A WEAKLY SUPERVISED CLUSTERING METHOD FOR CANCER SUBGROUP IDENTIFICATION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Duygu Özçelik

July 2016

A WEAKLY SUPERVISED CLUSTERING METHOD
FOR CANCER SUBGROUP IDENTIFICATION
By Duygu Özçelik
July 2016

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Öznur Taştan Okan (Advisor)

Tolga Can

Ercüment Çiçek

Approved for the Graduate School of Engineering and Science:

Levent Onural
Director of the Graduate School

ABSTRACT

A WEAKLY SUPERVISED CLUSTERING METHOD FOR CANCER SUBGROUP IDENTIFICATION

Duygu Özçelik
M.S. in Computer Engineering
Advisor: Öznur Taştan Okan
July 2016

Each cancer type is a heteregonous disease consisting of subtypes, which may be distinguished at the molecular, histopathological, and clinical level. Identifying the patient subtypes of a cancer type is critically important as the unique molecular characteristics of a particular patient subgroup reveal distinct disease states and opens up possibilities for targeted therapeutic regimens. Traditionally, unsupervised clustering techniques are applied on 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 lieu of this unsupervised framework, we propose a weakly supervised clustering framework, WS-RFClust, in which the clustering partitions are guided with the clinical outcome of interest. In WS-RFClust a random forest is trained to classify the patients based on a categorical clinical variable of interest. We use the partitions of patients on the tree ensemble to construct a patient similarity matrix, which is then used as input to a clustering algorithm. WS-RFClust inherently uses the nonlinear subspace of the original features that is learned in the classification step for clustering. In this study, we demonstrate the effectiveness of WS-RFClust on hand-written digit datasets, which captures salient structural similarities of digit pairs. Finally, we employ WS-RFClust to find breast cancer subtypes using mRNA,

protein and microRNA expressions as features. Our results on breast cancer subtype identification problem show that WS-RFClust could identify patients more effectively in comparison to the commonly used unsupervised clustering methods.

Keywords: Clustering, weakly supervised clustering, subspace clustering, cancer subtype identification, patient subgroup identification.

ÖZET

KANSER ALT GRUPLARININ KEŞFİ İÇİN ZAYIF GÖZETİMLİ BİR KÜMELEME METODU

Duygu Özçelik
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Öznur Taştan Okan
Temmuz 2016

Kanser heterojen bir hastalıktır; her bir kanser tipi moleküler, histopatolojik ve klinik olarak farklılıklar gösteren bir çok alt tipi barındırır. Bir kanser tipine ait alt gruplarının belirlenmesi, kişiye özel ve hedefe yönelik tedavi yöntemleri geliştirilebilmesini ve alt tiplerin moleküler karakteristiklerinin anlaşılmasıyla hastalığın mekanizmalarına dair bilgileri açığa çıkarabilmesini mümkün kıldığı için önemlidir. Geleneksel olarak kanser alt gruplarını keşfetmek için genomik veriler üzerinde gözetimsiz kümeleme teknikleri uygulanır ve bu yolla belirlenen gruplar, ancak hasta sağ kalımı gibi kritik bir parametre açısından ilişkililer ise anlamlı olarak değerlendirilirler. Biz bu gözetimsiz öğrenme çerçevesi yerine, WS-RFClust adını verdiğimiz, grupların ayrışmasına klinik parametrenin yön verdiği zayıf gözetimli bir kümeleme tekniği öneriyoruz. Bu yöntemde, rastgele orman sınıflandırıcısı kurulup, ormandaki ağaçların ara dallarında hastaların aynı gruplara düşüp düşmediği bilgisine dayalı olarak bir hasta benzerlik matrisi oluşturulmaktadır. Bu matris daha sonra bir kümeleme algoritmasına girdi olarak verilmekte ve hasta grupları bulunmaktadır. WS-RFClust, yapısı gereği sınıflandırma adımında oluşturulan, özniteliklerin doğrusal olmayan kombinasyonlarından oluşan öznitelik alt uzayını kullanmaktadır. WS-RFClust yöntemini el yazısı rakamlarında kullandığımızda, rakamların yapısal özelliklerini yakaladığını görmekteyiz. WS-RFClust'ın mRNA, protein ve

microRNA ifadeleme veri setlerini kullanarak meme kanserinin alt tiplerini bulmak için uyguladığımızda genel geçer kullanılan gözetimsiz kümeleme teknikleri ile oluşan kümelemelerden daha iyi çalıştığını göstermekteyiz.

Anahtar sözcükler: Öbekleme, zayıf gözetimli öbekleme, altuzay ile öbekleme, kanser alt tip keşfi, hasta alt gruplarının keşfi.

Acknowledgement

Foremost, I would like to thank my thesis supervisor Asst. Prof. Dr. Öznur Taşţan for her contionus support, motivation and immense knowledge. She is a very kind person and a highly qualified, hardworking, and brillant scientist. Her guidance helped me to conduct this research work.

I thank to my beautiful family, my mom Birgül and my father Dilaver, for their endless care and support.

Lastly, I would like to thank my boyfriend, Uğur for his tranquil and positive personality.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is the name of a group of related diseases characterized by uncontrolled growth of the cells [1]. All body cells in a healthy human follow a regular path; they divide, proliferate and programmatically die. Cancer cells, on the other hand, abnormally grow and divide. This uncontrolled cellular growth eventually leads to a transformation of normal cells into tumor cells that may invade the normal tissues and organs. Despite intensive efforts, cancer remains to be among the leading causes of death across the world [2].

A major hurdle in devising more effective cancer therapies is the accurate stratification of patients. For instance it is critical to identify at the time of diagnosis which patients harbor aggressive tumors and which of them will progress slowly. Aggressive treatment strategies exercised on the latter impairs

the quality of the patient life with no additional benefit [3, 4]. Therefore, patient stratification is a critical first step in developing personalized treatments.

Cancer is heterogeneous at the molecular level; seemingly similar tumors that are classified into the same histopathological subtype may have distinct genotypes resulting in distinct phenotypes [5]. Identifying patient subgroups of patients with similar genotype and phenotype may reveal the unique molecular characteristics of this group that shape different cancer states and opens up possibilities for targeted therapeutic regimens. Cancer is a disease of the genome. The cancer cell acquires several somatic aberrations during its lifespan. Recent developments in genomic sequencing technologies enabled the characterization of somatic alterations in the cancer genomes [6]. Using this rich source of genomic cancer data, this thesis focuses on developing a machine learning approach that allows the stratification of patients with similar molecular profiles and similar clinical outcomes.

Traditionally, unsupervised clustering analysis is applied on 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 rate of patients [7, 8, 9]. In lieu of this unsupervised framework, in this thesis, we propose a weakly supervised clustering framework (WS-RFClust). In this approach, the clustering partitions are weakly guided with the clinical outcome of interest. We achieve this by using similarity of patients under subsets of features created in a random forest ensemble that is trained with the label of interest.

In this study we have limited our analysis to breast cancer but the approach

presented herein can be applied to any cancer type and any clinical variable of interest. Breast cancer is the most commonly diagnosed malignancy and the second leading cause of cancer-related deaths among females [10]. Early diagnosis underlies every therapeutic strategy against breast cancer by improving the survival rate. Therefore, the clinical variable of interest we focus in this study is the survival rate of patients. We show that our approach lead to clusters of interest.

The thesis is organized as follows:

- Chapter 1 introduces the thesis and states the problem definition.
- Chapter 2 provides a brief summary of related biological concepts and description of the data used in our experiments.
- Chapter 3 describes related work in relation to our contributions.
- Chapter 4 provides the weakly supervised approach (WS-RFClust) we take.
- Chapter 5 elaborates on empirical results on digit dataset and the breast cancer dataset.
- Chapter 6 states conclusions and recapitulates the key findings in this study.

Chapter 2

Background

This section contains a high-level description of the key biological terminology; datasets used and relevant preliminary information on breast cancer.

2.1 Biological Background

Genetic information of humans are stored on DNA (deoxyribonucleic acid), a macromolecule made up of building blocks called nucleotides. The genetic information on a DNA is expressed through a process called transcription whereby a portion of the DNA sequence is copied to a RNA (ribonucleic acid) molecule. A major type of RNA, the messenger RNA (mRNA), encodes the amino acid sequence of a target protein and carries this information to the

ribosome where protein synthesis will take place. During protein synthesis, mRNA sequence is translated into a sequence of amino acids which will fold into a specific three dimensional structure. Since proteins in the cells are polymerized from the mRNA transcripts; the mRNA expression levels provide a good approximation of the abundance of proteins [11].

Another form of RNAs is microRNA. microRNAs are small non-coding RNAs of 21-25 nucleotides that regulate gene expression posttranscriptionally. miRNAs exert their regulatory role by changing the mRNA levels through degradation whereby mRNAs are completely silenced or partially inhibited through translational repression [12]. MicroRNAs have been implicated in almost all cellular processes and some microRNAs are also reported to act as oncogenes and tumor-suppressor genes [13].

2.2 Molecular Data Used

In this work datasets pertaining to mRNA, microRNA and protein expression levels on solid primary tumors are used. This study relies on breast cancer data made publicly available by the UCSC (University of California, Santa Cruz) Cancer Genomics Browser [14] retrieved from TGCA project [15]. In the following subsections, we provide details on each data type:

2.2.1 mRNA Expression

RNA-sequencing is a next-gen sequencing technology (NGS) that quantifies the level of RNA transcripts in a sample. Since measuring mRNA expression level provides information about the activity of the respective gene, it enables analyzing gene activities of a cell in different conditions [16]. The RNAseq expression data used in this thesis were obtained from UCSC Cancer Genomics Browser June 2016 data archive. We used RNA expression file for per patient and downloaded data from 'Cancer' menu clicking the 'Add Datasets' button. Then, we selected TCGA breast invasive carcinoma from list and downloaded compressed file named '`TCGA_BRCA_exp_HiSeqV2-2015-02-24`'. mRNA data contained 1196 patients and 20531 features.

TCGA data included two differently processed RNAseq data: RNASeq and RNASeqV2. RNASeq data reports RPKM (Reads Per Kilobase of exon model per Million mapped reads). RPKM uses number of sequence reads of an mRNA and normalizes the number of reads by dividing it to the total length of the transcript. RNASeqV2 is based on RSEM normalization technique. mRNA of a gene can have different isoforms. Isoform is a splice variant of an mRNA such that the transcription start site of an isoform is different from its base mRNA [17]. Alteration in a transcription start site affects gene expression behavior. RSEM technique accounts for mRNA splice and its isoforms. RSEM uses the measure of estimated fraction of transcripts comprising a given isoform or a gene [18]. We used the RSEM normalized RNAseqV2 data files.

2.2.2 miRNA Expression

microRNA expression data were downloaded from 'Cancer' menu clicking the 'Add Datasets' button. Then, we selected TCGA breast invasive carcinoma from list and downloaded compressed file named '`TCGA_BRCA_miRNA-2015-02-24`'. The miRNA data was taken from the June 2016 data archive. miRNA data comprised information from 1194 patients and 1046 features. We use Level 3 normalized data, where the normalization is conducted by calculating expression for all reads aligning to a particular miRNA [19].

2.2.3 Protein Expression

Reverse Phase Protein Array(RPPA) is a high-throughput method to obtain protein expression levels. We downloaded data from 'Cancer' menu clicking the 'Add Datasets' button. Then, we selected TCGA breast invasive carcinoma from list and downloaded compressed file named '`TCGA_BRCA_RPPA_RBN-2015-02-24`'. Protein expression data hold information on 747 patients and 131 proteins. The data were taken from the June 2016 data archive. Level 3 normalized data are used. Level 3 normalization is carried out by calculating the median absolute deviation within a protein sequence for each sample, then calculating median absolute deviation across samples for each protein type [20].

2.2.4 Clinical Data

The clinical data were taken from the June 2016 data archive. Clinical information is available in compressed files of mRNA, miRNA and protein expression datasets. The file named `"clinical_data"` bear the fields of clinical traits. Clinical traits are days to last follow up, days to death, tumor stage and age at inital diagnosis. The description of these fields is as follows:

1. **Days to last followup:** Numeric value in days that keeps last contact day of a patient.
2. **Days to death:** Number of days between initial diagnosis date of a patient and death day of the patient.
3. **Age at initial pathologic diagnosis:** Age of a patient at the time cancer is firstly diagnosed.
4. **Ajcc pathologic tumor stage:** AJCC staging criteria defines the extension of a cancer and how far from its originated tissue [21]. Tumor stages are defined according to originated tissue of tumor, its expansion area, size and whether they spread to neighboring lymph nodes or not. Stages are named according to TNM system. T represents size or extent of primary tumor, N is vicinity to lymph nodes, M is the flag of metastasis [22].

Days to last follow-up and days to death are used to extract overall survival time. If the vital status of a patient is alive, the last follow up date is used for

the survival time. On the other hand, if a patient is deceased then the days to death field is used. We predicate survival analysis on survival time in months, which we calculate by dividing the overall survival time in days to 30.

2.3 Breast Cancer

Breast cancer is the leading cause of cancer death in women [23]. Every year 1.3 million new incidences arise and 450,000 deaths worldwide are due to breast cancer [24]. Breast cancer is a group of heterogeneous diseases that in their morphology, molecular profile and responsiveness to therapy. Accurate grouping of breast cancer and understanding the underlying biology behind these subtypes is of particular importance for diagnosing patients and making therapeutic decisions.

There are numerous ways for classifying breast cancer based on different principles. Breast cancer are classified into stages based on the size of the tumor, the spread of the tumor to the nearby lymph nodes and whether it metastasize to other tissues or not [25]. Grade is another metric used for tumor classification and is based on the differentiation of cancer cells. The normal breast cells are well differentiated to conduct their specialized function; cancer cells lack this specialization. By comparing the cancer tissue with the normal tissue, the grade of the cancer is determined: grade 1 cancer cells have small difference to the normal cells, grade 2 cancer cells are moderately differentiated, while grade 3 cancer cells are completely lose their differentiation compared to normal cells. Grade 3 cancers tend to grow and spread more quickly [26].

Pathology-driven classification does not always provide sufficient information to evaluate the biological characteristics of individual tumors and it is not useful for guiding the treatment selection [27]. The status of three molecular markers has served as the basis for breast cancer classification. Receptors are membrane proteins that receive signals from outside of the cell by binding to signaling molecules such as hormones [28]. There are three major receptors that are used in classification of breast cancer: estrogen receptor (ER) [29], progesterone receptor (PR) [30] and human epidermal growth factor receptor 2 (HER2) [31]. Breast cancer cells, which have ER require estrogen for their growth and are denoted as ER+, while that have PR is denoted by PR+. Breast cancer cells that overexpress HER2 or have HER2 amplified are referred to as HER2+. Based on the status of these molecular markers breast cancer is divided into four molecular subtypes:

- **Luminal A:** This type of tumors tend to be ER+ and/or PR+ and often does not show over expressed HER2 protein. Of the four subtypes, luminal A tumors tend to have the most favorable prognosis, with fairly high survival rates and fairly low recurrence rate [32, 33].

- **Luminal B:** These tumors tend to be ER+ and/or PR+. types (ER+, PR+); unlike luminal A tumors HER2 is overexpressed in these tumors. Ki67 protein is also highly expressed in Luminal B subtype, while it is lowly expressed in Luminal A [34]. Luminal B subtype has the highest percentage of lymph node involvement. [35]. Luminal B tumors grow more quickly compared to Luminal A tumors and often lead to poorer prognosis [36].

- **HER2+:** This subtype of breast cancer is characterized by the absence of both ER and PR receptors and over-expression of HER2.

- **Triple-negative/basal-like:** This type of tumors lack expression of HER2 or amplification, they are also PR-negative and ER-negative. These tumors are aggressive, more likely to metastases, and tumors are often associated with poor prognosis and survival rates. This type of cancer bears similarities to basal-like tumors, but also represents a distinct subtype with heterogeneous properties.

Immunochemistry (IHC) markers for ER, PR and HER2, together with tumor size, grade and nodal involvement are used for patient prognosis and management. These classification approaches have been successful in reducing the breast cancer mortality during the past three decades; however, are not sufficient to derive individualized therapy. Thus, breast cancer has been extensively studied at the genomic and transcriptomic levels with the rationale that underlying gene expression patterns reflect the tumor characterictics at the molecular level [37]. With the development of microarrays gene expression analysis subtypes of patients are identified using gene expression profiling. Through unsupervised clustering analysis of gene expression Sørlie et al. (2001) reported five intrinsic subtypes with distinct clinical outcomes, i.e., luminal A, luminal B, HER2 over-expression, basal and normal-like tumors [38, 39]. These largely coincided with the IHC-defined subtypes. These five intrinsic subtypes have been validated by several other studies with different gene signatures. Parker et al. (2009) reported a set of 50 genes (referred to as PAM50) with good prognostic performance [40]. Applying unsupervised clustering on copy

number variation and gene expression data, Curtis et al. (2012) recently suggested there are 10 subtypes of breast cancer [41]; however, these subtypes are not yet been clinically accepted.

Chapter 3

Related Literature

Clustering analysis refers to a broad set of techniques that seeks to find subgroups or clusters in the data. The goal is to partition the observations such that observations that are assigned to the same group are similar while those in different groups are dissimilar. Clustering algorithms need a definition of what it means for two or more observations to be similar or different. Clustering approaches can be broadly categorized into two as unsupervised and semi-supervised methods. In this chapter we will not attempt to discuss all clustering algorithms but instead focus only on the clustering algorithms that are commonly used for finding subgroup of patients based on their molecular profiles and discuss the related work.

3.1 Unsupervised Clustering

Most of the traditional clustering methods are unsupervised. In the unsupervised learning setting, the learner is given only the unlabeled examples and aims to discover the underlying structure and categories in the input space, $\mathcal{X}$ [42]. Since clustering analysis is useful in a diverse set of applications a variety of clustering techniques have been developed. We will not attempt to review them all but instead focus on those that are commonly used in analyzing molecular expression profiles for the task of finding subgroups of patients. There are three methods that are widely adopted for this purpose. These include hierarchical clustering, non-negative matrix factorization and consensus-clustering. Below we will give a brief description of these methods.

3.1.1 Hierarchical Clustering

Instead of producing a single partitioning of the input items, hierarchical clustering produces a hierarchy of nested clusterings [43]. The resulting family of clusterings can be graphically represented in a tree-based representation, called a dendrogram. There are two main approaches to hierarchical clustering algorithms: divisive and agglomerative (bottom up). The most commonly adapted method in patient subgroup analysis is the agglomerative approach. In this approach, one assumes that there are n samples. In the first step, the clustering process the algorithm seeks sample pairs that are the most similar. Let the cluster that contains samples with lowest dissimilarity be C_1. Now,

there are $n-1$ clusters. Next, it finds the sample i which is closest to C_1 and puts them in together in new cluster C_2. This procedure is repeated. It repeats the same process until there is one cluster left.

Algorithm 1 Hierarchical Clustering [44]

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as its own cluster.
For $i = n, n-1, ...2$:
   (a) Examine all pairwise inter-cluster dissimilarities among i clusters and identify the pair of clusters that are the least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

In addition to a distance measure to evaluate the similarity of examples, hierarchical clustering algorithms need to define how to measure the distance between clusters; these are referred to as linkage methods. The commonly used linkage methods are:

Single linkage: The distance between two clusters is identified as the shortest distance between a point from cluster 1 and a point from cluster 2. These two points are the closest points of two clusters.

Complete linkage: Distance between two clusters is identified as the longest distance between a point from cluster 1 and a point from cluster 2. These two points are the farthest points of two clusters.

Average linkage: Mean distance between two clusters is computed by averaging all pairwise distances of points between clusters.

Hierarchical clustering is one of the most commonly applied algorithms to group samples [45, 46]. Although run as an automated tool, it is sensitive to the distance metric used and typically requires a subjective evaluation to choose the final clustering.

3.1.2 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is commonly used for clustering analysis with high-dimensional data. Given an $n \times m$ matrix V where each element $V_{i,j} \geq 0$ and a desired rank $k < min\{m, n\}$, NMF decomposes into two non-negative matrices W $(n \times k)$ and H $(k \times m)$ such that

$$V \approx WH \tag{3.1}$$

This factorization has a natural interpretation and an inherent clustering property. Each data vector v is approximated by a linear combination of the columns of W, weighted by the components of h. In other words, each data item can be explained by an additive linear combination of few basis components. Since fewer basis vectors are used to represent all data vectors, a good approximation can only be achieved if the basis vectors capture structure in the data. If this is satisfied, NMF automatically clusters the columns of input data.

The goodness of the approximation in Equation 3.2 can be measured by Frobenius norm and the W and H can be found by solving the following optimization problem:

$$\min_{W \geq 0, H \geq 0} f(W, H) = \frac{1}{2} \|A - WH\|_F^2 \tag{3.2}$$

This problem is not solved analytically in general. There are different algorithms suggest for to solve NMF [47, 48].

NMF is applied to find subtypes of cancer by several studies. Zhang et al. find uncovered pathways, clinically relevant subtypes and relation between different cellular activities in multi-dimensional omics data [49]. TCGA Network group focuses high-grade serous ovarian cancer; they process DNA copy number, microRNA and mRNA expression, promoter methylation and exons from coding genes. They revealed four transitional subtypes related with survival rate, role of BRCA1 and BRCA2 genes and NOTCH and FOXM1 signalling in ovarian cancer, by applying consensus clustering [50].They map different types of genomic data into same measurement system by using joint matrix factorization. There are also other studies that use NMF with network data. Hofree and Ideker (2013) [51] in their network-based stratification (NBS) technique integrate somatic tumor genomes with gene network. The mutations are propagated on the network and the network-smoothed patient profiles are clustered into a predefined number of subtypes via NMF.

3.1.3 Consensus Clustering

In finding the subgroups of patients with expression profiles, relatively small sample sizes and the high dimensionality of features render clustering methods sensitive to noise. This might lead to instabilities that would result in different clustering assignments if the input is slightly perturbed or different parameters are used. Bhattacharjee et al. (2001) [52] used a bootstrapping approach to validate the resulting clusters in analyzing human lung carcinoma gene expression dataset. They input the bootstrapped samples into a hierarchical clustering algorithm and assessed the stability of the cluster assignments. Monti et al. (2003) [53] generalized this approach into a method named as *consensus clustering*.

In consensus clustering, the perturbed versions of the original data are generated. Resampling techniques can be used for this purpose. In the ensuing step the clustering algorithm is run with each of the perturbed data input. The cluster assignments obtained at each run are aggregated in a $N \times N$ consensus matrix. The consensus matrix stores numerical entries that correspond to the proportion of cluster runs in which the two items are clustered as pairs to the total number of cluster runs. In this way the cluster assignments that are robust to sampling variability are identified [54].

Consensus clustering is a meta-method that can be wrapped around different clustering algorithms and used in variety of recent papers in cancer subgroup identification. Hayes and Verhaak (2010) [55] applied consensus clustering with hierarchical clustering on Glioblastoma Muliformae (GBM)

and identified four new subtypes of GBM: Classical, Proneural, Neural and Mesenchymal. TCGA Network group analyses primary breast tumors with consensus clustering, and they identified four intrinsic subtypes by processing DNA methylation, DNA copy number, microRNA and mRNA expression and RPPA data [56]. Another study is conducted on high grade endimetriod and clear cell ovarian cancer. Winterhoff et al. found correlation between transcriptional subtypes of high grade serous ovarian cancer and high grade endimetriod in advanced stage and they apply consensus NMF [57].

3.2 Semi-Supervised Clustering

Unlike unsupervised clustering, the repertoire of methods for semi-supervised clustering methods is fairly limited. Semi-supervised learning uses additional knowledge together with the feature information. This additional knowledge can be encoded in different forms. There could be a small set of labeled examples available. Alternatively it could be stated as constraints on such as the following; "two must be (must-link) or cannot (cannot-link) be in the same cluster". Or additional information about the properties that the instances of a cluster have to hold could be available.

There exist methods for situations where the class labels are known for a subset of the observations. For instance, Basu et al. (2008) [58] proposed constrained k-means. In this method the k-means initial clusters are initiated with the labeled examples and they are always kept in their initial clusters even if they are closer to another cluster centroid. Alternative to this method,

they have also suggested seeded k-means, which is identical to the former with the exception that the seeded k-means always assigns examples to the nearest cluster. Other clustering methods, also developed for the specific case of using small set of labeled examples, exist. However, those methods are not typically applied to the cancer subgroup identification problem because the class labels are not generally available. Similarly, numerous semi-supervised methods, which operate with known constraints, were also proposed [59], but they have not been used for patient subgroup identification.

The third class of semi-supervised learning approaches seek to find clusters by exploiting some additional information about the cluster properties. However, it may not be possible to identify clusters solely by using this additional information. This variable acts as noisy surrogate for the clusters [60] and this is the setting which we focus on in this thesis. Substantial amount of additional information is readily available on cancer patients and this work identifies clusters of potential interest if they differ in terms of clinical outcomes such as the absence/presence of metastasis or the grouping into high survivor or low survivor categories. In conventional unsupervised clustering, this information is used for validating the unsupervised clustering approaches. For example, once the clusters are found, the divergence in the survival distribution of the clusters are checked and the clusters are found interesting if it are different in terms of survival. This is the common approach that is used in all recent cancer subgroup identification work [61, 62, 63, 64].

Very few methods have been suggested for identifying clusters associated with an outcome variable. Bair and Tibshirani (2004) [65] was the first to address this problem. They referred to their method as supervised clustering.

For each feature in the data set, it tests the null hypothesis of no association between the feature and the outcome variable and uses a test-statistic. The algorithm proceeds as follows: let m be a feature in a dataset. For each feature in the dataset, the algorithm calculates a test statistic T_m between the feature m and the outcome variable. Then, it chooses a threshold M and filter features with $|T_m| \leq M$. With the remaining features, it performs clustering using a conventional clustering algorithm such as k-means or hierarchical clustering. Since clustering is performed using only a subset of the features, this method reduces the high-dimensional data sets into lower dimensions and perform clustering with a reduced feature set. Bair and Tibshirani [66] show that this relatively simple method can identify biologically relevant clusters in several data sets. Later, Bullinger et al. (2004) [67] applied this method to discover acute myeloid leukemia subtypes associated with patient survival.

Koestler et al. (2010) [68] proposed a method called semi-supervised recursively partitioned mixture models with the same rationale. It also calculates a score for each feature and measures the association between that feature and the outcome variable of interest in the first step. Next, clustering is carried out using only the features with the largest scores. The difference between the Bair et al. and Koestler et al. is that the semi-supervised method applies the recursively partitioned mixture models algorithm of Houseman et al. (2008) [69] instead of a standard clustering algorithm.

Gaynor and Bair (2013) also proposed an alternative method called supervised sparse clustering. This method adapts the sparse clustering method introduced in [70] which is motivated from the observation that although clusters might not be visible under all features, unsupervised clustering method

can be possible only under a subset of features. The sparse clustering algorithm of Witten and Tibshirani (2010) [71] maximizes a k-means objective function where the features are weighted. In this method, the feature weights are uniformly initialized. At each iteration of k-means these weights are updated. In Gaynor and Bair's work the tested null hypothesis is that the mean value of the feature j does not vary across the clusters. For features that fail the test, the weights are assigned to zero and whereas the remaining feature weights are assigned to non-zero values.

Semi-supervised methods are not used in any of the recent cancer subgroup identification work.

Chapter 4

WS-RFClust: Weak Supervised Random Forest Clustering

As discussed in chapter 3, unsupervised approaches are widely used for the task of patient stratification. However, they do not make use the critical variable of interest in the cluster partitioning step. In this study, we propose a semi-supervised approach in which the clusters are guided with a surrogate variable that accounts for the survival of the patients. We call this approach Weakly Supervised Random Forest Clustering (WS-RFClust). This chapter will introduce WS-RFClust.

4.1 WS-RFClust

WS-RFClust operates with that basic principle that clusterings, which show agreement with the variable of interest, are favored over the rest. Let $\mathcal{D} = \{\mathrm{x}^{(1)}, \mathrm{x}^{(2)}, \ldots, \mathrm{x}^{(N)}\}$ be the set of patients, where each patient feature vector $\mathrm{x}^{(i)} \in \mathbb{R}^d$ and is derived from cancer genomic profiles. We are also given $\mathrm{y} = y^{(1)}, y^{(2)}, \ldots, y^{(N)}$, where each $y^{(i)} \in \{1, 2, \ldots k\}$ and k is the number of classes.

We would like to find a partitioning C such that:

- $\mathcal{D}$ is grouped into a number of disjoint subsets C_j's where, $\mathcal{D} = \cup_{j=1}^{k} C_j$ and where $C_i \cap C_j = \emptyset$
- the C is guided with y.

The main steps of WS-RFClust are as follows:

- **Step 1:** Using $\mathcal{D}$ learn a random forest classifier that can predict target variable y. Call the forest of trees, $\mathcal{RF}$.
- **Step 2:** To calculate the similarity of i and j, sort down i and j in the forest trees for which both examples are in the bag, and check whether i and j fall onto the same internal node at a randomly drawn depth. Based on the fraction of occasions when they share the same internal node, calculate patient similarity.

- **Step 3:** Input this similarity matrix to a clustering algorithm to arrive at a clustering.

Here we assume that the target variable y is discrete; however, the approach can easily be extended to the cases where y is a continuous variable. Alternatively, the target variable can be cast as a survival variable by replacing the random forest classifier with a random forest regressor or a random survival forest.

The algorithmic details are provided in Algorithm 2.

Algorithm 2 WS-RFClust: Weakly Supervised Random Forest Clustering

Input: $\boldsymbol{D}$ data matrix with n observations, $\boldsymbol{X}$, $n \times p$ feature matrix, $\boldsymbol{y}$ associated class labels, $\mathcal{R}$, random forest classifier, d_l fraction determine the lower bound of the range where the depth will be sampled, d_u : will be used to determine the upper bound of the depth range, B number of trees in random forest, r random forest parameters, c clustering parameters.
Output:
$\mathcal{D}$ is grouped into a number of disjoint subsets C_j's where, $\mathcal{D} = \cup_{j=1}^{k} C_j$ and where $C_i \cap C_j = \emptyset$
1. $\mathcal{F} \leftarrow$ RFClassifier($\boldsymbol{X}$, y, B, r)
2. $S \leftarrow$ Calc-RFrds(F, d_l, d_u)
3. $C \leftarrow$ Cluster(S, c) // input other parameters for the clustering algorithm
4. return C

4.1.1 Step 1: Random Forest Classification

Random forest is an ensemble method that learns many decision trees and aggregates their results [72]. Each decision tree is independently trained using

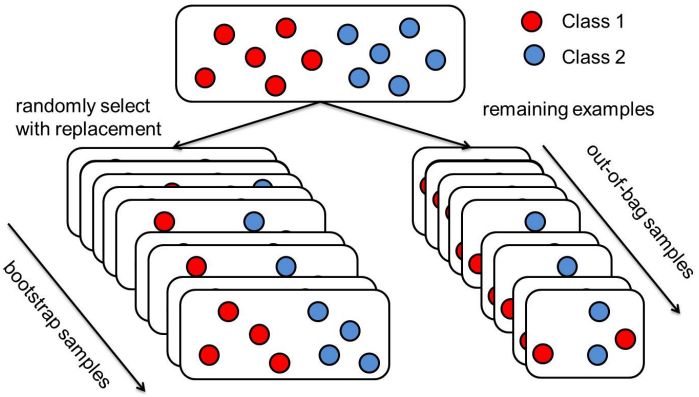


Figure 4.1: Bootstrapping samples in bag.

a bootstrap sample of the training examples. In the prediction step the sample is put down for class label prediction by each tree and the predicted labels from all trees are collected. The final class label is decided based on the majority vote of the trees.

In addition to constructing each tree using a different bootstrap sample of the data, random forest adds another layer of randomness in the tree construction step. In standard decision tree learning process, each node is split using the best split among all variables. In a random forest, each node is split using a subset of features randomly chosen at that node. If p is the total number of features and m is subset of features, then $m \ll p$ or $m \cong \sqrt{p}$. In this way the decision trees that are learned are decorrelated from each other [73] and the variance of the model is reduced. In a random forest model the impurity measure which is generally used as the split criterion is the gini index. An outline of the random forest algorithm is provided in Algorithm 3 and illustrated in figure 4.2.

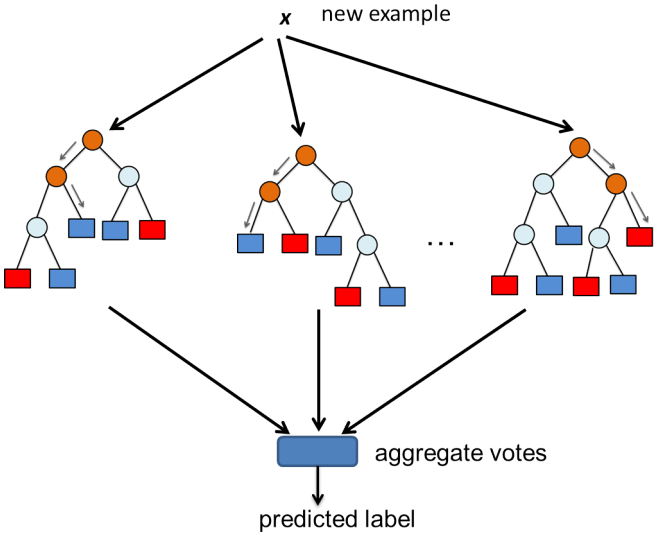


Figure 4.2: A schematic illustrating random forest classifier.

4.1.2 Step 2: Calculating Random Forest Random Depth Patient Similarity

This step is the critical step of our proposed WS-RFClust algorithm. Using the random forest ensemble of trees, we calculate a similarity metric, which we call random forest random depth similarity. Consider T_b, the b-th tree in the ensemble. T_b is trained with the bootstrap sample Z_b. For all pairs that are both in Z_b, we check whether they fall on the same internal node at a random depth. Given an interval range, we draw a depth, d_b uniformly at random. This depth is typically chosen from the mid level of the tree (in section 5.1.2 we discuss the effect of sampling from different depths). For a particular pair, we run down the examples on the tree and check if the examples land on the

Algorithm 3 Random Forest Algorithm [74]

1. For $b = 1$ to B
   (a) Draw a bootstrap sample Z_b of size N from the original training data D.
   (b) Grow a decision tree T_b with the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_min is reached.
   i. Select m features at random from the p variables.
   ii. Pick the best variable/split-point among the m features.
   iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\mathcal{R} = \{T_b\}_1^B$

To make a prediction at for a new example x:
Let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree. Then $\hat{C}_{rf}^B(x) = majority\ vote\{\hat{C}_b(x)\}_1^B$

same internal nodes at this given depth. For the pairs that end up at the same node, the similarity between them is incremented by 1. This is repeated for all trees and the similarities are finally normalized with the number of bootstrap samples where both examples were in the bag. We call this similarity metric random forest random depth similarity (RFrds). The steps of this calculation are of RFrds Algorithm 4.

RFrds similarity metric is similar to random forest proximity [75] suggested earlier. Random forest proximity is calculated based on how often the example pairs fall onto the same leaf node where as RFrds calculates the similarity based on internal nodes. The resulting proximities are different in that sense. By looking at the higher level internal nodes, we aim at finding the latent structure of the data based on features that are selected to have good prediction accuracy of y. The different depths provide different views of the samples. Checking

Algorithm 4 Calc-RFrds: Calculation of random forest random depth similarity.

Input: $\boldsymbol{N}$ size of observations, $\boldsymbol{D}$ set of $\boldsymbol{N}$ examples, $\boldsymbol{B}$ number of trees in the random forest, $\{T_b\}_1^B$ trees in the random forest, $c_{i,j}$ number of boostrap samples where i and j are in

Output: $\boldsymbol{S}$: $m \times m$ similarity matrix

1. For each i, j pair in D

    i. For all bootstrap samples b where i, j are both in the T_b

        (a) Get tree T_b of B

        (b) Get height h_b of T_b

        (c) Sample d from $[h_b \times d_s, h_b \times d_e]$ uniformly at random

        (d) Traverse i on T_b until depth d is reached and find the internal node p_i on which i falls

        (e) Traverse j on T_b until depth d is reached and find the internal node p_j on which j falls

        (f) **if** $p_i == p_j$ **then**

            $S(i,j) \leftarrow S(i,j) + 1$

    ii. $S(i,j) \leftarrow \frac{S(i,j)}{c_{i,j}}$

2. return S

whether a pair of observation would fall on the same internal node translates into a special partitioning of the data based on a nonlinear combination of features (nodes up to that point) and checking if they are in the same cluster based on that subspace of features. In that sense this calculation of RFrds is related to subspace and multi-view clustering methods [76]. On the other hand, since we are checking whether examples are similar or not on all the trees in the ensemble it is a consensus clustering approach wherein the trees are constructed by the bootstrap samples of the data.

4.1.3 Clustering

The last step of the algorithm uses the similarity matrix generated in the previous step to produce the desired clusterings. We convert the similarity matrix to a distance matrix by subtracting the values from 1 and we input it to the clustering algorithm. In this work we use the hierarchical clustering algorithm (explained in 3.1.1) with average linkage. We experiment with different k values.

4.2 Other Methods Employed

4.2.1 Feature Selection

In our experiments we used different feature selection methods to reduce the number of features employed in the random forest classifier.

4.2.1.1 Students t-test

The two-sample t-test checks if two population means are equal [77]. In the context of feature selection it is used to determine if the feature value distribution means for different classes differ. Let A and B are two feature value distributions to compare, n_A is number of samples in group A and n_B is number of samples in group B. μ_A is mean and σ_A^2 is variance of group A, μ_B is mean and σ_B^2 is variance of group B. T value is calculated as

$$t = \frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \tag{4.1}$$

4.2.1.2 ROC

ROC curves are typically used to compare performances of different classification models and displays the relation between true positive and true negative rates. In the case of feature selection, ROC curves are used as follows: consider

classifying examples based on a single feature, if the feature value is above a certain threshold it is classified in class 1 and if it is in the second class it is classified as class 2. By moving the threshold over all possible threshold values, one can obtain a ROC curve. The area between the ROC curve and the random line - in the case of binary classifier gives an assessment of how valuable this feature is by itself in predicting the class labels.

4.2.1.3 Relative entropy

Relative entropy as implemented in Matlabs `rankfeatures` method and is defined in Theodoridis et al. [78] is employed. The relative entropy can be used to measure purity of class labels.

4.2.1.4 Bhattacharyya distance

Bhattacharyya distance is used as a class separability measure. Assuming the feature values follow Gaussian distributions and class priors are equal, the Bhattacharyya Distance is calculated as follows:

$$P_e = \int_{R_2} \mathrm{p(x|c_1)}\,\mathrm{d}x + \int_{R_1} \mathrm{p(x|c_2)}\,\mathrm{d}x \tag{4.2}$$

4.2.1.5 Wilcoxon Signed Rank-Sum Test

Wilcoxon signed rank sum test hypothesis that two samples come from the same population against an alternative hypothesis. The test assumes it does not require the assumption of normal distributions.

4.2.2 Out-of-Bag-Error

In random forest classifier, an estimate of the error rate can be obtained using the out-of-bag samples. While constructing each tree, bootstrap samples are selected with replacement. In this process, approximately two-third of data is chosen as training set and the samples that are left out of the sample are referred to as out-of-bag (OOB) samples. For each of the example that is not in the bootstrap sample, the classifier makes a prediction with the tree grown on this sample. OOB error is obtained by averaging errors over the trees and the examples. Generally the OOB estimates are quite accurate given that enough number of trees are grown [73].

4.3 Validating Clusters

4.3.1 Kaplan-Meier Estimator

Kaplan-Meier method estimates survival function from survival time data of individuals [79]. Survival rate is at time t, proportion of patients survived from beginning of follow-up time. Probability of an event happening in short time interval is calculated by multiplying length of time and hazard rate of overall survival. Hazard rate is event rate at time t conditional on survival until time t or later. In order to interpret differences between survival groups statistically, we calculate hazard ratio, which is proportion of hazard in one group to hazard of another group. It uses log-rank test that is a statistical method to compare survival distribution of two cohorts. Null hypothesis is there are no difference between cohorts in terms of probability of an event occurring at any time point. Log rank method tests whether these populations are significantly different or not [80].

4.3.2 Silhouette Width

In order to validate clusters we use different metrics. One of them is the silhouette width. Silhouette width is a measure which shows relative quality of clusters [81]. It compares distance within a cluster with distance between clusters. Let $a(i)$ is the average dissimilarity within a cluster, $b(i)$ is average similarity between clusters. i is number of the cluster and $s(i)$ is average

silhouette with of all clusters. Then, silhouette width is:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{4.3}$$

Ideal case occurs when $a(i) = 0$ and $b(i) = 1$. In this condition, average silhouette with becomes 1. In worst case, $b(i) = 0$ and $a(i) = 1$, silhouette width becomes 0. These conditions show that $0 < s(i) < 1$. If average silhouette width is closer to 1, this means we obtained a good clustering.

Chapter 5

Results

5.1 Results on MNIST Digit Dataset

5.1.1 WS-RFClust Clusters

We applied WS-RFClust on MNIST handwritten digit dataset [82]. The dataset contains 60,000 images of 28×28 pixel handwritten digits. We used 5000 training digit samples, 500 from each class. The random forest classifier is generated with 200 trees, and trained with digit labels as class labels. In constructing the similarity matrix, we sample from $[(h * 1/3) - (h * 2/3)]$ interval depth. The digits are then clustered by inputting the corresponding similarity matrix to the hierarchical clustering algorithm.

Figure 5.1 shows a heatmap of clusters after applying hierarchical clustering with $k = 10$. The histogram shows the digit label distribution in each cluster. Our method reveals similar digits in the dataset and places them into the same cluster. Cluster 7 contains mostly digits 4 and 9. Both of these digits have very similar structures. Similarly, cluster 3 reveals that 3 and 5 are similar to each other and additionally 8 exhibits similarities to this digit pair. These results indicate that although the classifier is trained with the 10 digit labels, WS-RFClust is able to uncover similarity of digits other than the training class.

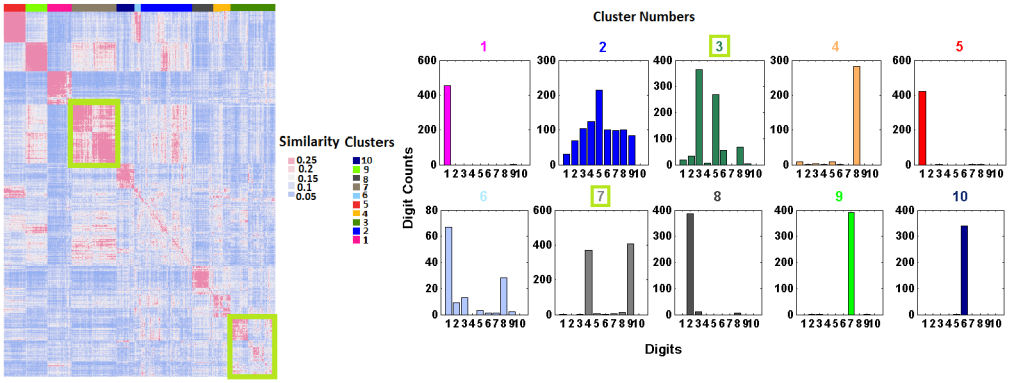


Figure 5.1: Clustering results of hand-written digit dataset with 5000 samples. Colors on the heatmap represent similarities computed for sample pairs (reds indicate high similarity, blue indicates low). The bars on top indicate different clustering. Each subplot that bears the same color on the histogram displays the digit content of the clusters based on their true class labels. x-axis of a histogram represents digits and y-axis represents the number of observed samples in each digit. The two interesting clusters, 3 and 7, are marked with green boxes.

To validate 10 digit class clustering, we calculated the silhouette width of resulting hierarchical clustering. Figure 5.2 shows the silhouette width of each cluster. The width of many clusters are positive, and assuming the average silhouette width is closer to 1, this means we obtain good clustering.

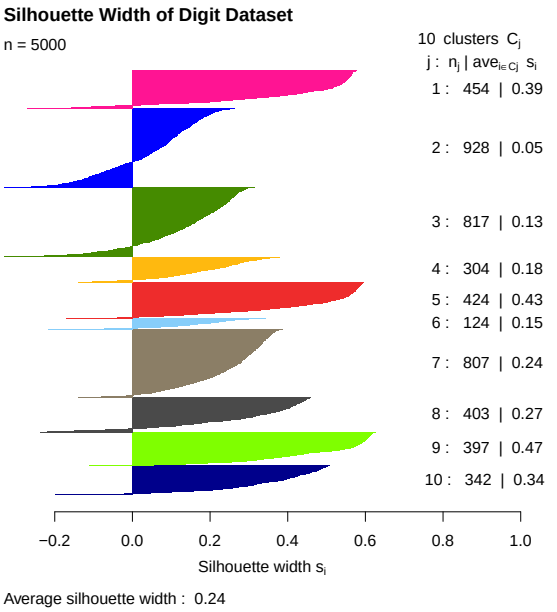
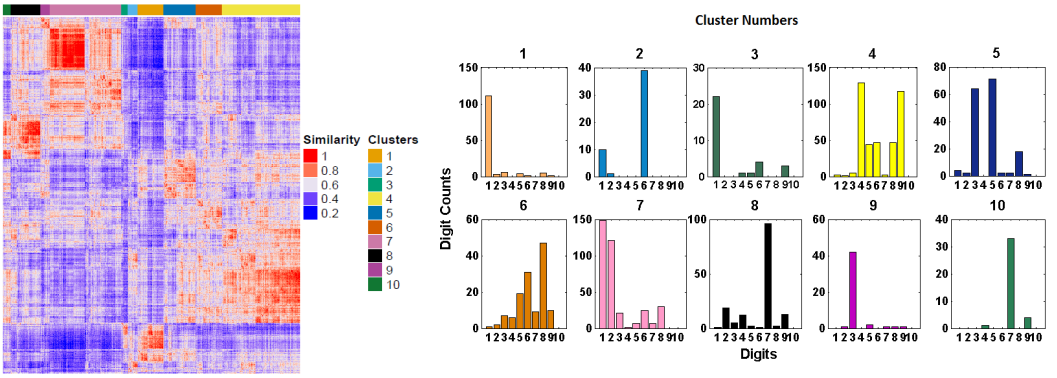


Figure 5.2: The silhouette width of each cluster in 10 digit classification where 5000 MNIST handwritten digit samples are used for training. y axis shows number of members in each cluster and its silhouette width. x axis is a ruler showing the silhouette width of each cluster.

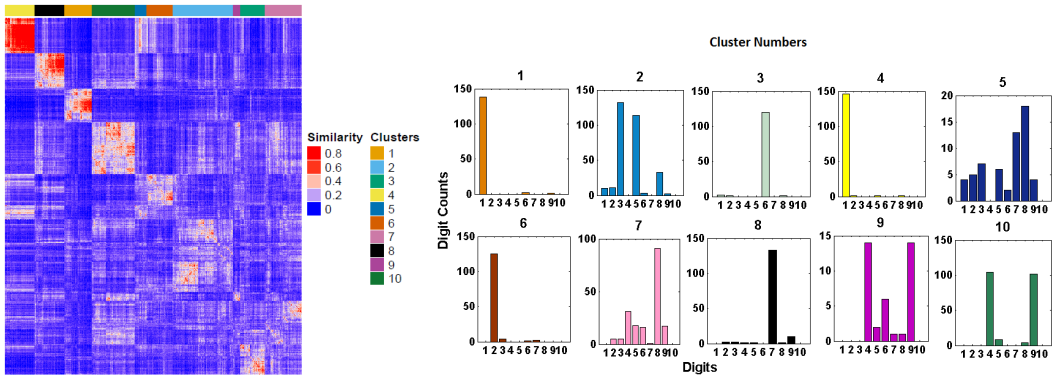
5.1.2 Effect of Sampling from Interval Nodes at Different Depths

In applying WS-RFClust, the depth levels in each tree are chosen randomly but within a predefined depth range. To understand the effect of the sampling depth we experiment with different ranges: let h be the height of a tree in the forest. We experiment with selecting d from the interval lower part of the tree that is from (0-$(h * 1/3)$] , the middle part from [$(h * 1/3)$-$(h * 2/3)$] and third interval is from $[(h * 2/3), h]$. To speed up calculations, for training random forest classifiers, we sample 1500 examples in these experiments.

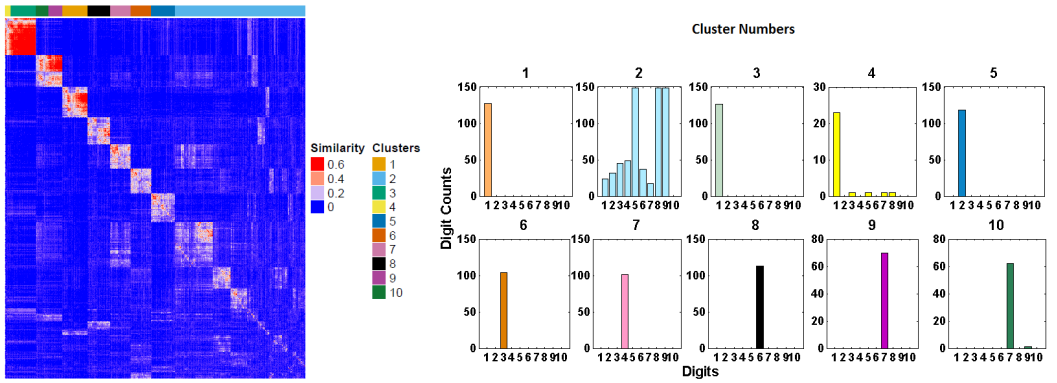
Figure 5.3a displays the results where the intervals are selected from the interval nodes closer to the root. In cluster 8, we observe that the digits 4 and 9 are in the same cluster, these are digit pairs with very similar shapes. In cluster 2, in figure 5.3b, where the intervals are sampled in the medium part of the trees, the grouping of digit 4 and 9 is more clear (cluster 9). Similarly, cluster 2 reveals not only that 3 and 5 are similar but also digit 8 is similar to these digits. Figure 5.3c shows that clustering is not possible when we sample depths from nodes closer to the leaves. Classification accuracy of 1500 samples is 90%, so we can obtain reliable results.



a. $0 < d \leq \frac{h}{3}$



b. $\frac{h}{3} \leq d \leq \frac{2h}{3}$



c. $\frac{2h}{3} \leq d \leq h$

Figure 5.3: Heatmaps and histograms of digit clustering computed at different depth levels and with 1500 digit samples. The similarity column of heatmap shows the similarity rate of paired samples obtained from the distance algorithm. Reds show high similarity rates, while blues show low similarity rates. Each colorful rectangle in the Clusters column represents a cluster. Histograms show the distribution of digit amounts in each cluster. x axis of a histogram represents digits, y axis represents the count of each digit in that cluster. When $\frac{h}{3} \leq d \leq \frac{2h}{3}$, we obtain most effective clustering that cluster 2 reveals similar 3-5-8 digits, cluster 9 and 10 reveals similar 4-9 digits.

After comparing different interval ranges, we observe that sampling d close to the leaves of the trees (the third interval range) results in clusters that are consistent with class labels. On the other hand, running WS-RFClust with depths that are closer to the root results with too impure clusters. We think this is because the depths chosen close to the roots leads to feature combinations that are too general, thus revealing the similarities of the samples is not possible. Therefore, the sampling in the medium part of the tree is more likely to reveal different clusterings.

5.1.3 Discovering Clusters Under Uniform Label Noise

Biological and clinical data are often noisy. Therefore we wanted to test how WS-RFClust will perform when the labels that guide the clusters are noisy. We corrupt the label information of digit dataset by adding uniform random noise. Let p be the predefined noise level and $\bar{y}_i$ be the corrupted label for the instance i. We sample a new class label $P\{\bar{y}_i = \bar{c}\ y_i = c\} = p$ where $\bar{c} \in \{0, 1, \ldots, 9\}\ c$. Inserting noise is not enough to reduce the accuracy; therefore, we also reduce the training set size to 1000. We take equal amount (100) samples from each digit class. The test set size remains as 1000. Table 5.1 includes the range of noise values and the corresponding test accuracy of the different models trained with the noisy label set.

Noise Probability	Accuracy
0.05	0.87
0.10	0.87
0.20	0.86
0.25	0.86
0.30	0.86
0.35	0.85
0.40	0.82
0.45	0.80
0.50	0.77

Table 5.1: Accuracy with different noise values.

We apply WS-Clust on the label set where the noise level is 0.5. The corresponding Heatmap and histogram of clusters are shown in Figure 5.4. The results display that although the guiding labels are noisy, the feature representations reveal the structure of the digits. For example, the cluster 3 points out that two digits, 3 and 5 are similar. Cluster 7 points out that digits 4 and 9 are similar in their structure. We conclude although the classifier that is learned is not a very good one, the feature representations are able to decode the structure.

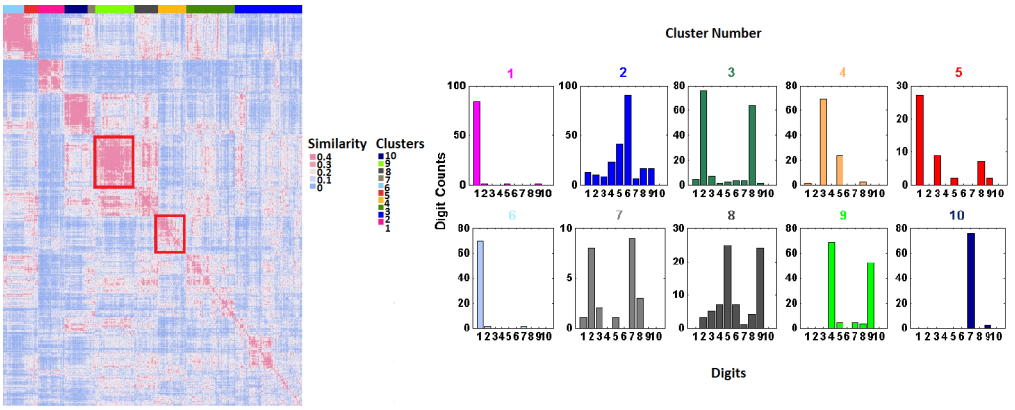


Figure 5.4: Similarity matrix and corresponding histogram in noisy MNIST handwritten digit data. noise=0.5, number of samples=1000. Colors in clusters column are consistent with the heatmap annotation and histogram. Cluster 4 points out that two digits, 3 and 5 are similar. Cluster 9 points out digits 4 and 9 are similar in their structure. Cluster 4 and Cluster 9 are marked with red boxes.

5.2 Results in Cancer Dataset

In this section we apply WS-RFClust in breast cancer expression datasets. We also run the widely adapted method NMF-Consensus clustering on each of these datasets. Results from these runs are elaborated in the ensuing subsections.

5.2.1 mRNA Results in WS-RFClust

mRNA expression data contain 20531 genes' expression values on 1196 patients samples. We first dichotomize the survival time of patients into two classes. To this end, we calculate the 25% lower and 75% upper quantiles; patients with survival time shorter than the 25% quartile are labeled as low survivors, whereas patients with survival times longer than the 75% quartile are labeled as high survivors. These labels constitute the true class labels for random forest classification. 1196 patients are reduced to 599 after selective filtering for high and low survivor patients. Number of long survivors are 299 and number of short survivors are 300.

We apply different feature selection criteria including ttest, ROC, Entropy, Chernoff and Wilcoxon statistical tests to reduce the number of features. Let n be number of top-ranked features in our dataset. We experiment with different n values and select the top n feature sets that produce the best 5-fold cross-validation accuracy. We apply 5-fold cross validation to the training data 10 times and form decisions based on the average accuracy over 10 runs. Results

of different rank tests are listed in Table 5.2.

# of selected features	ttest	roc	bhattacharyya	entropy	wilcoxon
25	0,69	0,61	0,50	0,51	0,67
50	0,70	0,62	0,50	0,51	0,69
100	0,72	0,63	0,59	0,59	0,69
200	0,72	0,64	0,63	0,57	0,69
500	0,71	0,64	0,65	0,60	0,70
750	0,70	0,62	0,64	0,61	0,69

Table 5.2: Accuracy with different feature selection method and number of features in mRNA expression data.

We first want to check if the clustering based on this methodology can put the low and high survivors into the right classes. We divide the expression data into two parts as training and test matrices. Test samples are generated by getting 1/5 of all low-survivor and high-survivor patients. Accordingly, we operate with 480 training samples and 119 test samples. We train the random forests with 200 trees and apply RF-WSClust by sampling from depths from the interval [1/3 2/3] x height of the trees uniformly at random. We cluster samples using hierarchical clustering with cluster number 2.

The confusion matrix of test sample classification is provided in table 5.3 and KM survival plot is shown in Figure 5.5. Cluster 1 represents the low survivor patients; cluster 2 represents the high survivor patients. The accuracy of predicting test samples is 0.57. We also plot the survival distributions of these test samples and check their clusterings; survival distributions of two clusters are not well separated from each other. Using the log-rank test at a

significance level of 0.05, we also test the null hypothesis that the two clusters are not different from each other in terms of survival distribution. p-value is not lower than 0.05, therefore we cannot reject the null hypothesis. This is somewhat expected as patients are not perfectly stratified into two groups even with the random forest classifier (whose accuracy is at most 70%), therefore the clustering which does not focus on the prediction of the two classes cannot achieve better accuracy. This might also point that there are more than two subgroups that are different at the molecular level. microRNA and protein expression data give similar results.

		Predicted	
		Cluster 1	Cluster 2
True	Cluster 1	38	22
	Cluster 2	28	31

Table 5.3: Confusion matrix of class low survivor and high survivor. Accuracy of overall prediction is 0.57.

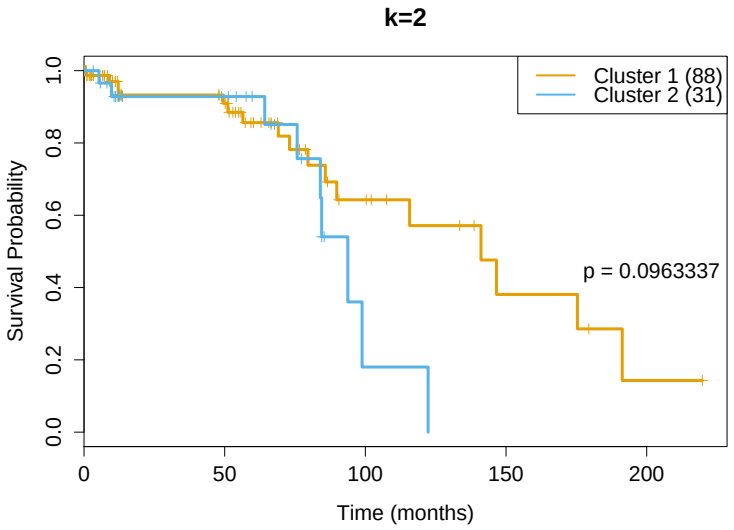
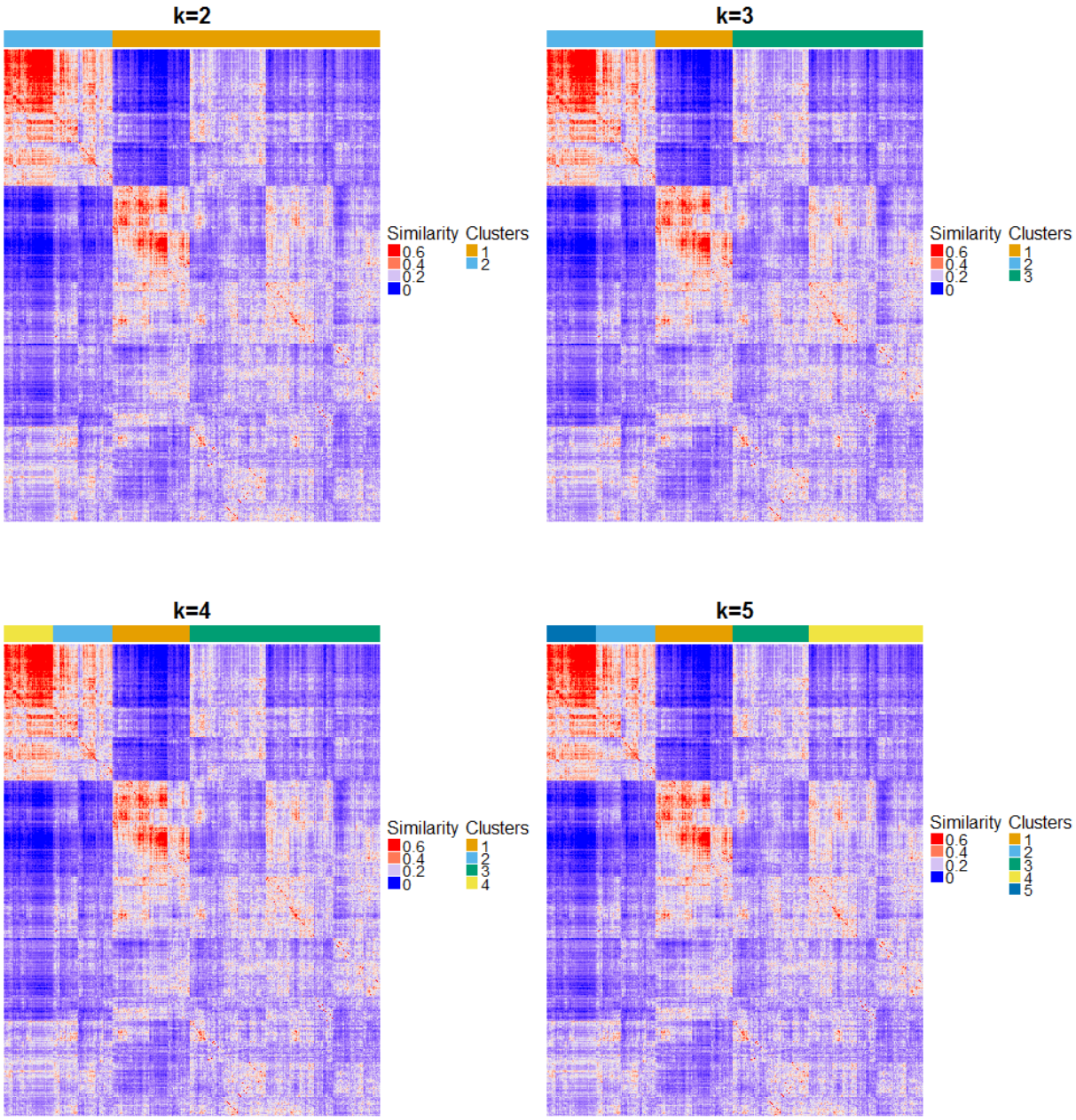


Figure 5.5: KM survival plots of the clusters obtained on 119 test samples used. Test samples are assigned to clusters based on WS-RFClust model with $k = 2$. The model is trained with mRNA expression data.

Finally, we apply WS-RFClust on all 1196 samples. Here we use the RF model that is trained with the Random Forest classifier trained on the low and high survival patients. Using the random forest we calculate their similarity matrix. With the output similarity matrix, we apply hierarchical clustering with different numbers. Let k denote the number of clusters, we try clustering with $k = 2, 3, 4, 5, 6$. We use our training model to cluster all the samples. Table 5.6 shows the heatmaps of different k values.


k=2
Similarity
0.6
0.4
0.2
0
Clusters
1
2
k=3
Similarity
0.6
0.4
0.2
0
Clusters
1
2
3
k=4
Similarity
0.6
0.4
0.2
0
Clusters
1
2
3
4
k=5
Similarity
0.6
0.4
0.2
0
Clusters
1
2
3
4
5

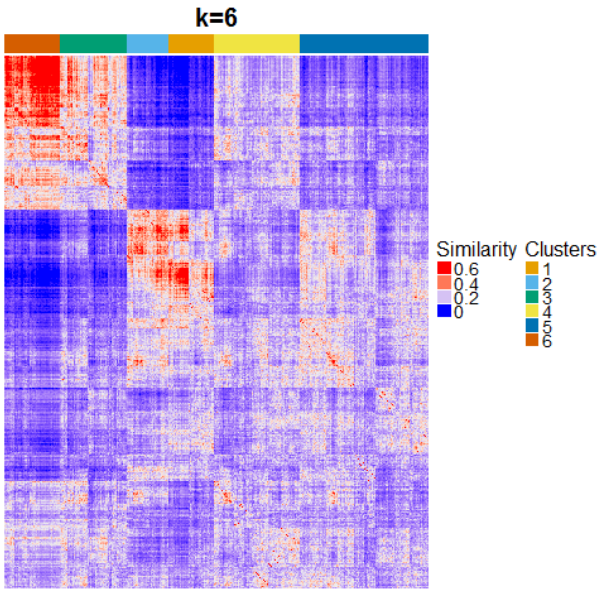


Figure 5.6: Heatmaps for different $k = 2, 3, 4, 5, 6$ for 1196 x 1196 patient similarity matrix in mRNA expression dataset. Colorful bars on top of heatmaps represent clusters, red color denotes high similarity, blue color denotes low similarity.

The silhouette width of the clustered data shows a degree of purity within a cluster and the quality of separation between clusters. Figure 5.7 indicates the silhouette width graph of clustered patients.

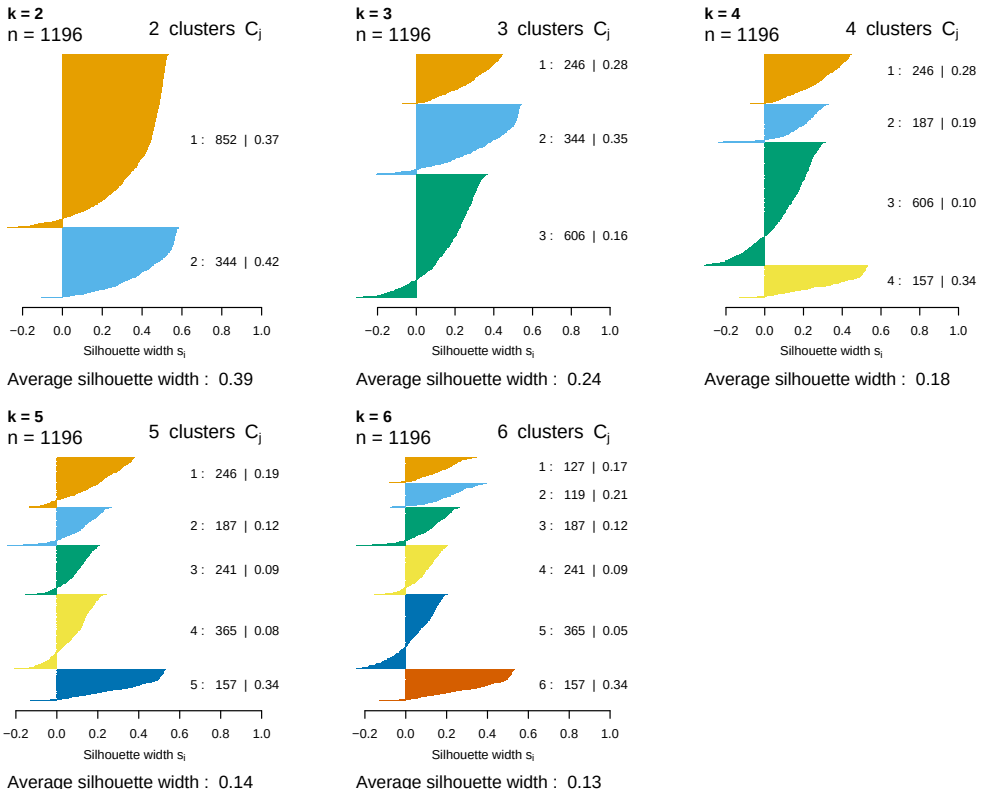


Figure 5.7: Silhoutte width graphics for k=2,3,4,5,6 in mRNA dataset. x axis is the ruler that shows the width of each cluster. j is cluster id, n_j is number of patients in cluster C_j and S_i is silhouette width of C_j. y axis shows $j : n_j|ave_{i \in C_j} S_i$ for each cluster. Average silhouette width is overall average of all clusters.

5.2.1.1 Comparison of survival distributions

The survival plot of a clustered dataset demonstrates the survival distribution of each subgroup. Using the log-rank test we test the clusters that do not differ in terms of cluster validity. Figure 5.8 shows the Kaplan-Meier survival plots

for $k = 2, 3, 4, 5, 6$. We achieve the best separation at $k = 5$ (4.5017e−05).

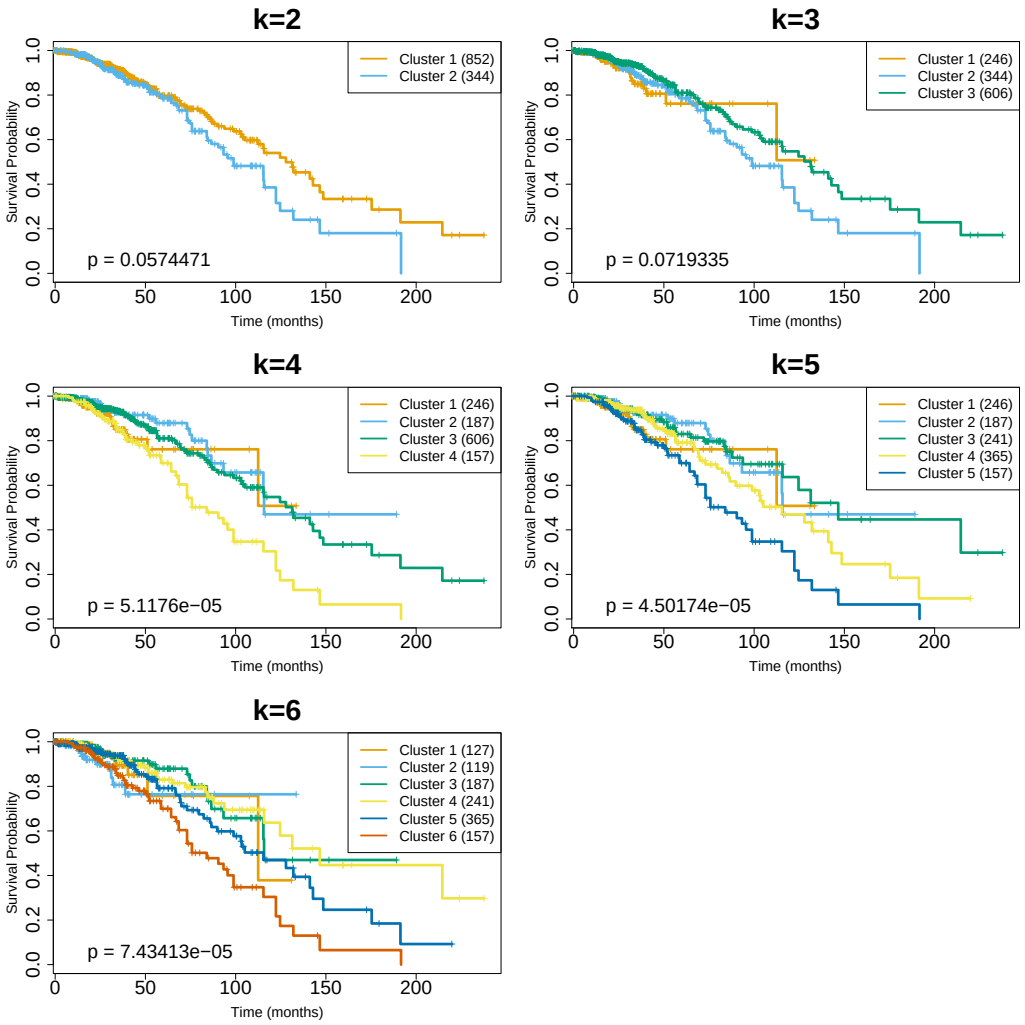


Figure 5.8: Survival plots for different k values pertaining to the model trained with mRNA. x axis shows the time of survival in months. y axis shows the survival probability at a given time. $k = 5$ gives smallest p-value, 4.5017e−05. Survival distributions of clusters are distinctive from each other at $k = 5$. There are five subgroups that are statistically different from each other.

5.2.1.2 Comparing clusters in terms of age

We applied one-way ANOVA test to compare the difference of mean ages between clusters. We reject the null hypothesis that their ages do not differ with $p = 6.5e{-}04$. Based on this test, these clusters are found to be significantly different from each other in terms of age. Figure 5.9 bears the box-plot of age distributions of clusters when $k = 5$. The variance is high in these box-plots. Although the statistical test rejects the null hypothesis, this difference could be attributed to different times when people are diagnosed at the clinic.

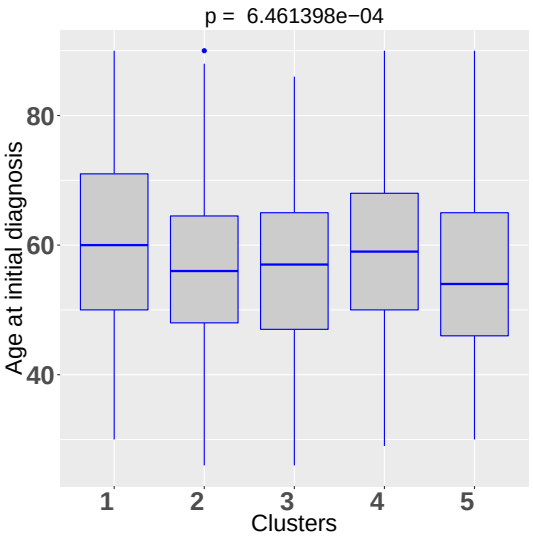


Figure 5.9: ANOVA comparison of age when k=5. y axis labels are patient ages, x axis labels are cluster ids. The start edge and the end edge of a boxplot indicates the range of ages in a cluster and line at the middle of the box shows the mean age value of patients in the cluster. Mean differences of clusters are significantly different.

5.2.1.3 Comparison with tumor stages

We also check if the clusters are significantly associated with the tumor stage. We tabulate the data into clusters and stages, and apply χ^2 test of independence. Null hypothesis is that WS-RFClust subgroups are independent of tumor stages. We deleted stages Stage IB, Stage II, Stage III, Stage Tis, Stage X, Stage IV; because only a few patients belong to this stages. Table 5.4 shows the relation between the tumor stage and the resulting cluster in k=5, $p = 0.03 < 0.05$.

	WS-RFClust Clusters				
Tumor Stages	**1**	**2**	**3**	**4**	**5**
Stage I	9	18	25	23	28
Stage IA	22	15	16	28	7
Stage IIA	75	63	83	119	50
Stage IIB	61	46	48	86	35
Stage IIIA	31	30	31	50	22
Stage IIIB	8	5	6	10	4
Stage IIIC	16	4	17	25	6

Table 5.4: Contingency table of tumor stages and WS-RFClust clusters. $\chi^2 = 38.569$, $df = 24$, $p = 0.03029$

5.2.1.4 Comparison with PAM50 subtypes

Pam50 subtypes are known as clusters of breast cancer. We apply χ^2 test of independence. $p < 2.2\text{e}{-16}$ of test is considerably smaller than 0.05, therefore WS-RFClust clusters have strong correlation with the intrinsic molecular subtypes. Table 5.5 shows the contingency table of WS-RFClust clusters and PAM50 subtypes.

	WS-RFClust Clusters				
PAM50 subtypes	**1**	**2**	**3**	**4**	**5**
Basal	2	42	47	29	22
Her2	10	8	14	34	1
LumA	25	90	121	155	43
LumB	41	24	25	99	5
Normal	1	20	12	1	85

Table 5.5: Contingency table of PAM50 subtypes and WS-RFClust clusters. $\chi^2 = 439.39,\ df = 16,\ p < 2.2e-16$

5.2.1.5 mRNA results in consensus NMF

We apply Consensus NMF [83] with the same mRNA dataset. We run the consensus NMF algorithm dataset with 1196 samples containing all the patients. We select 200 features by implementing ttest. Figure 5.10 demonstrates heatmaps derived from consensus NMF for k=2,3,4,5,6.

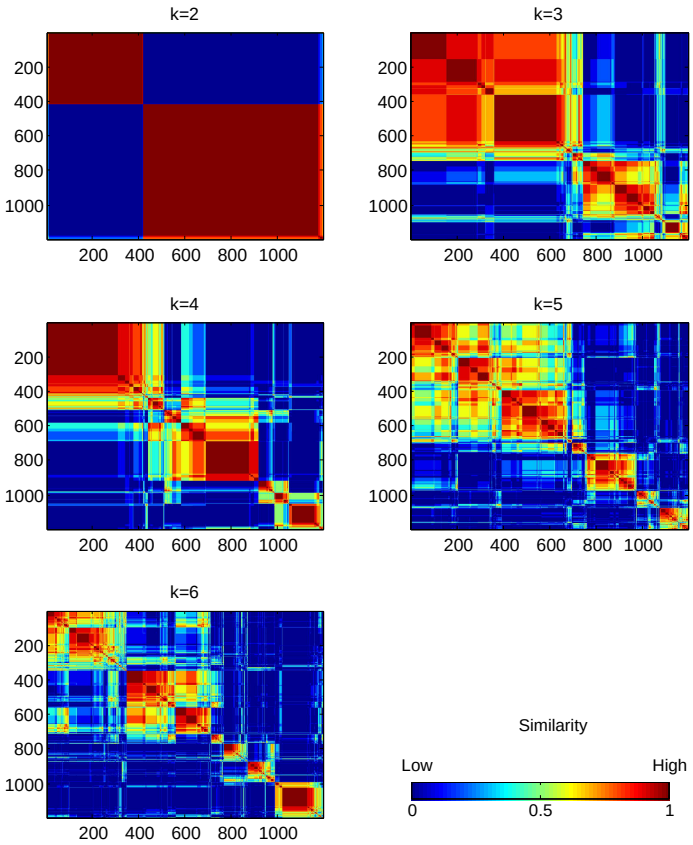


Figure 5.10: Heatmaps of consensus NMF run for k=2,3,4,5,6 on mRNA dataset. x and y axes show number of patients. Red regions show high similarity, while blue regions show low similarity rate.

We show that there is a correlation between the clusters found in WS-RFClust method and the survival rates of patients. Figure 5.11 demonstrates KM survival plots for each k value when consensus NMF is applied. Smallest p-value is achieved when k=6. $p = 0.000353525 \cong 3e-04$ value is larger than

p-value we obtained in WS-RFClust which is 4.50174e−05 in k=5. In $k = 5$, it is even 100 magnititude larger, clusters with $p = 4e-03$, which is 100 times less than WS-RFClust sensitivity. The smaller the p-value is, the more significantly the different clusters are obtained. WS-RFClust outperforms Consensus NMF in terms of survival rate differentiation between subgroups.

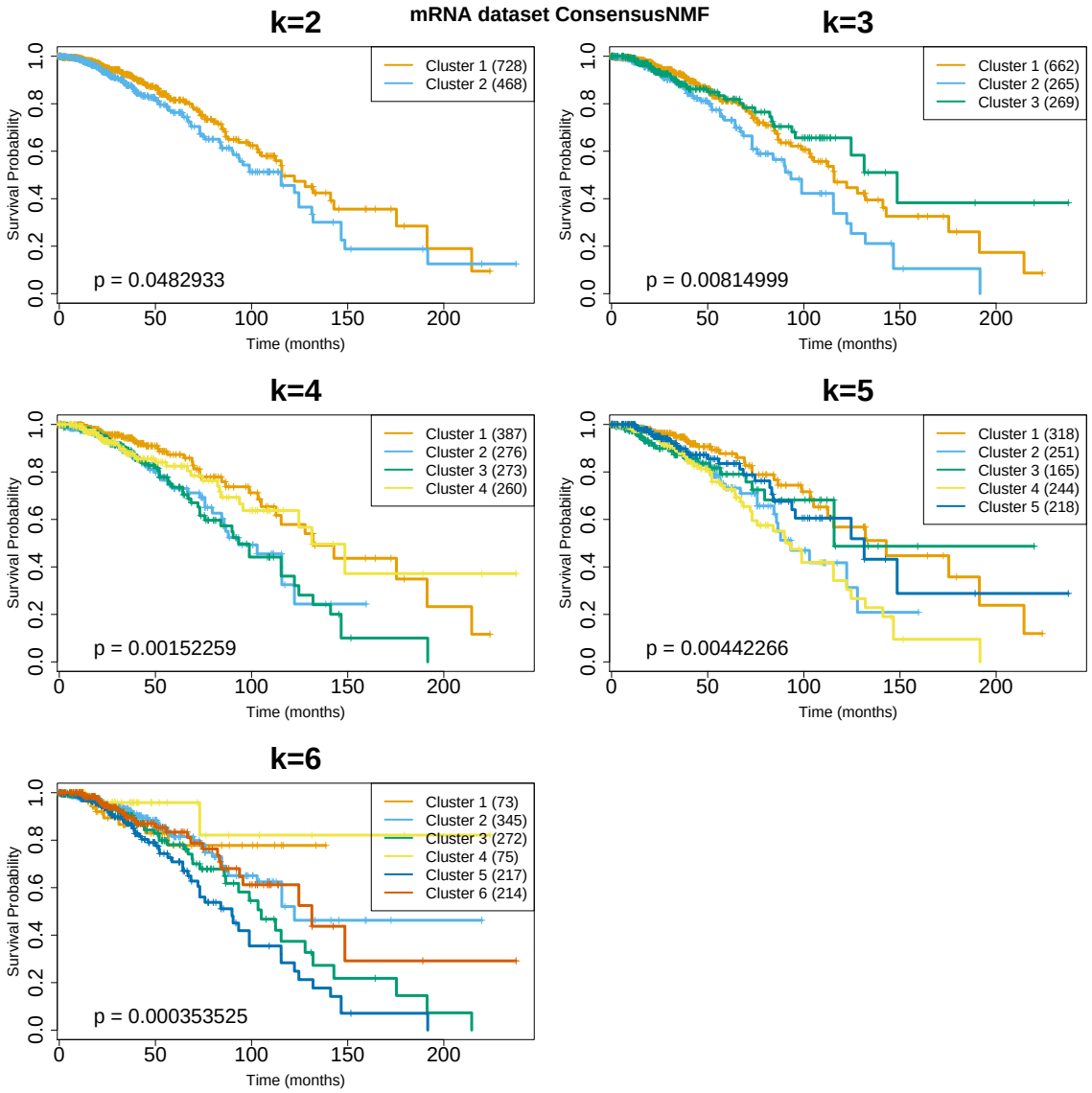


Figure 5.11: Survival plots of consensus NMF run for $k = 2, 3, 4, 5, 6$ on mRNA data.

5.2.2 microRNA Results in WS-RFClust

microRNA expression matrix contains entries for 1172 available patients and 1046 features. We apply our method to the miRNA dataset to measure the data dependency of WS-RFClust. We aim to verify that WS-RFClust detects intrinsic molecular subtypes that are independent of data type. We follow the same steps in training the mRNA expression data. We select only the low and high survivor 587 patients from miRNA expression data and select 200 features with t-test. We divide these patients into 476 training and 116 test examples. Test examples are classified with WS-RFClust. The accuracy of predicting test examples is 68%. The confusion matrix of test prediction results is given in Table 5.6. The survival plot of two classes is given in Figure 5.12. The accuracy of predicting test samples is 68%. p-value is $0.15505 > 0.05$, therefore we cannot state that this is a good stratification of low and high survivor patients.

		Predicted	
		Low Survivor	High Survivor
True	Low Survivor	44	14
	High Survivor	23	35

Table 5.6: Confusion matrix of class low survivor and high survivor. Accuracy of overall prediction is 0.68.

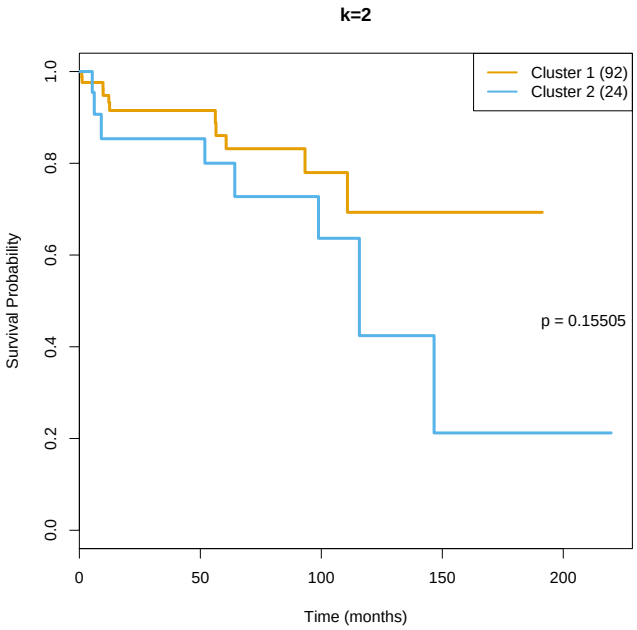
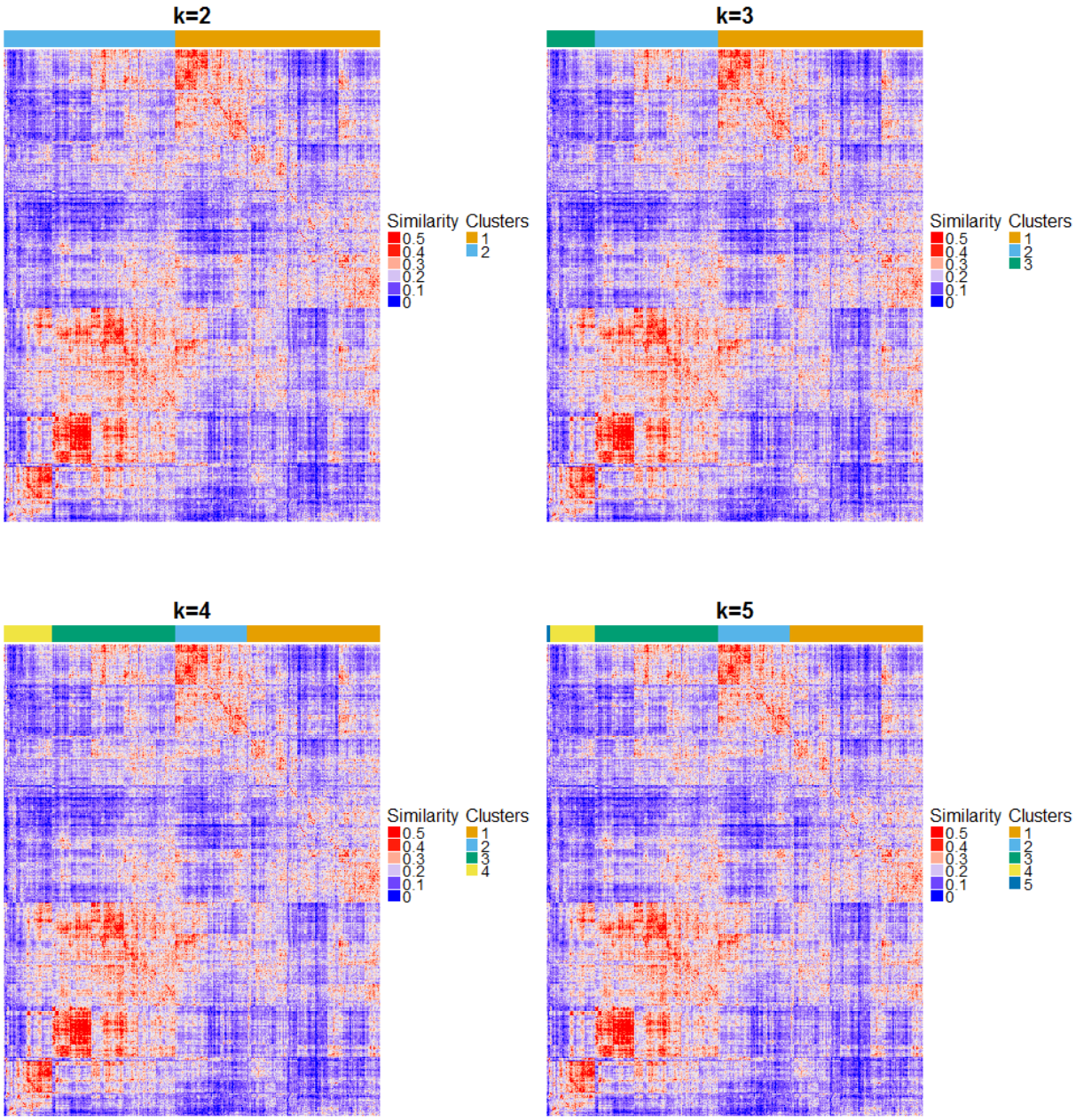


Figure 5.12: KM survival plot for 116 test samples in microRNA data. Cluster 1 represents high survivor patients; cluster 2 represents low survivor patients.

After random forest classification, we have a bag containing 200 trees. Then, all the patients (1172) that are available in the dataset are input to train model and WS-RFClust constructs a similarity matrix of patients. We apply hierarchical clustering for $k = 2, 3, 4, 5, 6$ as shown in Figure 5.13. Resulting clusters are compared with respect to survival rate, age, tumor stage and PAM50 subtypes.


k=2
Similarity
0.5
0.4
0.3
0.2
0.1
0
Clusters
1
2
k=3
Similarity
0.5
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
k=4
Similarity
0.5
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
4
k=5
Similarity
0.5
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
4
5

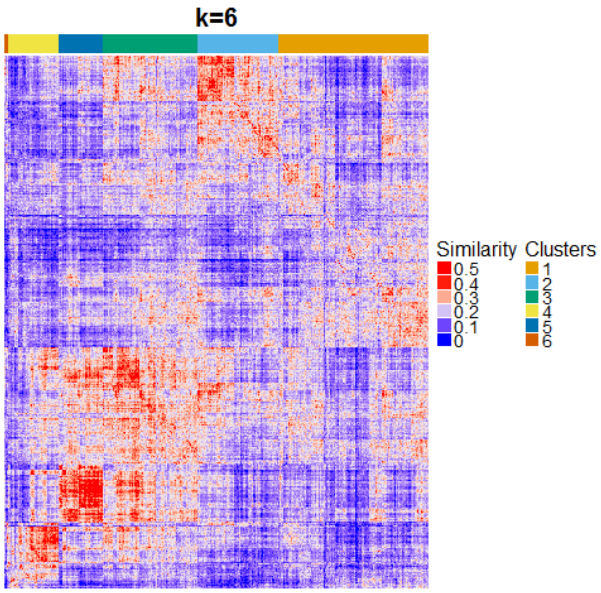


Figure 5.13: Heatmaps for different k=2,3,4,5,6 on 1172 x 1172 patients similarity matrix in microRNA expression data. Colorful bars on top of heatmaps represent clusters. Red color denotes high similarity, blue color denotes low similarity.

Figure 5.14 indicates silhouette width graph of clustered patients in microRNA expression dataset.

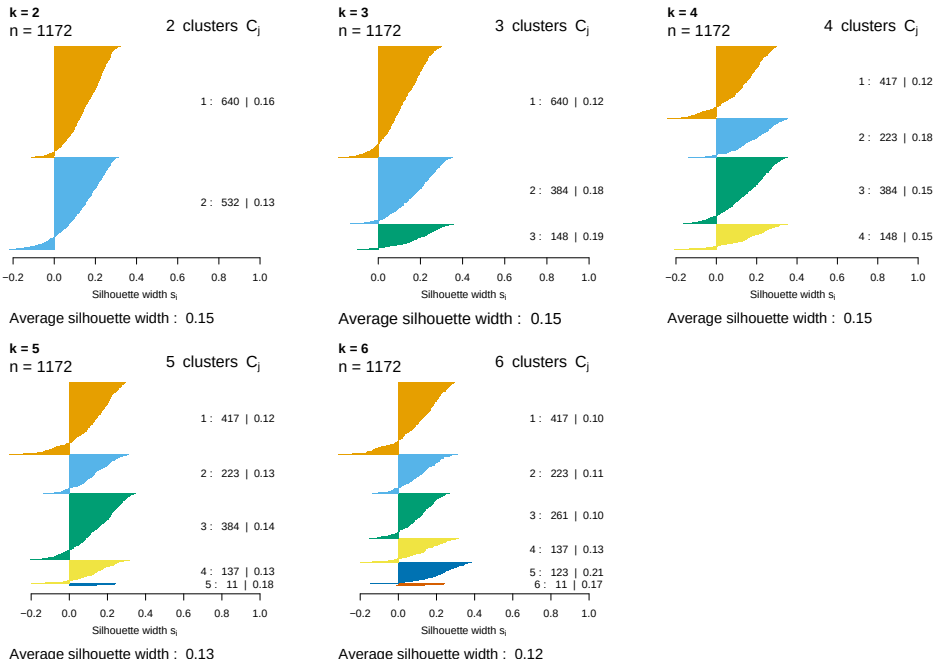


Figure 5.14: Silhouette width graphics for $k = 2, 3, 4, 5, 6$. x axis is the ruler shows the width of each cluster. j is cluster id, n_j is number of patients in cluster C_j and S_i is the silhouette width of C_j. y axis shows $j : n_j | ave_{i \in C_j} S_i$ for each cluster. Average silhouette width is the overall average computed over all clusters.

5.2.2.1 Comparison of survival distributions

Survival plots for all k values in Figure 5.15 point out that resulting subtypes have better separation in terms of survival when $k = 6$. Age, tumor stage and PAM50 subtype comparison is done between found clusters when k=6.

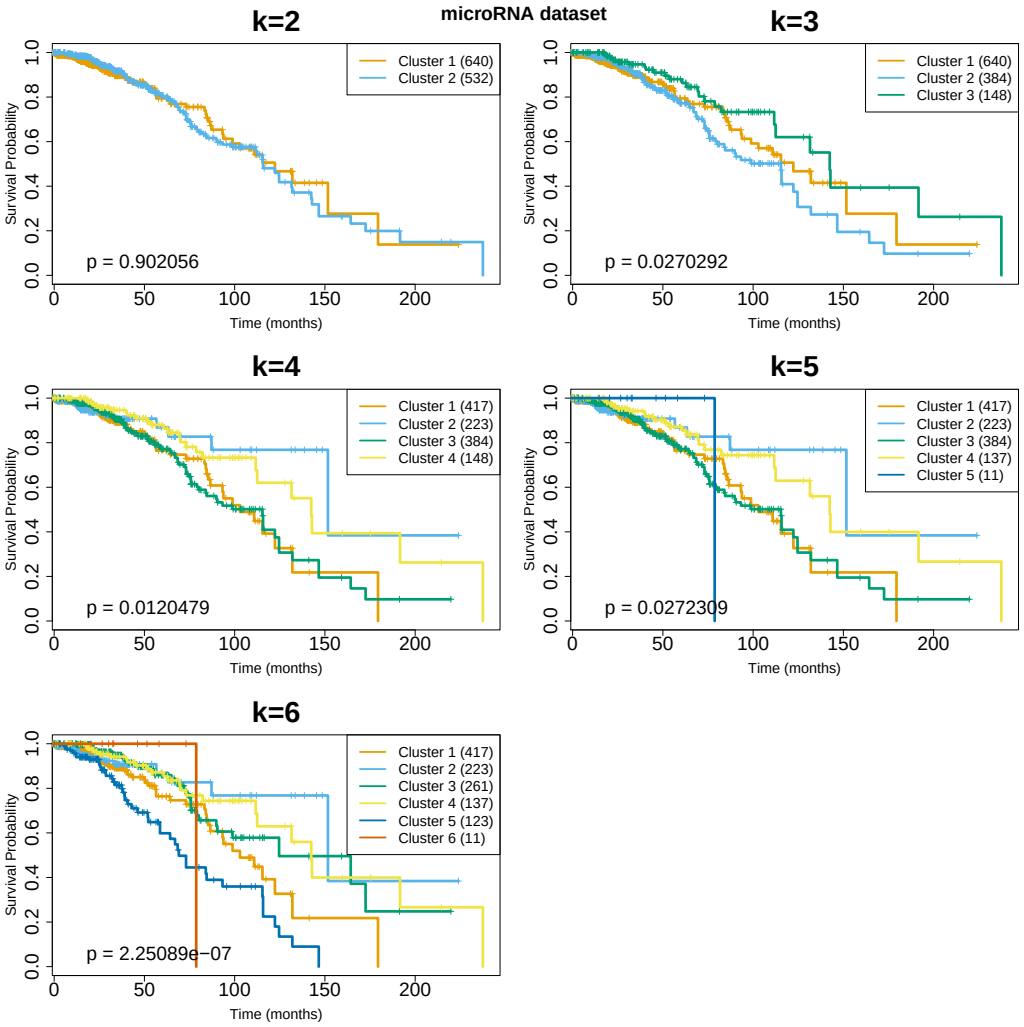


Figure 5.15: Survival plots of microRNA dataset for k=2,3,4,5,6. x axis shows time of survival in months. y axis shows survival probability at a time. $k = 6$ gives smallest p-value, 2.25089e−07. Survival distributions of clusters are distinctive from each other at $k = 6$, there are five subgroups that statistically different from each other.

5.2.2.2 Comparison of age distributions

We applied one-way ANOVA test to compare difference of mean ages between clusters. One tailed test is preferred to increase detection power. Figure 5.16 denotes boxplot of age distributions of clusters when k=6. $p = 2.55e{-}02 < 0.05$, we can conclude in 95% confidence interval that subgroups are significantly different in terms of age.

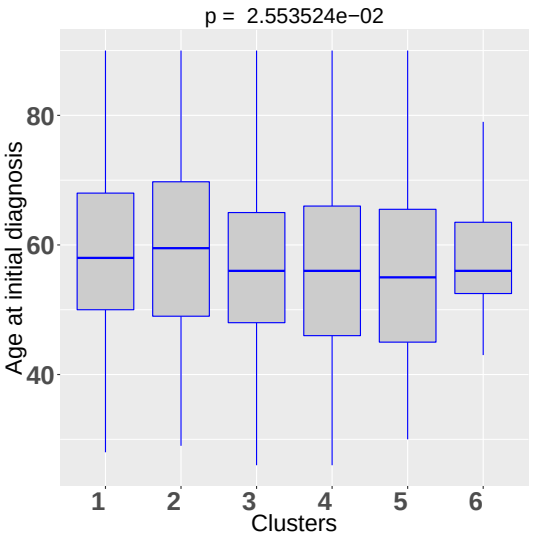


Figure 5.16: ANOVA comparison of age for $k = 6$, microRNA dataset. y axis labels are patient ages, x axis labels are cluster ids. Start edge and end edge of a boxplot shows range of ages in a cluster and line at the middle of the box shows mean age value of patients in the cluster.

5.2.2.3 Comparison with tumor stages

We tabulate the data into clusters and stages, and apply χ^2 test of independence for k=6. The null hypothesis is that the WS-RFClust subgroups are independent of tumor stages. We delete stages Stage IB, Stage II, Stage III, Stage IIIB, Stage Tis, Stage X, Stage IV; because there are only small numbers of patients belonging to those stages. Table 5.7 shows the relation between tumor stage and the resulting cluster in k=6. $p = 0.002 < 0.05$, therefore we reject the null hypothesis and we can conclude that tumor stages are correlated with WS-RFClust subtypes in miRNA dataset.

	WS-RFClust Clusters					
Tumor Stages	**1**	**2**	**3**	**4**	**5**	**6**
Stage I	33	10	29	10	18	0
Stage IA	24	23	25	9	5	1
Stage IIA	157	70	70	46	38	3
Stage IIB	90	52	65	28	30	2
Stage IIIA	60	22	40	17	22	1
Stage IIIC	21	24	10	5	3	2

Table 5.7: Contingency table of tumor stages and WS-RFClust clusters. $\chi^2 = 51.127$, $df = 25$, $p - value = 0.001544$

5.2.2.4 Comparison with PAM50 subtypes

We tabulate the data into clusters and subtypes, and apply χ^2 test of independence for the clustering results with $k = 6$. The resulting $p < 2.2e - 16$ of test is considerably smaller than 0.05. Therefore, WS-RFClust clusters show strong correlations with the intrinsic molecular subtypes. Table 5.8 shows the contingency table of WS-RFClust clusters and PAM50 subtypes.

	WS-RFClust Clusters					
PAM50 subtypes	**1**	**2**	**3**	**4**	**5**	**6**
Basal	95	5	11	24	1	0
Her2	41	8	7	10	0	0
Luma	91	61	174	72	27	5
LumB	104	23	27	26	5	2
Normal	9	7	5	3	83	4

Table 5.8: Contingency table of PAM50 subtypes and WS-RFClust clusters. $\chi^2 = 646.56,\ df = 20,\ p - value < 2.2e - 16$

5.2.2.5 microRNA results in consensus NMF

We apply Consensus NMF to the microRNA dataset to compare clustering performance of WS-RFClust with Consensus NMF. We run the consensus NMF algorithm dataset with 1172 samples containing all patients. We select 200 features by implementing t-test in order to make a fair comparison. Figure 5.17 demonstrates heatmaps derived from Consensus NMF for $k = 2, 3, 4, 5, 6$.

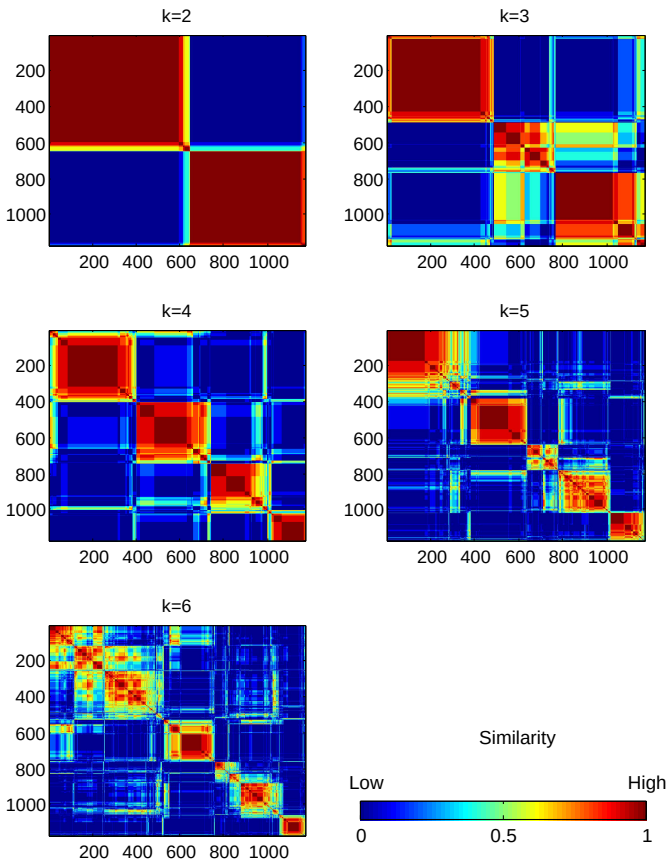


Figure 5.17: Heatmaps of consensus NMF run on microRNA dataset for $k = 2, 3, 4, 5, 6$. x and y axes show the number of patients. The similarity matrix stores 1172 patients. Red regions show high similarity, while blue regions show low similarity rate.

Figure 5.18 demonstrates kaplan-meier survival plots for each k value when consensus NMF is applied. Smallest p-value($p = 6.17428ee{-}05$) is achieved when k=5. This value is larger than p-value we obtained in WS-RFClust which is $p = 2.25089e{-}07$ in k=5. WS-RFClust provides more sensitive seperation

of clusters.

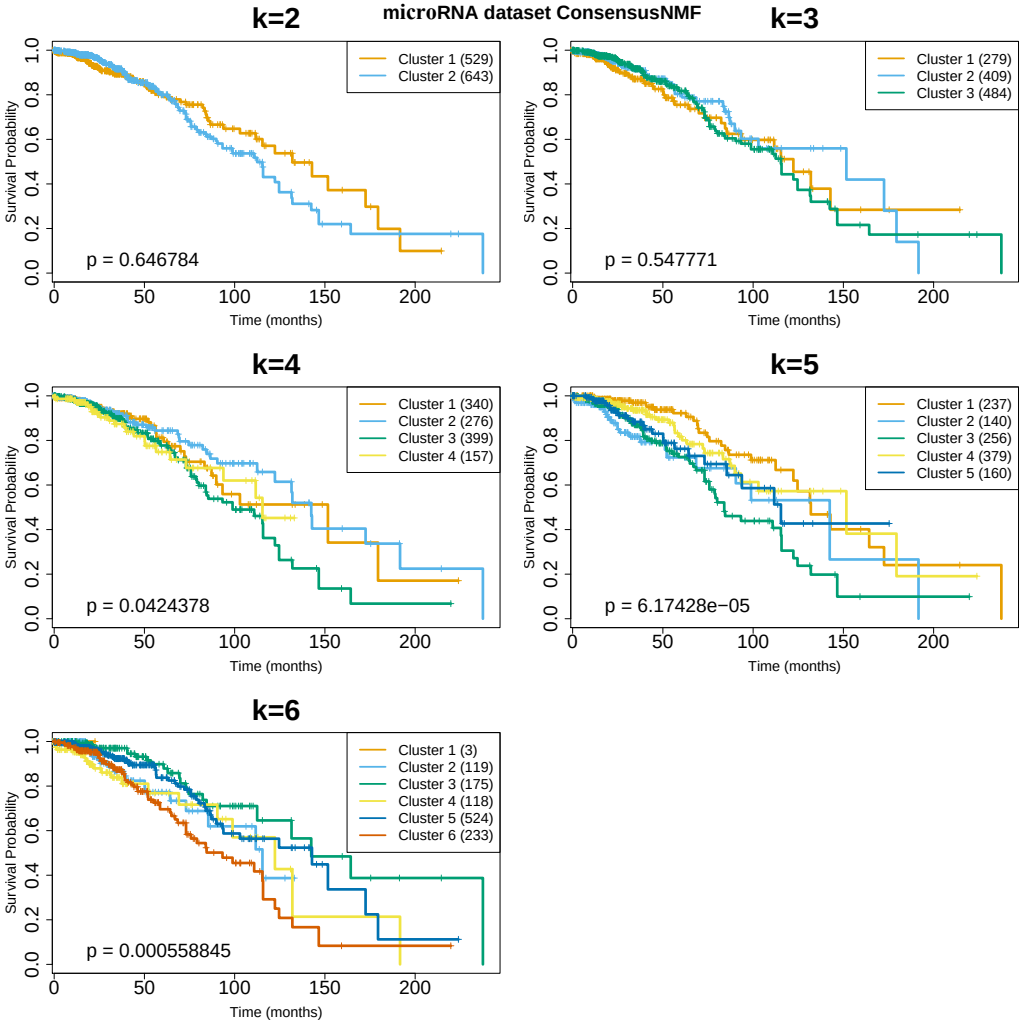


Figure 5.18: Survival plots of consensus NMF run on microRNA dataset for k=2,3,4,5,6. pvalue of ConsensusNMF when $k = 5$ is 100 times larger than the p value of WS-RFClust Therefore, WS-RFClust exhibits better performance in finding the clinically relevant survival subgroups.

5.2.3 RPPA Results in WS-RFClust

Protein expression matrix comprises information on 744 available patients and 131 features. The protein expression data contains 373 low and high survivor patients. We divide 299 of them as the training set and 74 of them as the test set. There are only 131 features in the RPPA dataset; therefore, we select all the features without any feature selection. The training set is input to WS-RFClust and the model accuracy is calculated from the test examples, and accordingly they are classified into low and high survivor labels. The accuracy of test example classification is 67%. The confusion matrix of actual and expected low and high survivor patients are listed in Table 5.9. Figure 5.19 demonstrates the survival distribution of low and high survivors among test examples.

		Predicted	
		Low Survivor	High Survivor
Actual	Low Survivor	27	10
	High Survivor	14	23

Table 5.9: Confusion matrix of class low survivor and high survivor. Accuracy of overall prediction is 0.67.

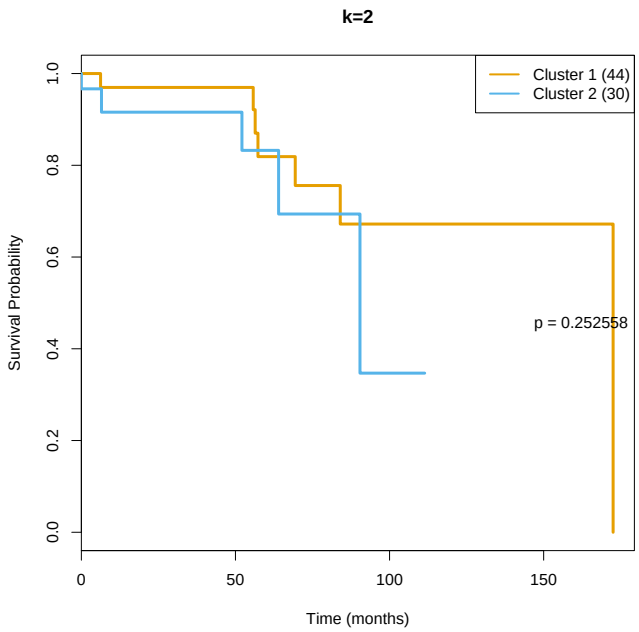
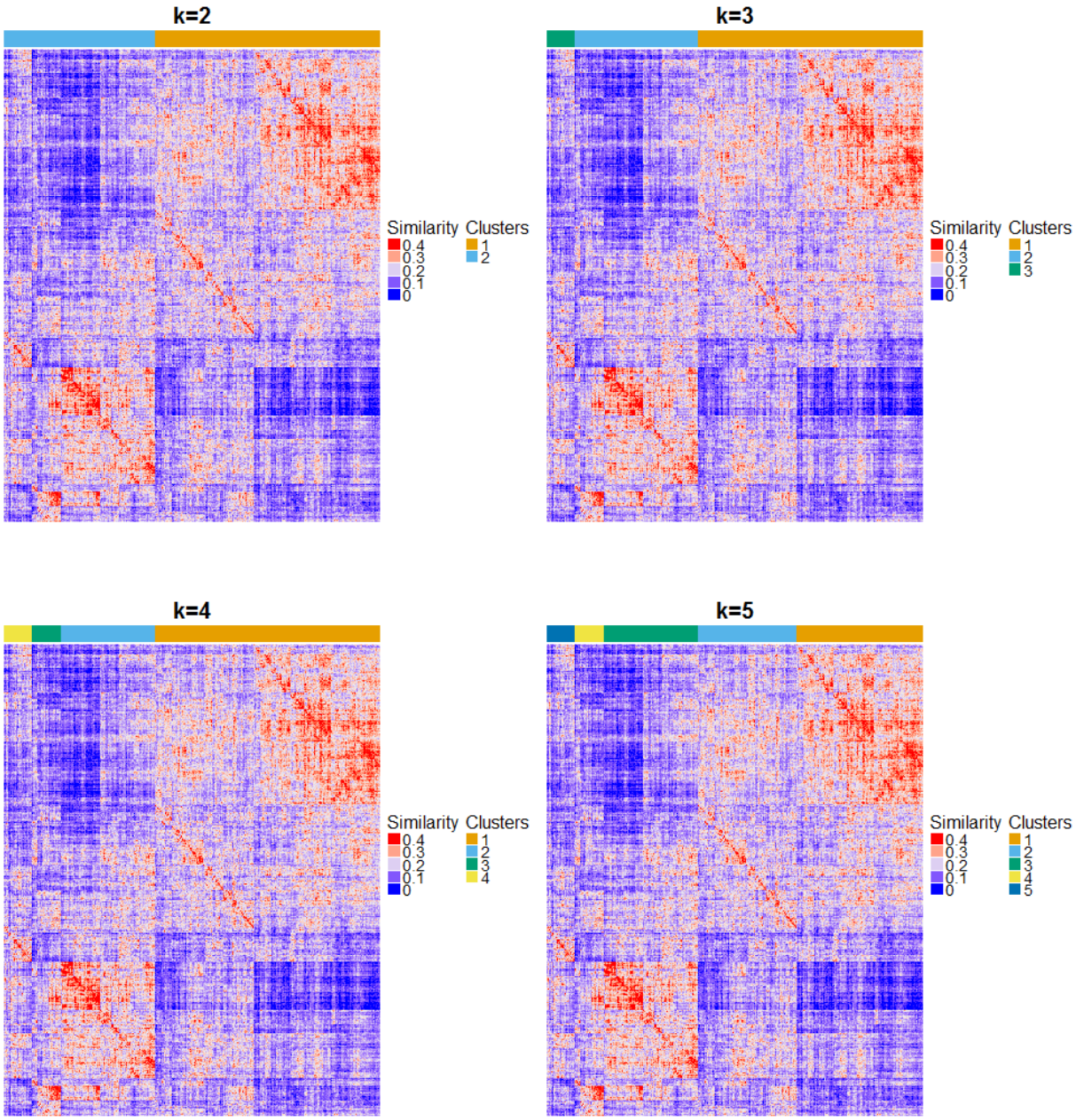


Figure 5.19: KM survival plot for 74 test samples in RPPA data. Cluster 1 represents high survivor patients, cluster 2 represents low survivor patients. Accuracy of predicting test samples is 67%. p-value is $0.252558 > 0.05$, therefore we cannot state that this is a good stratification of low and high survivor patients. However, accuracy of prediction is not ignorable and another point is steep accuracy is not a requirement in the success of WS-RFClust.

After random forest classification, we end up with a bag containing 200 trees. Then, all patients (744) available in dataset are input to train model and WS-RFClust constructs similarity matrix of patients. We apply hierarchical clustering for $k = 2, 3, 4, 5, 6$. The resulting clusters are compared with respect to survival rate, age, tumor stage and PAM50 subtypes.


k=2
Similarity
0.4
0.3
0.2
0.1
0
Clusters
1
2
k=3
Similarity
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
k=4
Similarity
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
4
k=5
Similarity
0.4
0.3
0.2
0.1
0
Clusters
1
2
3
4
5

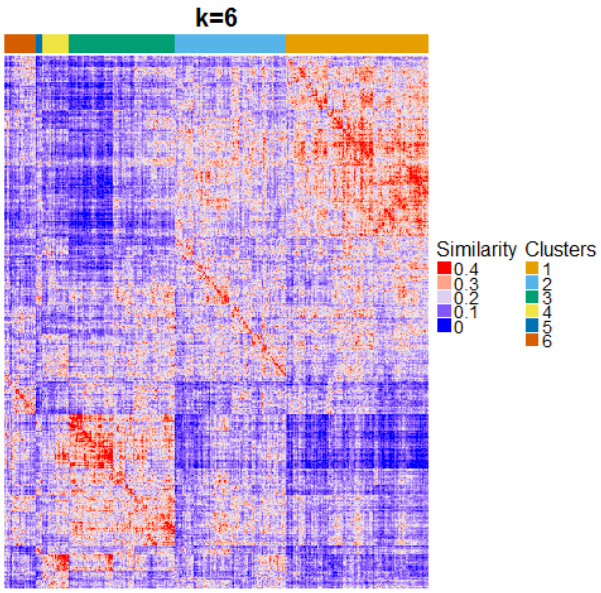


Figure 5.20: Heatmaps for different k=2,3,4,5,6 on 744 x 744 patients similarity matrix in RPPA expression data. Colorful bars on top of heatmaps represent clusters and "Clusters" column with rectangles maps cluster ids to colors. "Similarity" column shows similarity rate of patients resulted from Calc-RFrds. Red color denotes high similarity, blue color denotes low similarity.

Figure 5.21 indicates silhouette width graph of clustered patients in protein expression dataset.

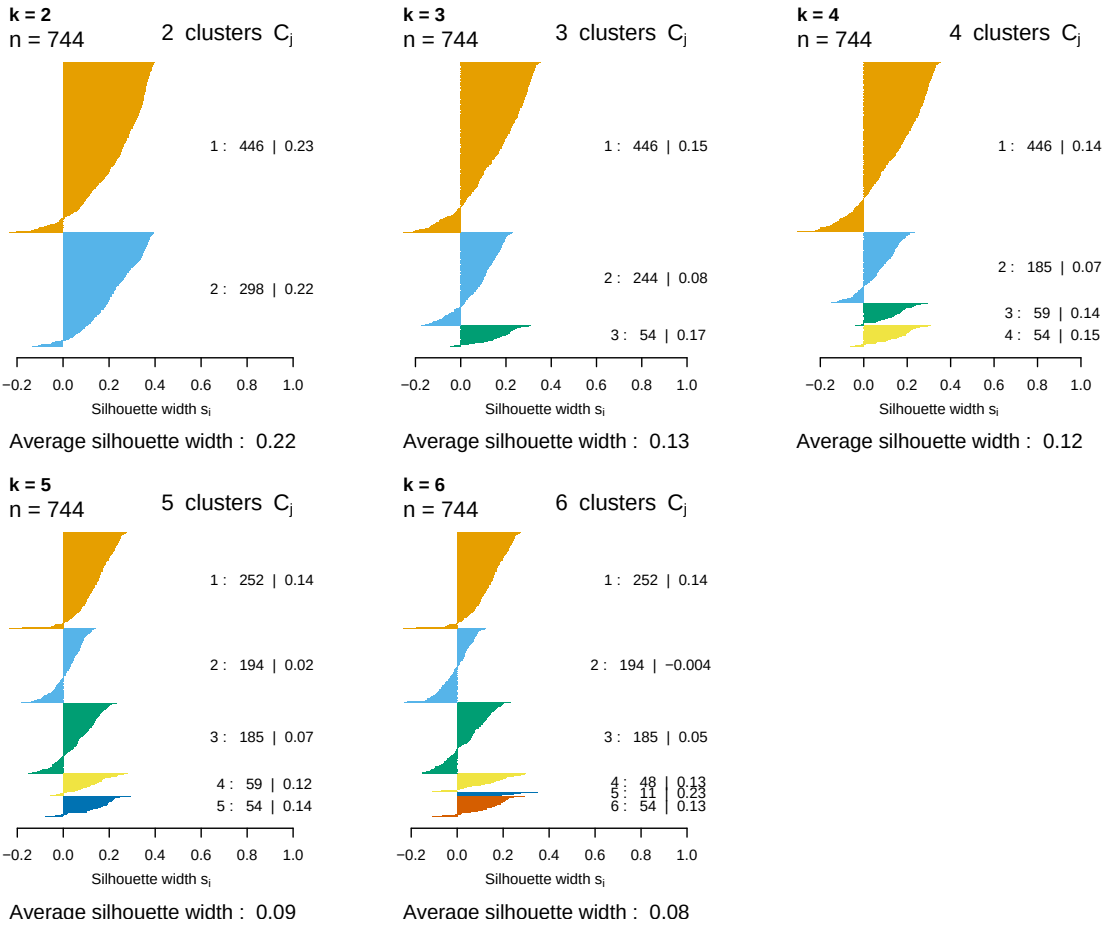


Figure 5.21: Silhoutte width graphics for k=2,3,4,5,6. x axis is the ruler shows width of each cluster. j is cluster id, n_j is number of patients in cluster C_j and S_i is silhouette width of C_j. y axis shows $j : n_j | ave_{i \in C_j} S_i$ for each cluster. Average silhouette width is the overall average of all clusters.

5.2.3.1 Comparison of survival distributions

Survival plots for k=2,3,4,5,6 are demonstrated in Figure 5.22. p-value is considerably small for each k-value, we select k=5 to correlate subtypes with mRNA results and PAM50 subtypes. Age, tumor stage and PAM50 subtype comparison is done to compare the clusters that are computed when k=5.

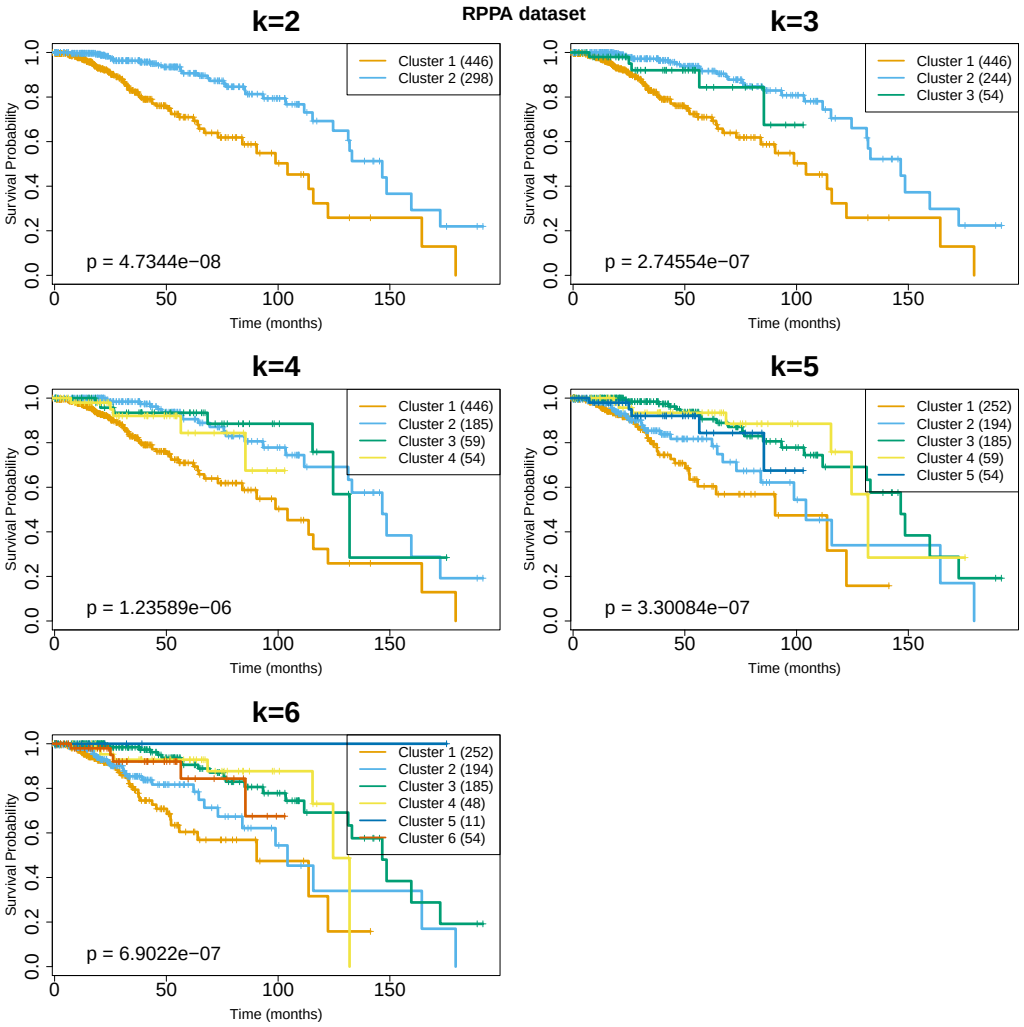


Figure 5.22: Survival plots of RPPA dataset for k=2,3,4,5,6. x axis shows the time of survival in months. y axis shows survival probability at a time. All k values give considerably small p-values, we select the case $k = 5$ to be consistent with mRNA dataset and PAM50 subtypes. The survival distributions of clusters are distinctive from each other at $k = 5, p =$ 3.30084e−07, there are five subgroups that are statistically different from each other.

5.2.3.2 Comparison of age distributions

We apply one-way ANOVA test to compare the difference of mean ages between clusters. Figure 5.23 shows the box-plot of age distributions of clusters when k=5. $p = 1.327781\text{e}{-03} < 0.05$, we can conclude in 95% confidence interval that subgroups are significantly different in terms of age.

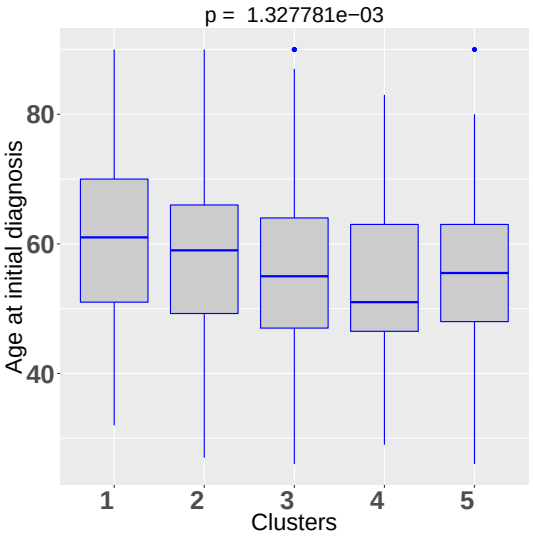


Figure 5.23: ANOVA comparison of age when k=5, RPPA dataset. y axis labels are patient ages, x axis labels are cluster ids. The start edge and the end edge of a box-plot indicates the range of ages in a cluster and the line in the middle of the box shows the mean age value of patients in the cluster. Mean differences of clusters are significantly different.

5.2.3.3 Comparison with tumor stages

We tabulate data into clusters and stages, and apply χ^2 test of independence for k=5. Null hypothesis is that WS-RFClust subgroups are independent of tumor stages. We delete stages Stage IB, Stage II, Stage III, Stage IIIB, Stage Tis, Stage X, Stage IV; because there are only a few patients belonging that stages. Table 5.10 shows relation between tumor stage and resulting cluster in k=5. $p = 0.02 < 0.05$, therefore we can safely reject the null hypothesis and we can conclude that tumor stages are correlated with WS-RFClust subtypes.

	WS-RFClust Clusters				
Tumor Stages	**1**	**2**	**3**	**4**	**5**
Stage I	12	12	17	8	8
Stage IA	24	8	9	3	8
Stage IIA	84	73	52	23	16
Stage IIB	58	42	47	16	12
Stage IIIA	43	28	35	4	6
Stage IIIC	9	15	8	2	2

Table 5.10: Contingency table of tumor stages and WS-RFClust clusters. $\chi^2 = 34.76$, $df = 20$, $p - value = 0.02142$

5.2.3.4 Comparison with PAM50 subtypes

We tabulate the data into clusters and subtypes, and applied χ^2 test of independence for k=5. $p < 2.2e - 16$ of test is considerably smaller than 0.05,

therefore WS-RClust clusters have strong correlation with intrinsic molecular subtypes. Table 5.11 shows the contingency table of WS-RFClust clusters and PAM50 subtypes.

	WS-RFClust Clusters				
PAM50 subtypes	**1**	**2**	**3**	**4**	**5**
Basal	8	44	19	43	6
Her2	16	29	7	5	4
LumA	130	54	111	0	35
LumB	87	51	18	3	5
Normal	1	1	13	2	1

Table 5.11: Contingency table of PAM50 subtypes and WS-RFClust clusters. $\chi^2 = 299.16,\ df = 16,\ p < 2.2e-16$

5.2.3.5 RPPA results in consensusNMF

We apply Consensus NMF to the protein expression dataset to compare the clustering performance of WS-RFClust with ConsensusNMF. We run the consensus NMF algorithm dataset with 744 samples containing all the patients. We select 200 features by implementing ttest in order to make a fair comparison. Figure 5.24 demonstrates the heatmaps derived from consensus NMF for k=2,3,4,5,6.

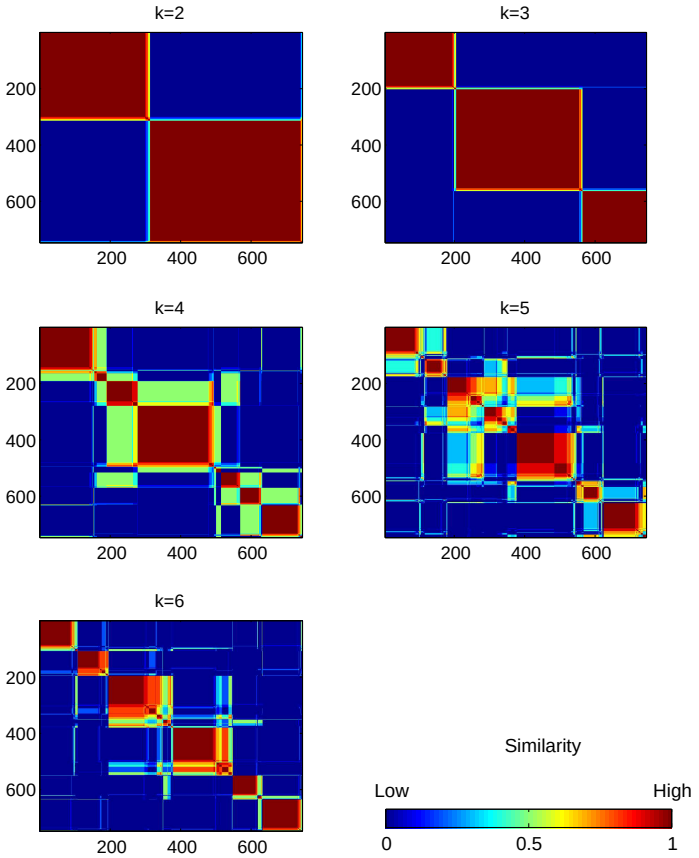


Figure 5.24: Heatmaps of consensus NMF run on RPPA dataset for k=2,3,4,5,6. x and y axes show the number of patients. Similarity matrix contains data for 744 patients. Red regions show high similarity, while blue regions show low similarity rate.

Figure 5.25 demonstrates kaplan-meier survival plots for each k value when consensus NMF is applied. p-value is $p = 0.0550391$ when k=5, overall p-value range is between $0.01 - 0.2$. Consensus NMF results are not confidently below

$\alpha = 0.05$, therefore we conclude that Consensus NMF clusters are not significantly different in terms of survival rate. WS-RFclust outperforms Consensus NMF in terms of survival rate differentiation between subgroups.

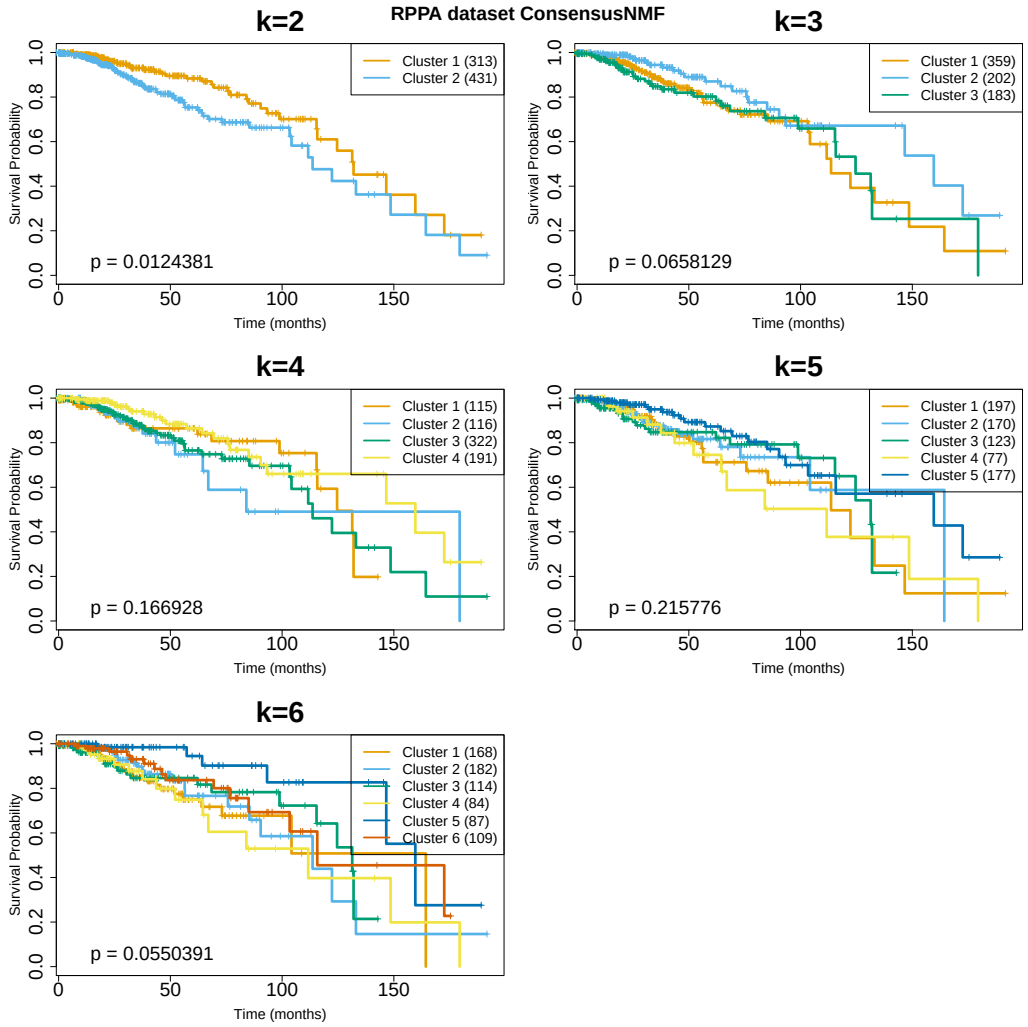


Figure 5.25: Survival plots of consensus NMF run on RPPA dataset for k=2,3,4,5,6. For all k values, Consensus NMF results are not confidently below $\alpha = 0.05$. Correspondingly, pvalue range of WS-RFClust is between e−06 and e−08. Therefore, WS-RFClust performs considerably better in stratification of patients.

Chapter 6

Conclusion and Future Work

Cancer is a complex disease that includes multiple genomic alterations or dysfunction in molecular systems. Cancer is a heterogeneous disease with multiple subtypes and lack of knowledge on subtypes hinders developing effective targeted therapies and realizing the personalized medicine objective. With the advent of next-gen sequencing technologies, now it is possible analyze a large cohort of patients and record patient genomic alterations and expression dysregulations. This opens up opportunities to redefine the subtypes of cancers.

The widely adapted approach for finding such subtypes is to apply unsupervised clustering techniques on genomic data of patients. The clusters are deemed interesting if they are found to be associated with a clinical variable of interest. These clinical variable of interest; therefore, do not participate in clustering decisions. We propose a new approach, WS-RFClust, where the

clustering process is supervised with the clinical variable of interest. The supervision is achieved by learning a similarity metric with features that are selected to predict the clinical variable of interest. Specifically, WS-RFClust involves a random forest classifier training step to predict the clinical variable of interest. Then the internal nodes are used to derive a similarity metric among pairs of samples. This similarity metric is based on the subsets of feature representations within the random forest classifier. By sorting down the examples on to random depths in the tree and checking how often they appear in the same partition in the tree, we can we construct a similarity matrix. This similarity matrix can be input to any traditional clustering algorithm.

We applied WS-RFClust to handwritten digit datasets to understand the effect of several parameters. WS-RFClust reveals clusters that have structural similarities, for example 4 and 9 are often found in the same cluster. To understand, how the sampling from different levels of the tree would affect clustering, we vary the interval range from which we sample the random depths. We observe that if the depths are close to the tree height, the resulting partitions are found to be close to the leaves and therefore these clusters correspond to the 10 classes. If we choose depths that are near the root, then the structure information is lost. Thus, we conclude that the sampling from the interval range is critically important. As a second experiment, we investigate how the classifier performance affects the cluster identification step. Digit classification has accuracy around 90%, and by adding uniform label noise we degrade the class label quality. We repeat the same experiments with WS-RFClust algorithm using a classifier of 50% test accuracy. The results show that we are still able to capture clusters with similar shapes although the learned classifier accuracy degrades significantly.

Finally to identify breast cancer subtypes, we apply WS-RFClust to TCGA breast cancer miRNA, mRNA and protein expression datasets separately. A widely adapted technique in TCGA cancer papers is the NMF-Consensus clustering approach. We also run NMF-Consensus on the same datasets to see if we are able to capture better subgroupings of patients. We vary the number of clusters and analyze these clusters in terms of internal cluster validity metrics, such as silhouette width and external clinical data such as tumor stage, PAM50 classification and age of the patients. Experiments with mRNA, miRNA and RPPA data shows the separation quality of clusters in terms of survival rate, age, tumor stages and PAM50 subtypes. The relations of clusters to other clinical variables such as tumor stages are found to be statistically significant. When the data are clustered to 5 or 6 subgroups, the resulting survival rates of subgroups are shown to significantly differ from each other.

There are several routes to follow as part of future work:

- The similarity matrices that are found from each different data type can be combined in a single similarity matrix and clusters can be found using this combined similarity matrix.
- WS-RFClust can be applied to other cancer datasets.
- Current study investigates how the WS-RFClust classifier accuracy affects the predicted clusters. Other experiments could be designed to thoroughly investigate the relationship of the classifier accuracy to cluster validity.

Bibliography

[1] National Cancer Institute, "Defining cancer," jun 2014.

[2] American Cancer Society, 250 Williams Street, NW, Atlanta, GA, *Cancer Facts & Figures*, 2013.

[3] A. P. Trape and A. M. Gonzalez-Angulo, "Breast cancer and metastasis: On the way toward individualized therapy," *Cancer Genomics & Proteomics*, vol. 9, pp. 297–310, 2012.

[4] K. Madhu, "Personalized oncology: recent advances and future challenges," *Metabolism*, vol. 62, jan 2013.

[5] C. M. Perou et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–52, 2000.

[6] P. J. Stephens et al., "The landscape of cancer genes and mutational processes in breast cancer," *Nature*, vol. 486, no. 7403, pp. 400–4, 2012.

[7] D. N. Hayes and R. G. W. Verhaak, "An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1," *Cancer Cell*, jan 2010.

[8] T. Ideker and M. Hofree, "Network-based stratification of tumor mutations," *Nature*, pp. 1108–1115, 2013.

[9] N. K. Speicher and N. Pifeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, 2015.

[10] American Cancer Society, 250 Williams Street, NW, Atlanta, GA, *Cancer Facts & Figures*, 2013.

[11] S. Clancy and W. Brown, "Translation: Dna to mrna to protein," *Nature Education*, vol. 1, no. 1, p. 101, 2008.

[12] "microRNAs." `http://www.yale.edu/giraldezlab/miRNA.html`, 2016. [Online; accessed 27-March-2016].

[13] K. B. Reddy, "Microrna (mirna) in cancer," *Cancer Cell International*, vol. 15, no. 38, 2015.

[14] UC Santa Cruz, "UCSC Cancer Genomics Browser," June 2016.

[15] National Cancer Institute, "The cancer genome atlas," 2011.

[16] Z. Wang, M. Gerstein and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, oct 2010.

[17] B. P. Lewis, R. E. Green and S. E. Brenner, "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans," *PNAS*, vol. 100, pp. 189–192, 2003.

[18] K. Wang and J. Liu, "Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery," *Oxford Journals*, 2010.

[19] TCGA, mar 2016.

[20] MD Anderson Cancer Center, "The cancer proteome atlas," mar 2016.

[21] American Joint Commitee on Cancer, "What is cancer staging?," apr.

[22] National Cancer Institute, "What is the tnm system?," sep 2015.

[23] American Cancer Society, "Cancer facts for women," 2015.

[24] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.

[25] American Cancer Society, "How is breast cancer staged?," 2014.

[26] National Cancer Institute, "Tumor grade," 2013.

[27] G. Viale, "The current state of breast cancer classification," *Journal of Clinical Oncology*, vol. 21, no. 10, pp. 1973–1979, 2003.

[28] Nature Education, "Cell signaling," 2014.

[29] S. Sommer and S. A. W. Fuqua, "Estrogen receptor and breast cancer," *Seminars in Cancer Biology*, vol. 11, pp. 339–352, 2001.

[30] V. J. Bardou et al., "Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases," *Journal of Clinical Oncology*, vol. 21, no. 10, pp. 1973–1979, 2003.

[31] P. G. Natali et al., "Expression of the p185 encoded by her2 oncogene in normal and transformed human tissues," *International Journal of Cancer*, vol. 45, pp. 457–461, 1990.

[32] K. D. Voduc et al., "Breast cancer subtypes and the risk of local and regional relapse," *J Clin Oncol*, vol. 28, no. 10, pp. 1684–91, 2010.

[33] O. Metzger-Filho, Z. Sun and G. Viale et al., "Patterns of recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from international breast cancer study group trials viii and ix," *J Clin Oncol*, vol. 31, no. 25, pp. 3083–90, 2013.

[34] M. Yanagawa and K. Sasaki, "Luminal A and luminal B (HER2 negative) subtypes of breast cancer consist of a mixture of tumors with different genotype," *BMC Research Notes*, vol. 5, no. 376, 2012.

[35] Zorka Inic et al., "Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information", journal = "Clinical Medicine Insights: Oncology," vol. 8, no. 107-111, 2014.

[36] K.D. Voduc et al., "Breast cancer subtypes and the risk of local and regional relapse," *J Clin Oncol*, vol. 28, no. 10, pp. 1684–91, 2010.

[37] S. J. Schnitt, "Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy," *Modern Pathology*, vol. 23, pp. 60–64, 2010.

[38] T. Sørlie, C. M. Perou and R. Tibshirani et al, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[39] C. M. Perou et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–52, 2000.

[40] J. S. Parker et al., "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Clin Oncol*, vol. 27, no. 8, pp. 1160–1167, 2009.

[41] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, pp. 346–352, June 2012.

[42] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, ebook 10. 103, Springer, 2013. page 385-386.

[43] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, ebook 10. 103, Springer, 2013. page 395.

[44] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, ebook 10. 103, Springer, 2013. page 395.

[45] A. A. Alizadeh et al., "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, feb 2000.

[46] C. M. Perou et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, may 2000.

[47] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*, pp. 556–562, MIT Press, 2000.

[48] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM Journal on Scientific Computing*, 2011.

[49] S. Zhang et al., "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, 2012.

[50] The Cancer Genome Atlas Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, pp. 609–615, 2011.

[51] T. Ideker and M. Hofree, "Network-based stratification of tumor mutations," *Nature*, pp. 1108–1115, 2013.

[52] A. Bhattacharjee et al., "Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses," *PNAS*, vol. 98, no. 24, pp. 13790–13795, 2001.

[53] S. Monti et al., "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, pp. 91–118, 2003.

[54] S. Monti et al., "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, pp. 91–118, 2003.

[55] D. N. Hayes and R. G. W. Verhaak, "An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1," *Cancer Cell*, jan 2010.

[56] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.

[57] B. Winterhoff et al., "Molecular classification of high grade endometrioid and clear cell ovarian cancer using tcga gene expression signatures," *Mayo Clinic*, vol. 141, no. 6, pp. 95–100, 2016.

[58] A. B. S. Basu and R. Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*, 2002.

[59] S. Basu, I. Davidson and K. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 1 ed., 2008.

[60] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLOS Biology*, vol. 2, April 2004.

[61] D. N. Hayes and R. G. W. Verhaak, "An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1," *Cancer Cell*, January 2010.

[62] T. Ideker and M. Hofree, "Network-based stratification of tumor mutations," *Nature*, pp. 1108–1115, 2013.

[63] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, 2015.

[64] The Cancer Genome Atlas Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, pp. 609–615, 2011.

[65] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLOS Biology*, vol. 2, April 2004.

[66] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLOS Biology*, vol. 2, April 2004.

[67] L. Bullinger et al., "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia," *the New England Journal of Medicine*, vol. 350, pp. 1605–1616, April 2004.

[68] D. C. Koestler et. al., "Semi-supervised recursively partitioned mixture models for identifying cancer subtypes," *Bioinformatics*, vol. 26, no. 20, pp. 2578–85, 2010.

[69] E. A. Houseman et al., "Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions," *BMC Bioinformatics*, vol. 9, p. 365, 2008.

[70] S. Gaynor and E. Bair, "Identification of biologically relevant subtypes via preweighted sparse clustering," *Biostatistics*, pp. 1–33, 2013.

[71] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 713–726, 2010.

[72] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[73] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*. Springer, second ed., 2013.

[74] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[75] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[76] L. Parsons, "Abstract subspace clustering for high dimensional data: A review ."

[77] A. Zhang, *Advanced Analysis of Gene Expression Microarray Data*. World Scientific Publishing Co., 2006.

[78] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier, fourth ed., 2009.

[79] E. L. Kaplan, and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[80] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports*, vol. 50, no. 3, pp. 163–70, 1966.

[81] Peter J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[82] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits,"

[83] J. P. Brunet et al., "Metagenes and molecular pattern discovery using matrix factorization," *Proc Natl Acad Sci USA*, vol. 101, no. 12, pp. 4164–9, 2004.