pathway and hidden layers. For both models, we limited our analysis to GBM cancer data but can be used to

detect any cancer subtypes. Our models R-PathCluster and SPACL showed the significant difference in terms of survival rates by Kaplan-Meier survival analysis for two clusters. The results for silhouette score, KM analysis, ANOVA test prove that our models outperform the other conventional clustering methods. Though both methods can detect cancer subtypes, because of deep neural net frame work and sparsity in SPACL model, it resolves the challenges faced by HDLSS data and can represents the nonlinear hierarchical representations of genes and pathways.

CHAPTER VIII

REFERENCES

[1] V. Pudata and S. V v, “A Short Note on Cancer,” *J. Carcinog. Mutagen.*, 2012.

[2] A. Fadaka, B. Ajiboye, O. Ojo, O. Adewale, I. Olayide, and R. Emuowhochere, “Biology of glucose metabolization in cancer cells,” *J. Oncol. Sci.*, 2017.

[3] W. Mair, “How normal cells can win the battle for survival against cancer cells,” *PLoS Biol.*, 2010.

[4] A. Sudhakar, “History of Cancer, Ancient and Modern Treatment Methods,” *J. Cancer Sci. Ther.*, 2009.

[5] P. Anand *et al.*, “Cancer is a preventable disease that requires major lifestyle changes,” *Pharmaceutical Research*. 2008.

[6] C. M. Dimitrakopoulos and N. Beerenwinkel, “Computational approaches for the identification of cancer genes and pathways,” *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. 2017.

[7] S. H. Hassanpour and M. Dehghani, “Review of cancer from perspective of molecular,” *J. Cancer Res. Pract.*, 2017.

[8] J. H. Folley, W. Borges, and T. Yamawaki, “Incidence of leukemia in survivors of the atomic bomb in Hiroshima and Nagasaki, Japan,” *Am. J. Med.*, 1952.

[9] D. Richardson *et al.*, “Ionizing radiation and leukemia mortality among Japanese Atomic Bomb Survivors, 1950-2000,” *Radiat Res*, 2009.

[10] J. A. Newton-Bishop *et al.*, “Relationship between sun exposure and melanoma risk for tumours in different body sites in a large case-control study in a temperate

climate," *Eur. J. Cancer*, 2011.

[11] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2017.," *CA. Cancer J. Clin.*, 2017.

[12] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA. Cancer J. Clin.*, 2018.

[13] A. Li, X. Yin, and Y. Pan, "Three-Dimensional Gene Map of Cancer Cell Types: Structural Entropy Minimisation Principle for Defining Tumour Subtypes," *Sci. Rep.*, 2016.

[14] T. Xu, T. D. Le, L. Liu, R. Wang, B. Sun, and J. Li, "Identifying cancer subtypes from miRNA-TFmRNA regulatory networks and expression data," *PLoS One*, 2016.

[15] J. Clark *et al.*, "Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays," *Oncogene*, 2003.

[16] I. P. Touw and S. J. Erkeland, "Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia," *Molecular Therapy*. 2007.

[17] L. S. Friedman *et al.*, "Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families," *Nat. Genet.*, 1994.

[18] J. N. Cancer Genome Atlas Research Network *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project.," *Nat. Genet.*, 2013.

[19] International Cancer Genome Consortium, "International network of cancer genome projects," *Nature*, 2010.

[20] R. Kumar, R. Srivastava, and S. Srivastava, "Detection and Classification of Cancer

from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features," *J. Med. Eng.*, 2015.

[21] J. B. Sørensen, F. R. Hirsch, A. Gazdar, and J. E. Olsen, "Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma," *Cancer*, 1993.

[22] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Comput. Methods Programs Biomed.*, 2012.

[23] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns.," *Proc. Natl. Acad. Sci. U. S. A.*, 1998.

[24] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science (80-.).*, 1999.

[25] K. Hanif, F., Muzaffar, S. M. & Perveen, K., Malhi, and S. U. Simjee, "Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment," *Asian Pacific J. Cancer Prev. J Cancer Prev*, 2017.

[26] E. Domingo-Musibay and E. Galanis, "What next for newly diagnosed glioblastoma?," *Futur. Oncol*, 2015.

[27] D. Hambardzumyan and G. Bergers, "Glioblastoma: Defining Tumor Niches," *Trends in Cancer*. 2015.

[28] K. Urbanska, J. Sokolowska, M. Szmidt, and P. Sysa, "Glioblastoma multiforme - An overview," *Wspolczesna Onkologia*. 2014.

[29] F. B. Furnari *et al.*, "Malignant astrocytic glioma: Genetics, biology, and paths to treatment," *Genes and Development*. 2007.

[30] D. Parry, D. Mahony, K. Wills, and E. Lees, "Cyclin D-CDK subunit arrangement

is dependent on the availability of competing INK4 and p21 class inhibitors.," *Mol. Cell. Biol.*, 1999.

[31] D. N. Louis *et al.*, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*. 2007.

[32] Q. T. Ostrom *et al.*, "CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010," *Neuro. Oncol.*, 2013.

[33] "Brain Tumor: Statistics | Cancer.Net." [Online]. Available: https://www.cancer.net/cancer-types/brain-tumor/statistics. [Accessed: 10-Jul-2018].

[34] J. Lu, M. C. Cowperthwaite, M. G. Burnett, and M. Shpak, "Molecular Predictors of Long-Term Survival in Glioblastoma Multiforme Patients," *PLoS One*, 2016.

[35] M. Hingorani, W. P. Colley, S. Dixit, and A. M. Beavis, "Hypofractionated radiotherapy for glioblastoma: Strategy for poor-risk patients or hope for the future?," *British Journal of Radiology*. 2012.

[36] "Glioblastoma Multiforme (GBM) | GBM Tumor Treatment." [Online]. Available: https://tocagen.com/patients/brain-cancer/glioblastoma-multiforme/. [Accessed: 10-Jul-2018].

[37] "Glioblastoma clinical trial shows combined therapy extends life for patients 65 and older - FirstWord Pharma." [Online]. Available: https://www.firstwordpharma.com/node/1457270. [Accessed: 10-Jul-2018].

[38] S. A. Grossman and J. F. Batara, "Current management of glioblastoma multiforme," *Semin. Oncol.*, 2004.

[39] R. Stupp *et al.*, "Radiotherapy plus Concomitant\nand Adjuvant Temozolomide for

Glioblastoma," *N. Engl. J. Med.*, 2005.

[40] C. Louis *et al.*, "The 2007 WHO Classification of Tumours of the Central Nervous System The 2007 WHO ClassiWcation of Tumours of the Central Nervous System," *Acta Neuropathol*, 2007.

[41] P. Kleihues and H. Ohgaki, "Primary and secondary glioblastomas: from concept to clinical diagnosis," *Neuro. Oncol.*, 1999.

[42] D. A. Reardon and P. Y. Wen, "Therapeutic advances in the treatment of glioblastoma: rationale and potential role of targeted agents.," *Oncologist*, 2006.

[43] R. G. W. Verhaak *et al.*, "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, 2010.

[44] H. Noushmehr *et al.*, "Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma," *Cancer Cell*, 2010.

[45] T. M. Kim, W. Huang, R. Park, P. J. Park, and M. D. Johnson, "A developmental taxonomy of glioblastoma defined and maintained by microRNAs," *Cancer Res.*, 2011.

[46] a a Alizadeh *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.," *Nature*, 2000.

[47] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, 2001.

[48] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, 2003.

[49] Z. Chang *et al.*, "eMBI: Boosting Gene Expression-based Clustering for Cancer

Subtypes," *Cancer Inform.,* 2014.

[50] J. A. Nepomuceno, A. Troncoso, and J. S. Aguilar-Ruiz, "Biclustering of gene expression data by correlation-based scatter search," *BioData Min.,* 2011.

[51] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: Coclustering genes and conditions," *Genome Research*. 2003.

[52] Y. Cheng and G. M. Church, "Biclustering of expression data.," *Proc. Int. Conf. Intell. Syst. Mol. Biol.,* 2000.

[53] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC Bioinformatics,* 2014.

[54] S. Huang, H. Wang, D. Li, Y. Yang, and T. Li, "Spectral co-clustering ensemble," *Knowledge-Based Syst.,* 2015.

[55] T. M. Kim, S. H. Yim, Y. B. Jeong, Y. C. Jung, and Y. J. Chung, "PathCluster: A framework for gene set-based hierarchical clustering," *Bioinformatics,* 2008.

[56] B. D. B. Lehmann *et al.*, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *J. Clin. Invest.,* 2011.

[57] M. W. I. Schmidt *et al.*, "Persistence of soil organic matter as an ecosystem property," *Nature*. 2011.

[58] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach," *IEEE/ACM Trans. Comput. Biol. Bioinforma.,* 2015.

[59] A. F. Syafiandini, I. Wasito, S. Yazid, A. Fitriawan, and M. Amien, "Cancer subtype

identification using deep learning approach," in *Proceeding - 2016 International Conference on Computer, Control, Informatics and its Applications: Recent Progress in Computer, Control, and Informatics for Data Science, IC3INA 2016*, 2017.

[60] R. Shen *et al.*, "Integrative subtype discovery in glioblastoma using iCluster," *PLoS One*, 2012.

[61] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, 1986.

[62] G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines," *Computer (Long. Beach. Calif).*, 2010.

[63] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, 2006.

[64] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "An interactive approach to multiobjective clustering of gene expression patterns," *IEEE Trans Biomed Eng*, 2013.

[65] T. Grotkjær, O. Winther, B. Regenberg, J. Nielsen, and L. K. Hansen, "Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm," *Bioinformatics*, 2006.

[66] L. Hopp, H. Löffler-Wirth, and H. Binder, "Epigenetic heterogeneity of B-cell lymphoma: DNA methylation, gene expression and chromatin states," *Genes (Basel).*, 2015.

[67] a Brazma and J. Vilo, “Gene expression data analysis.,” *Microbes Infect.*, 2001.

[68] S. A. Armstrong *et al.*, “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia,” *Nat. Genet.*, 2002.

[69] D. Glez-Peña, M. Reboiro-Jato, R. Domínguez, G. Gómez-López, D. G. Pisano, and F. Fdez-Riverola, “PathJam: a new service for integrating biological pathway information.,” *J. Integr. Bioinform.*, 2010.

[70] A. Alibés, A. Cañada, and R. Díaz-Uriarte, “PaLS: filtering common literature, biological terms and pathway information.,” *Nucleic Acids Res.*, 2008.

[71] J. Feher, “Quantitative Human Physiology,” *Quant. Hum. Physiol.*, 2012.

[72] H. J. Abuhusain *et al.*, “A metabolic shift favoring sphingosine 1-phosphate at the expense of ceramide controls glioblastoma angiogenesis,” *J. Biol. Chem.*, 2013.

[73] J. O’Campo *et al.*, “Aberrant p53, mdm2, and proliferation differ in glioblastomas from long-term compared with typical survivors,” *Clin. Cancer Res.*, 2002.

[74] I. Paul, S. Bhattacharya, A. Chatterjee, and M. K. Ghosh, “Current Understanding on EGFR and Wnt/β-Catenin Signaling in Glioma and Their Possible Crosstalk,” *Genes and Cancer*. 2013.

[75] S. Nobusawa, R. Stawski, Y. H. Kim, Y. Nakazato, and H. Ohgaki, “Amplification of the PDGFRA, KIT and KDR genes in glioblastoma: A population-based study,” *Neuropathology*, 2011.