Pınar GÜNER

A Master's Thesis

AGU 2021

DEVELOPING A LABEL PROPAGATION APPROACH FOR CANCER SUBTYPE IDENTIFICATION PROBLEM

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Pınar GÜNER
July 2021

DEVELOPING A LABEL PROPAGATION APPROACH FOR CANCER SUBTYPE IDENTIFICATION PROBLEM

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Pınar GÜNER
July 2021

SCIENTIFIC ETHICS COMPLIANCE

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

Name-Surname: Pınar GÜNER
Signature :

REGULATORY COMPLIANCE

M.Sc. thesis titled Developing a Label Propagation Approach for Cancer Subtype Identification Problem has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By
Pınar GÜNER

Advisor
Assist. Prof. Dr. Burcu BAKIR-GÜNGÖR

Co-Advisor
Assist. Prof. Dr. Mustafa COŞKUN

Head of the Electrical and Computer Engineering Program
Assoc. Prof. Dr. Kutay İÇÖZ

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled Developing a Label Propagation Approach for Cancer Subtype Identification Problem and prepared by Pınar GÜNER has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

13 /07 / 2021

JURY:

Advisor : (Assist. Prof. Dr. Burcu BAKIR-GÜNGÖR)

Co-Advisor : (Assist. Prof. Dr. Mustafa COŞKUN)

Member : (Assist. Prof. Dr. Ahmet SORAN)

Member : (Prof. Dr. Bahriye AKAY)

Member : (Assist. Prof. Dr. Özkan Ufuk NALBANTOĞLU)

APPROVAL:

The acceptance of this M.Sc. thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / …….. and numbered ……..……. .

…….. /…….. / ……..

(Date)

Graduate School Dean
Prof. Dr. Hakan USTA

ABSTRACT

DEVELOPING A LABEL PROPAGATION APPROACH FOR CANCER SUBTYPE IDENTIFICATION PROBLEM

Pınar GÜNER
M.Sc. in Electrical and Computer Engineering
Advisor: Assist. Prof. Dr. Burcu BAKIR-GÜNGÖR
Co-Advisor: Assist. Prof. Dr. Mustafa COŞKUN
July 2021

The term of cancer is used to describe diseases in which abnormal cells that grow out of control and invade other tissues. There are multiple types of cancer and many types of cancer have various subtypes with different clinical and biological implications. These differences show that diverse methods should be followed for the treatment of different subtypes of cancer. Discovering cancer subtypes is an important problem in bioinformatics, as it can help improve personalized medicine. Knowing the subtype of cancer is useful for determine the treatment steps and prognosis. Computational bioinformatics methods help performing cancer analysis to design targeted treatments by exposing the common molecular pathology of different cancer subtypes. Thus far, several computational methods have been proposed to discover cancer subtypes or to stratify cancer into informative subtypes. However, existing works do not consider the sparseness of data, and result in ill-conditioned solution. To resort this shortcoming, in this thesis, we propose an alternative unsupervised computational method to stratify cancer into subtypes using applied numerical algebra techniques. More specifically, we applied this label propagation-based approach to stratify somatic mutation profiles of colon, head and neck, uterine, bladder and breast tumors. We then evaluated the performance of our method by comparing it to the baseline methods. Extensive experiments demonstrate that our approach highly renders tumor classification tasks by largely outperforming the state-of-the-art unsupervised and supervised approaches.

Keywords: Machine Learning, Label Propagation, Cancer Subtype, Personalized Medicine

ÖZET

KANSER ALT TİPİ TANIMLAMA PROBLEMİ İÇİN BİR ETİKET YAYMA YAKLAŞIMI GELİŞTİRME

Pınar GÜNER
Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Yöneticisi: Dr. Öğr. Üyesi Burcu BAKIR-GÜNGÖR
Eş Danışman: Dr. Öğr. Üyesi Mustafa COŞKUN
Temmuz 2021

Kanser terimi, anormal hücrelerin kontrolden çıkıp diğer dokuları istila ettiği hastalıkları tanımlamak için kullanılır. Çok sayıda kanser türü vardır ve birçok kanser türü, farklı klinik ve biyolojik etkileri olan çeşitli alt tiplere sahiptir. Bu farklılıklar, kanserin farklı alt tiplerinin tedavisi için farklı yöntemlerin izlenmesi gerektiğini göstermektedir. Kişiselleştirilmiş tıbbın geliştirilmesine yardımcı olabileceğinden, kanser alt tiplerini keşfetmek biyoinformatikte önemli bir problemdir. Kanserin alt tipinin bilinmesi, tedavi basamaklarının ve öngörünün belirlenmesinde faydalıdır. Hesaplamalı biyoinformatik yöntemler, farklı kanser alt tiplerinin ortak moleküler patolojisini ortaya çıkararak hedeflenen tedavileri tasarlamak için kanser analizi yapmaya yardımcı olur. Şimdiye kadar, kanser alt tiplerini keşfetmek veya kanseri bilgilendirici alt tiplere ayırmak için çeşitli hesaplamalı yöntemler önerildi. Ancak, mevcut çalışmalar verilerin seyrekliğini dikkate almamakta ve kötü koşullu (tersi alınamayan) çözümle sonuçlanmaktadır. Bu eksikliği gidermek için, bu tezde, uygulamalı sayısal cebir tekniklerini kullanarak kanseri alt tiplerine ayırmak için alternatif bir denetimsiz hesaplama yöntemi öneriyoruz. Daha detaylı olarak, bu etiket yayma tabanlı yaklaşımı kolon, baş ve boyun, rahim, mesane ve meme tümörlerinin somatik mutasyon profillerini sınıflandırmak için uyguladık. Sonra, yöntemimizin performansını temel yöntemlerle karşılaştırarak değerlendirdik. Kapsamlı deneyler, yaklaşımımızın, modern denetimsiz ve denetimli yaklaşımlardan büyük ölçüde daha iyi performans göstererek tümör sınıflandırma görevlerini yüksek oranda yerine getirdiğini kanıtlamaktadır.

Anahtar kelimeler: Makine Öğrenmesi, Etiket Yayma, Kanser Alt Tipi, Kişiselleştirilmiş Tıp

Acknowledgements

I would like to express my sincerest appreciation to my advisor Assist. Prof. Dr. Burcu BAKIR-GÜNGÖR for her support and guidance throughout my thesis study. I would also like to extend my deepest appreciation to my co-advisor Assist. Prof. Dr. Mustafa COŞKUN for sharing his invaluable knowledge, for his patience and full support.

I also would like to give my warmest thanks to my family. I would like to thank my mother Emine and my father Erkan for their endless patience, tolerance and love. I am grateful to my sister Mediha for listening to all my complaints and for her guidence. They have been always there for me, especially during the tough times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BLCA	Urothelial Bladder Carcinoma
BRCA	Breast Cancer
COAD	Colon Adenocarcinoma
GNMF	Graph Regularized Non-negative Matrix Factorization
HNSC	Head-Neck Squamous Cell Carcinoma
NMF	Non-negative Matrix Factorization
RWR	Random Walk with Restart
TCGA	The Cancer Genome Atlas
t-SNE	t-distributed Stochastic Neighbor Embedding
UCEC	Uterine Corpus Endometrial Carcinoma

Chapter 1

Introduction

Cancer is a complex and deadly disease, according to the report published by the World Health Organization (WHO) in 2018, 18.1 million new cancer cases were predicted and 9.6 million people around the world died from the disease. There are many factors that causes cancer to be deadly, such as tumor type, stage of cancer, clinical factors and among many others. Thus, to improve the survival rates, cancer patients should be given the best possible treatment plan based on the aforementioned factors [1].

Currently, people who are diagnosed with cancer generally receive the same treatment as others who have the same type of cancer. However, different patients may respond to the same treatment differently since the types of tumors of patients have the different genetic changes that cause cancer to grow and spread differently. The genetic changes in one person's cancer may not occur in other with the same type of cancer. Furthermore, changes that cause the same cancer can also be found in the other cancer types. The area of personalized medicine has been established to deal with personalized treatment with one of the objectives, personalized cancer treatment [2]. The first step to apply personalized medicines is to stratify (classify or cluster) cancer patients into meaningful subtypes based on the tumor molecular profiles and specific mutations. To facilitate the personalized medicine, computational methods have been soaring many research attentions to remedy the cancer subtype identification problem since *in vitro* experiments and clinical trials are costly and time consuming [3-7].

In this thesis, we aim at developing such an effective computational method for cancer subtype stratification. More specifically, we propose an alternative unsupervised computational method based on the idea of sparsity the given cancer data and by capitalizing applied numerical algebra techniques to cluster tumors into meaningful subtypes with a better clustering accuracy comparing the existing supervised and unsupervised approaches [3, 7]. To evaluate the proposed method, we used various cancer

datasets, such as colon, head and neck, uterine, bladder and breast tumors extracted from the TCGA (The Cancer Genome Atlas) project [8]. Extensive experiments on these cancer datasets from TCGA show that our unsupervised stratification method significantly outperforms the state-of-the-art unsupervised computational methods [3] for identifying cancer subtypes; even its performance exceeds the supervised approaches [7].

1.1 Problem Definition: Cancer Subtype Identification Problem

Subtyping (stratifying) cancer or identification of cancer subtypes is one of the essential steps in personalized medicine for determining the accurate diagnosis and most effective follow-up therapy to increase survival of cancer patients. To date, various subtypes of cancer tumors have been comprehensively investigated [9-12] and it has been shown that different subtypes of cancer are often caused by different genetic mutations [13]. There are two main approaches for cancer subtype discovery: (i) *in vitro* experiments based on examination of a biopsy and (ii) computational methods based on advanced machine learning techniques. The first and traditional methods are time consuming as well as the results of the analysis are subject to human error [14]. Thus, the second approach, computational methods, have been gaining many research attentions to mitigate the associated costs of the experimental approaches to identify cancer subtypes. To do so, the computational approaches rely on basic premise: cluster patients into different subgroups based on their genetic profiles and clinical symptoms [15] and they are validated on The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) which generate genetic profiles thousands of patients from several tumor types [16,17].

1.2 Thesis Organization

This thesis organized as follows. In Chapter 2, background information for cancer and cancer subtype is given, including the omics datasets used in cancer research. Also, computational background of the cancer subtype identification and performance evaluation methods is given. Chapter 3 reviews methods in the literature about identifying tumor subtypes. Chapter 4 defines the input data and presents the algorithmic background behind our proposed method. Chapter 5 covers the results of using the proposed method on different data, compared to state-of-the-art methods. Chapter 6 discusses the results of the proposed method. And finally, Chapter 7 summarizes the thesis, gives main contributions of the thesis, and explains where the results lead us and what might be the future studies.

Chapter 2

Background

2.1 Biological Background

2.1.1 Cancer and Cancer Subtypes

Cancer is a genetic and complex disease for which body cells divide indefinitely and spread to neighboring tissues. It can develop in different tissues and cells. Normally, a healthy person's cells grow and die when the cells become old or damaged so that new cells can be substituted with the dead ones. However, a cancer patient cells do not die when they grow older; rather they continue to spread other tissues. These cells grow with an uncontrolled way, termed as tumor [18]. Here, in a specific cancer type, cancer subtype is defined as the smaller group in the cancer type that is formed based on molecular profiles and specific mutations [19]. Many cancer types have various subtypes each of which with its own clinical implications, such as patient survival time and the response to drug resistance. Thus, knowing these subtypes of cancers is important to determine the treatment steps, prognosis, and response to treatment [3]. As a result, separating patients into different groups based on their cancer subtypes can guide the selection of drugs that minimize side effects and provide more effective outcomes.

2.1.2 Omics Data for Cancer

With the advent of the recent advanced technologies in genomics have contributed to the collection of high-throughput molecular datasets by large-scale projects, such as The Cancer Genome Atlas (TCGA). These types of large amount of datasets curated in public repositories is a very important source of information for cancer researchers [20]. Furthermore, as one of the important issues for cancer researchers is cancer subtype

identification, along this line, the subtype datasets have been generated, such as mRNA and microRNA expression levels, methylation data, copy number alterations, somatic mutation data [21]. In the following subsections, these datasets are given in details.

2.1.2.1 Gene Expression Data

Gene expression is a term used for the process of conversion of genetic information from genes to functional protein structures via messenger RNA (mRNA), which links genes to proteins [22]. These mRNAs play crucial roles in producing proteins which are vital for human body. Additionally, Micro RNAs (miRNA) are non-coding RNA molecules that have a function in gene expression regulation. Thus, miRNAs have been at the heart of cancer research, as they are identified as potential biomarkers for human cancer diagnosis [23]. Thanks to the advancement of DNA microarray technology, it is possible that expression levels of thousands of genes can be measured simultaneously under specified experimental environments and conditions. This technology also enables the production of large-scale gene expression data ready to be analyzed. The use of the high-throughput technologies for gene expression analyses have provided new classifications of cancer patients. Cancer subtypes based on gene expression have been comprehensively investigated, as they are associated with different cellular, molecular and clinical properties [11, 24, 25].

2.1.2.2 DNA Methylation Data

Another important data used in the cancer research is DNA methylation, which is defined as an epigenetic mechanism involving a methyl group is added to the DNA molecule [26] as DNA methylation patterns are altered in many diseases, such as cancer. Several research projects on DNA methylation have specifically focused on the cancer and tumor suppressor genes [27-29]. Since DNA hyper-methylation can accurately characterize type of a specific tumor, the usage of DNA methylation markers has been demonstrated to be promising for identification of tumor subtypes [29].

2.1.2.3 Copy Number Alterations

Copy number alterations (CNAs) are the repetition or deletion of large parts of the genetic code. CNAs are common in many cancer types and by examining these alterations, one can infer which cancers will be more deadly [30]. The ability to examine copy number alterations in tumors may also guide doctors to tailor treatment for a specific tumor [31].

2.1.2.4 Somatic Mutation Data

Alterations (mutations) in the DNA sequence of the genomes of cancer cells cause cancer. Broadly speaking, these mutations are classified into two types, germline and somatic. Germline mutations occur in sperm and egg cells, so they are inherited from parents and transmitted to children [13].

Somatic mutations, on the other hand, can arise in any part of the body except the germ cell, thus, they are not transmitted to descendants [32]. These mutations are the most common cause of cancer, so they have an important part in cancer development and disease progression. Therefore, subtype classification based on somatic mutation profiles could be informative to identify subgroups of patients who might respond to different treatments [21]. In recent years, as a result of the development of high-throughput platforms, somatic mutation profiles have become a new and promising data source for tumor classification. However, the somatic mutation profiles of tumors are heterogeneous, meaning that there are many differences between and within different cancers. To alleviate the heterogeneity, recent studies have been incorporating protein-protein interaction (networks) information as an additional information source to somatic mutation data [3,7].

2.2 Computational Background

2.2.1 Machine Learning for Cancer Subtype Identification Problem

Recent advances in Machine learning have been used for cancer subtype identification problem. In the context of biology, various machine learning techniques have been used for many purposes, such as diagnosis, prognosis, screening, treatment in cancer [33]. We can broadly generalize, machine learning techniques used in cancer research into 4 sub-categories: supervised learning, semi-supervised learning, unsupervised learning and reinforcement learning [34].

Supervised learning is a machine learning method in which a supervised model is trained on a labeled dataset. In supervised learning there are both input and output data. In general setting, the supervised learning methods can be seen as a function which maps the dataset onto label set [35]. To do so, the dataset is divided into training and test data and the function is trained on the training data and evaluated on the test data, these processes are named as classification or regression depending on the label.

Another machine learning technique is called semi-supervised learning (SSL) where label information is limited [36]. The objective of SSL is using some propagation rules to expand the label set information by using small labeled examples with the idea of "guilt-by-association".

Unsupervised learning refers to machine learning methods that try to explain the relationship in the data without knowledge of label. To build an unsupervised model, we create an objective function that tries to measure the latent "closeness" of data. In unsupervised learning, the goal is to discover hidden patterns in the input data and to detect which samples belong to which class [35].

Reinforcement learning is another widely adopted machine learning approach that can be seen as supervised method without label and changing dataset. In a trail-error fashion, labels are generated based on a predefined reward function that aims at maximizing (or minimizing) a pre-defined goal [35].

Among these computational machine learning approaches, supervised classification and unsupervised clustering are well-adopted for molecule-based cancer subtype discovery, which is an important topic in personalized medicine. More specifically, to mitigate the aforementioned heterogeneity, see Section 2.1.2.4, network-based supervised

and unsupervised methods are popular methods used for cancer subtype identification problem as following [3,7].

2.2.1.1 Unsupervised Tumor Stratification

Tumor stratification problem can be loosely defined as clustering patients (tumors), via molecular data, into classes or subcategories. As obtaining labeled data is mostly difficult, costly, and time consuming, unlabeled molecular data, on the other hand, is easy to access; thus, unsupervised clustering machine learning methods have been employed for the tumor stratification problem [37]. While solely using molecular data has been shown to be effective for the tumor stratification, heterogeneity of this data limits ability of clustering methods [37]. Thus, as additional protein-protein interaction network information has been invoked to smooth and resolve heterogeneity problem of the data.

More specifically, somatic mutation profiles are combined with molecular network information in network-based tumor stratification approaches. Using the network propagation techniques, such as random walk with restarts (RWR), the influence of each somatic mutation profile is propagated across its network neighborhood and clustering approaches are applied over the smoothed somatic mutation profiles. We can summarize the clustering procedure as follows:

1- A certain part of the rows (patients) and columns (mutated genes) of the binary somatic mutation data are subsampled at random without replacement.
2- Binary somatic mutation data is propagated over the network.
3- Quantile normalization technique is applied to the network smoothed mutation data.
4- Graph (or network) regularized non-negative matrix factorization (GNMF) is used to decompose network data into k clusters.

Finally, consensus clustering is applied over network smoothed mutation profiles.

- **Network Smoothing:** Network propagation is at the core of a large number of network analyses tasks, such as protein function prediction, gene prioritizing and disease module discovery. In the context of computational biology, network propagation algorithms are based on the assumption that information on known disease genes flows over the network via nearby proteins [38]. In terms of tumor stratification, network propagation has been used for serving the same purpose: to capture the similarity among nodes in the molecular network [39]. The basic idea

behind network propagation is to employ a random walk [40] model to diffuse information about tumor mutations using molecular interaction networks associations. Mathematically, we can define this network propagation as follows:

$$\mathbf{F_{t+1}} = \boldsymbol{\alpha}\mathbf{F_t}\mathbf{A} + (\mathbf{1} - \boldsymbol{\alpha})\mathbf{F_0} \quad \textbf{(2.1)}$$

$\mathbf{F_0}$ (patient-by-gene binary matrix) is each tumor's mutation profile, **A** is a degree normalized adjacency matrix of the molecular interaction network. The parameter $\boldsymbol{\alpha}$ controls the random walker that determines how much a mutation signal should diffuse on the network and set to be in the range between 0 and 1. Iterative computation is performed via $\mathbf{t}$ values (0, 1, 2, …) until $\mathbf{F_{t+1}}$converges. At convergence ($\mathbf{F_{t+1}} \approx \mathbf{F_t}$), $\mathbf{F_t}$ (propagated mutation profiles) denotes a patient - by- gene binary matrix in which each tumor's mutation profile has been smoothed across the network. After obtaining propagated mutation profiles, the non-negative factorization is applied to this matrix to identify clinically and biologically meaningful subtypes [3, 7].

- **Non-negative matrix factorization (NMF):** In machine learning problems, the high dimensional problem in the input data matrix adversely affects the learning task. Therefore, to resolve the high dimensionality problem matrix factorization techniques, dimensionality reduction, have become frequently used for data representation [41]. As such, one of the dimensionality reduction technique, non-negative matrix factorization (NMF) has been used for many learning tasks for which the constraint, lower dimensional matrices have to be positive, i.e., the input matrix **V** is factorized into a feature set **W** and hidden variables $\boldsymbol{H}$. In the matrix factors **W** and **H** do not contain negative elements [42].

$$\mathbf{V} \approx \mathbf{WH} \quad \textbf{(2.2)}$$

Non-negative matrix factorization (NMF) has proven to be one of the best methods for learning the components of objects such as text documents [43]. However, vanilla NMF ignores the geometric structure of the data space, which might be useful for data classification and clustering tasks. Thus, in the context of unsupervised tumor stratification problem, graph (or network) regularized version

of non-negative matrix factorization (GNMF) is used to minimize the objective function given in the Equation 2.3 [3, 41, 43, 44].

$$\|\mathbf{F} - \mathbf{WH}\|^2 + \boldsymbol{\lambda}\, \mathbf{Tr}(\mathbf{HLH}^{\mathbf{T}}) \qquad \textbf{(2.3)}$$

In GNMF, vanilla NMF is extended by graph topological information as in the last equation. If we do not have readily available graph, we can employ K-nearest-neighbor (KNN) to create a network from influence matrix [45] of the reference molecular network. Then, the graph Laplacian of this KNN network is used as the regularizer in the NMF steps.

In the Equation 2.3, **W** and **H** are a decomposition of **F** (patient-by-gene matrix) formed as a result of network smoothing. **W** (genes-by-k) and **H** (k-by-patients) are the basis and patient cluster matrices. $\mathbf{Tr}()$ represents the trace of a matrix and $\boldsymbol{L} = \mathbf{D} - \mathbf{A}$ is the graph Laplacian of the K-nearest-neighbor network, where **D** denotes diagonal degree matrix of the KNN network and **A** is the adjacency matrix of the KNN network. $\boldsymbol{\lambda}$ is the regularization constant to scale network regularizer (**L**) term in GNMF.

Multiple instances of **H** matrix will be combined together during the consensus clustering step of the algorithm.

- **Consensus clustering:** This methodology serves to represent consensus among multiple clustering algorithm runs. Also, it is used to detect the number of clusters in the data and evaluates the stability of the identified clusters. Resampling techniques can be used to simulate data perturbations. The clustering algorithm can then be implemented to each perturbed dataset and the consensus among the multiple runs can be evaluated. To represent the agreement among the clustering a consensus matrix (N × N) is defined (The number of elements in a dataset is denoted by N). A consensus matrix is calculated for each cluster and each element in the matrix denotes the proportion of clustering runs in which two samples clustered together [46]. In unsupervised tumor stratification problem, GNMF is performed multiple times on subsamples of the dataset. Then, the set of clustering outcomes is transformed into a co-clustering matrix. Each element in the co-clustering matrix represents the frequency with each two tumors was discovered

that belong to the same cluster. In order to obtain co-clustering matrix, the following procedure is implemented [47]:

1- For each **H** matrix, generated after multiple iterations, the column-wise argmax of each row is computed and the patient is assigned to that cluster number.
2- A matrix is created to count how many times each patient pair appears in the same **H** matrix.
3- A matrix is generated to count how many times each patient pair has been assigned to the same cluster.
4- The matrix obtained in step 3 is divided by the matrix obtained in step 2 (element-wise division) for normalization.

Then, a patient linkage map is created from this co-clustering matrix and patient clusters are assigned from the patient link map hierarchy.

2.2.1.2 Supervised Tumor Classification

As opposed to the unsupervised methods, supervised methods classify tumors into predefined subtypes using labeled datasets. Inspired by supervised random walk approach presented in [48] cancer classification is used to identify potential biomarkers as well as predict patient survivability and cancer prognosis [49]. In this section, we first explain the basic ideas behind the supervised learning-based tumor classification [7].

Given a graph $\mathcal{G}$ with nodes and edges, the nodes denote genes, and the edges denote molecular interactions between genes. Random Walk with Restarts (RWR) procedure is given in the Equation 2.4 is conducted iteratively as follows:

$$\mathbf{P^{(t+1)}} = (\mathbf{1} - \boldsymbol{\alpha})\mathbf{P^{(t)}Q} + \boldsymbol{\alpha}\mathbf{P^{(0)}} \qquad \textbf{(2.4)}$$

In this equation, $\mathbf{P^{(0)}}$ represents a tumor-by-gene matrix and $\mathbf{Q}$ denotes degree normalized adjacency matrix of $\mathcal{G}$. With the RWR, an activation score is computed for each edge and a weighted transition matrix is calculated.

The center of each subtype cluster is learned from training data during the network based supervised classification training procedure. Each tumor sample is assigned to one of these subtypes during the validation stage. This operation is performed according to

the shortest Euclidean distance to the centers. Based on this explanation, the value of the cost function is calculated.

Total number of tumors is represented by $\mathbf{m}$ and row $\mathbf{u}$ of $\mathbf{P}$ is represented by $\mathbf{p_u}$. The centroid vector of tumor $\mathbf{u}$'s true subtype $\mathbf{a}$ is $\mathbf{c_a}$, described as follows, is and $\mathbf{m_a}$ represents the number of tumors in subtype $\mathbf{a}$.

$$\mathbf{c_a} = \frac{\mathbf{1}}{\mathbf{m_a - 1}} \sum_{\mathbf{v \epsilon a, v \neq a}} \mathbf{p_u} \tag{2.5}$$

Training the model is performed iteratively through gradient descent. To obtain local optimum of the optimization problem, the gradient of each parameter with respect to the edge feature weights $\mathbf{w}$ is calculated using the chain rule and $\mathbf{w}$ is updated correspondingly. When convergence is achieved, the final feature weights, transition matrix, and propagated mutation profiles, are produced.

Finally, to predict the subtype of a new tumor $\mathbf{z}$ with mutation profile $\mathbf{P_z^{(0)}}$ following equation is used ($\mathbf{s}$ is the predicted subtype of $\mathbf{z}$ and $\mathbf{A}$ is the set of all subtypes):

$$\underset{\mathbf{s \in A}}{\mathbf{argmin}} \|\mathbf{p_z} - \mathbf{c_s}\|_2^2 \tag{2.6}$$

2.2.2 Performance Evaluation Metrics

In supervised learning, the performance of the model can be tested in a reserved evaluation set, as there are labels for each sample. Therefore, there are numerous well-adopted evaluation metrics for supervised learning. In unsupervised learning, cluster evaluation is not well developed because of there are unlabeled data. Even so, there are many metrics that analyze the quality of clustering results of model without labeled data. In this thesis, following evaluation metrics are used [46, 50-53]. We use two evaluation schemes to visualize the results: t-Distributed Stochastic Neighbor Embedding (t-SNE) and co-clustering heat map. More evaluation and computational methods for gene clustering from the perspective of cluster quantification can be found [54]. Clustering validity indexes mentioned in this paper such as Silhouette coefficient, Davies-Bouldin

index and average of sum of intra-cluster distances were used to evaluate the quality of the proposed approach in this thesis.

2.2.2.1 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique, visualizes high dimensional data in a low-dimensional space of two or three dimensions. It models objects in a dataset, by keeping the low-dimensional representations of dissimilar data points far apart and bringing similar data points close together. t-SNE is also able to recover well-separated clusters [50].

Figure 2.1 shows the experiment results of Mateen and Hinton with t-SNE [50]. They compared the results with the existing visualization techniques, and they found that t-SNE produced a map in which the distinction between digit classes was almost perfect. Each class is represented in a different color on the map. The use of coloring helps to assess how well the map preserves the similarities within each class. As shown in the figure, most classes are grouped into a single heap. Hence, t-SNE is useful for finding a representation that can distinguish between classes.

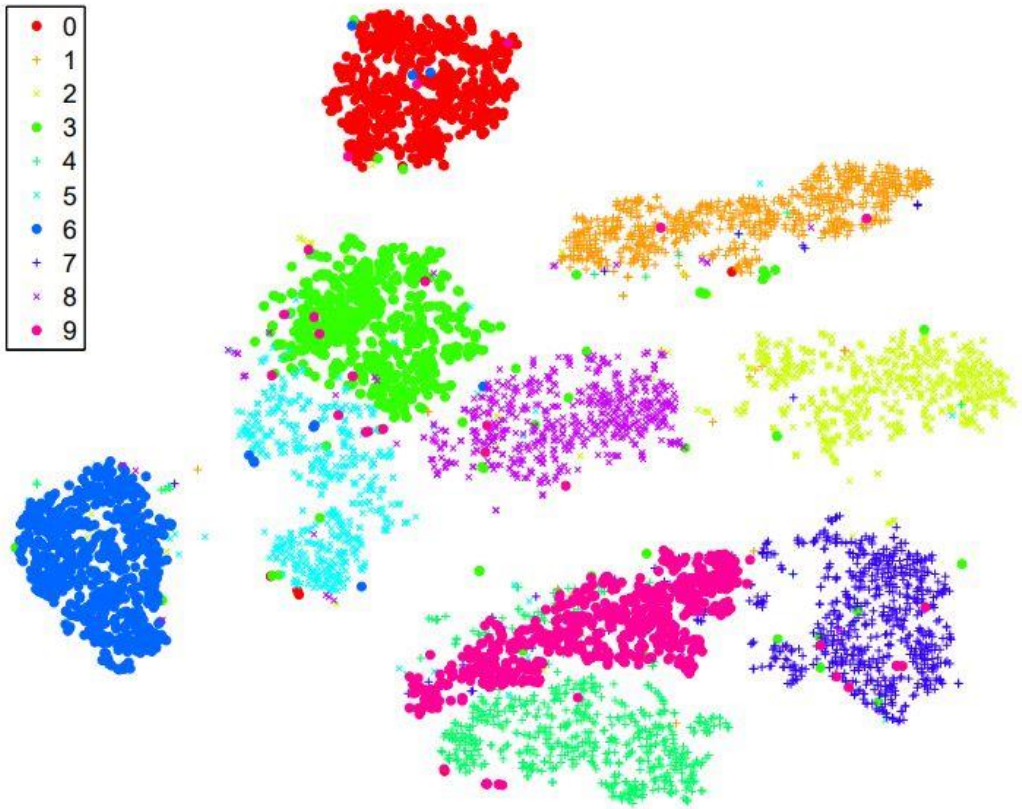

Figure 2.1 Visualization by t-SNE [50]

The t-SNE algorithm calculates a similarity measure between pairs of samples and uses a cost function to optimize these two similarity measures. In this study, this method was used to observe the distribution of stratified tumor subtypes.

2.2.2.2 Co- Clustering Heat Map

The purpose of using consensus clustering is to assess the consistency of the detected clusters. The consensus matrix mentioned in the section 2.2.1.1 can be visualized and used to evaluate composition and number of the clusters. Figure 2.2 is an example of consensus clustering heat map [46]. It shows two heat maps obtained by performing consensus clustering to two different datasets. Here, it is assumed that the 0 to 1 range of real numbers are represented by a color scale, so that 0 corresponds white color and 1 corresponds red color. A darker red color indicates higher co-clustering. A color-coded heat map with red blocks along the diagonal on a white background symbolizes a perfect consensus since items in the same cluster are located close to each other. In the example of Figure 2.2, the heat map for the Gaussian3 consensus matrix shows a well-defined 3-cluster structure, while the heat map for the Uniform1 consensus matrix displays no such structure.

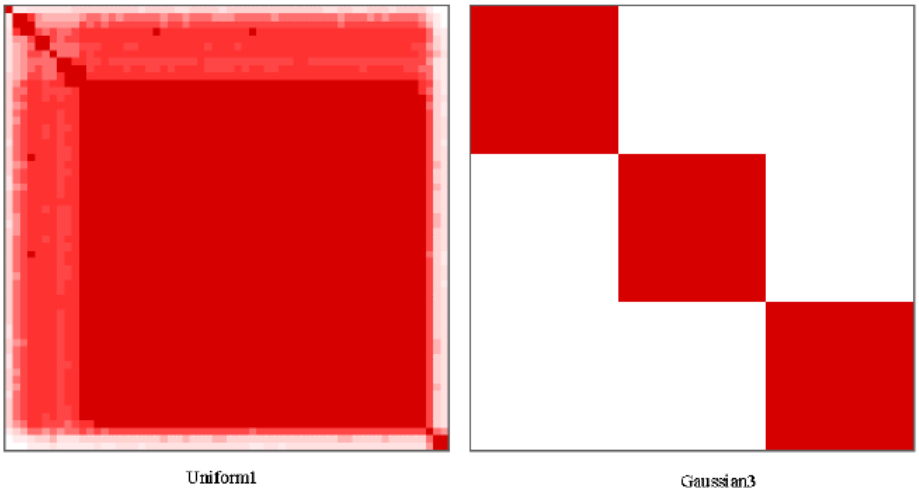

Figure 2.2 Co-clustering heat map example [46]

2.2.2.3 Silhouette Coefficient

The silhouette coefficient is a metric that indicates each data point's proximity to its own cluster and the separation between different clusters. It is a famous method of evaluating the clustering quality [51]. The silhouette value of data point **i** (**s**(**i**)) is calculated as Equation 2.7. Here, **a**(**i**) is the average distance of object **i** to all other objects in its cluster and **b**(**i**) is the minimum of all average distances of object **i** to all objects in any other cluster.

$$\mathbf{s(i)} = \frac{\mathbf{b(i)} - \mathbf{a(i)}}{\mathbf{max}\,(\mathbf{a(i)}, \mathbf{b(i)})} \tag{2.7}$$

The value of silhouette coefficient ranges from -1 to 1. Values closer to -1 means that clusters are not assigned correctly. Values closer to 1 indicates that clusters are well apart from each other, and the better clustering result. Silhouette coefficient is calculated for all data points and an average value is obtained as an overall measure ($\mathbf{s_k}$). In the Equation 2.8, the number of clusters is **k**, and the Silhouette coefficient is calculated for all data points and an average value is obtained as an overall measure number of data points is **n**.

$$\mathbf{s_k} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i \neq 1}}^{\mathbf{n}} \mathbf{s(i)} \tag{2.8}$$

2.2.2.4 Davies-Bouldin Index (DB Index)

The Davies-Bouldin Index is used to estimate the clustering results and it measures average similarity between each cluster and the cluster to which it is most similar. [52]. The formula for DB index is given in the Equation 2.9.

$$\mathbf{DB} = \frac{\mathbf{1}}{\mathbf{k}} \sum_{\mathbf{i,j=1}}^{\mathbf{k}} \underset{\mathbf{i \neq j}}{\mathbf{max}} \left\{ \frac{\hat{\mathbf{d}}_\mathbf{i} + \hat{\mathbf{d}}_\mathbf{j}}{\hat{\mathbf{d}}_\mathbf{i,j}} \right\} \tag{2.9}$$

Here, **k** is the number of clusters and $\hat{\mathbf{d}}_\mathbf{i}$ is the average distance from each data point in cluster **i** to the centroid of cluster **i**; $\hat{\mathbf{d}}_\mathbf{j}$ is the average distance from each data point in cluster **j** to the centroid of cluster **j**; and $\hat{\mathbf{d}}_\mathbf{i,j}$ is the Euclidean distance between the

centroids of cluster **i** and **j**. The minimum value of DB Index is zero and values closer to zero indicate a better partition [54].

2.2.2.5 Intra-cluster Distances

Intra-cluster distance is the distance between data points belonging to same cluster. Intra-cluster distance should be minimum to get the best clustering result. Three methods are used to calculate intra-cluster distance [53].

The *complete diameter distance* calculates the distance between two most remote data points belonging to the same cluster. It is defined as the Equation 2.10. Here, **S** is the cluster formed using partition, $\mathbf{d}(\mathbf{x}, \mathbf{y})$ is the distance between two data points, **x** and **y**, belonging to cluster **S**.

$$\Delta_1(\mathbf{S}) = \max_{\mathbf{x},\mathbf{y} \in \mathbf{S}} \{\mathbf{d}(\mathbf{x}, \mathbf{y})\} \tag{2.10}$$

The *average diameter distance* is the average distance between all the data points belonging to the same cluster. It is defined as the Equation 2.11. Here, $|\mathbf{S}|$ is the number of data points in cluster **S**.

$$\Delta_2(\mathbf{S}) = \frac{\mathbf{1}}{|\mathbf{S}|.(|\mathbf{S}| - \mathbf{1})} \sum_{\substack{\mathbf{x},\mathbf{y} \in \mathbf{S} \\ \mathbf{x} \neq \mathbf{y}}} \{\mathbf{d}(\mathbf{x}, \mathbf{y})\} \tag{2.11}$$

The *centroid diameter distance* represents the double average distance between all the data points and the center of cluster. It is defined as the Equation 2.12.

$$\Delta_3(\mathbf{S}) = \mathbf{2}\left(\frac{\sum_{\mathbf{x} \in \mathbf{S}} \mathbf{d}(\mathbf{x}, \bar{\mathbf{v}})}{|\mathbf{S}|}\right) \tag{2.12}$$

Here, $\bar{\mathbf{v}}$ is calculated as:

$$\bar{\mathbf{v}} = \frac{\mathbf{1}}{|\mathbf{S}|} \sum_{\mathbf{x} \in \mathbf{S}} \mathbf{x} \tag{2.13}$$

Chapter 3

Literature Review

Large-scale cancer omics studies aim to understand the molecular mechanisms of cancer and have demonstrated that cancer subtypes have a strong association with clinical outcomes. Several studies have been proposed to identify tumors into subtypes. Cancer subtyping studies in the literature can be divided into basically two: unsupervised clustering and supervised classification. Prominent approaches for cancer subtype identification use different data types. Some of them stratify tumors with molecular profiles using mRNA expression data [11, 12, 56, 57]. As a result of these studies subtypes of breast and glioblastoma cancers are identified. The other data used for tumor stratification is somatic mutation profiles that is essential in the development of cancer research. Similarities and differences in patient tumor mutation profiles provide information for tumor subtype stratification. However, some challenges are created by the fact that somatic mutation profiles are sparse and heterogeneous. To overcome these challenges, some studies used the network-based approach to discover cancer subtypes [3-6, 58, 59]. The following are some studies in the literature related to cancer subtype identification.

Lee et al. introduced a disease classification technique based on pathway activities inferred for each patient. This gene expression-based classification technique shows that pathway markers can increase the classification accuracy [60].

The Cancer Genome Atlas (TCGA) used mRNA and miRNA expression and DNA methylation of ovarian adenocarcinomas for subtype identification. In a TCGA project, four ovarian cancer transcriptional subtypes, three microRNA subtypes, four methylation subtypes were obtained using non-negative matrix factorization consensus clustering [11].

Yuan et al. presented a nonparametric Bayesian model that combines copy number and expression data to discover cancer subtypes. This discovery method combining

genomics and transcriptomics, provides more comprehensive understanding of the functional components and pathway regulations for each cancer subtype [61].

Shen et al. presented an integrative subtype analysis of glioblastoma (TCGA GBM) dataset. They demonstrated that this clustering analysis method provided a biologically diverse source for subtype discovery [62].

Levine et al. developed a clustering algorithm, named SuperCluster to obtain overall subtypes for the samples based on their cluster memberships of different data types. And their results classified endometrial cancers based on integrated genomics data [63].

Verhaak et al. described a gene-expression-based molecular classification of GBM into four subtypes. To detect robust clusters, they used consensus clustering [64].

Cho and Przytycka developed a new technique for modeling of cancer heterogeneity. This unsupervised method models the individual cancer cases as mixtures of subtypes [65]. In their study, they suggested that GBM could be better explained by three subtypes instead of four subtypes as previously propose [64].

Hofree et al. introduced a new approach named Network-based stratification (NBS) to stratify tumors into meaningful subtypes by cluster patients with mutations in similar network regions [3]. NBS is known as the first method in which somatic mutation profiles have been used for stratifying patients. This method integrates gene networks with tumor molecular profiles to overcome some problems, such as somatic mutation profiles are very sparse and unusually heterogeneous. NBS considers the sparsity of mutations at network level, and it is used to identify subgroups of patients by spreading the influence of each mutation profile in a gene interaction network. As described in the section 2.2.1.1, basic algorithm of NBS is a network propagation [38] that aggregates mutations impacting the same subnetwork regions. It uses a random walk model (Random Walk with Restarts, RWR). Also, in their study to derive a stratification of the input cohort a variant of NMF (GNMF) [42] was used. Finally, the technique of consensus clustering [46] was used to identify robust cluster assignments.

Wang et al. introduced an approach named ‘Similarity Network Fusion’ (SNF) that provides clinically relevant patient subtypes. SNF constructs networks of patients for each individual data type and then fuse these data into one single network. In this way it improved the performance of popular integrative approaches at the time [66].

Speicher NK and Pfeifer N proposed that unsupervised multiple kernel learning be used to discover biologically meaningful subtypes for five different cancer types in their

work [67]. This method provided more flexibility for each data type than SNF since it generates a variety of dimension reduction approaches [66].

Yang et al. [68] used a protein-protein interaction network and somatic mutation profiles to classify patients into molecular subtypes. In their study, to overcome the heterogeneity of mutation profiles network propagation algorithm was employed. And the final smoothed mutation profiles of prostate cancer were input into the graph regularized NMF (GNMF) algorithm. Lastly unsupervised consensus clustering was used to identify a predefined number of subtypes. However, since the true subtypes for prostate cancer were unknown until then, no clear result could be given about validation performance of the stratification.

intNMF (an integrative approach for disease subtype classification based on NMF) was proposed by Chalise et al. and it aims to classify disease into distinct subtypes. It was stated that it has advantages over other clustering algorithms that need distributional assumptions because it makes no assumptions about the data's distributional form [69].

A study published in 2017, combined somatic mutation (endometrial cancer) and gene expression data to identify patient clusters [70]. Unlike the NBS method [3], cancer-type-specific significant co-expression networks (SCNs) were created instead of using a fixed gene network in all cancers.

Kuijjer et al. described a new method to identify cancer subtypes using tumor somatic mutation profiles. This method uses biological pathways to overcome sparseness and heterogeneity of the somatic mutation data [21].

Zhang et al. [7] introduced a supervised method named Network-Based Supervised Stratification to classify tumors as described in the section 2.2.1.2. It extends the Supervised Random Walk algorithm by including a new loss function used for classification of cancer subtypes. It uses the network propagation technique for aggregation of mutations affecting the same subnetwork regions, same as other methods. Unlike these methods, this method uses supervised learning to adjust the weight of each molecular interaction.

Mun et al. [71] presented proteogenomics analysis of diffuse gastric cancers (GCs) in young populations. In this study, four subtypes of diffuse gastric cancers were identified by integrating analysis of mRNA and protein data.

Xu, et al. (2020) [72] introduced a partial multi-omics integrative technique for cancer subtyping. The method named MSNE (Multiple Similarity Network Embedding)

is a network embedding based integrative method and it can capture the similarity of samples, even though some samples are not found in the same omics as others.

Rohani and Eslahchi (2020) [73] presented a study to discover subtypes of breast cancer using the somatic mutations and CNAs data. They used network propagation method to make the somatic mutation profiles dense. Then, they used the deep embedded clustering method to classify breast tumors into subtypes.

Liu et al. (2021) [74] developed a network-based deep learning algorithm to identify patient subtypes from somatic mutation profiles. This stratification methodology based on network embedding and argues that two tumors can be classified into the same subtypes if the somatic mutated genes of these tumors are found in similar network regions of interaction network.

Although all these existing studies provide valuable contributions to cancer subtype discovery, they have some limitations. Subtypes derived from expression profiles are not associated with clinical outcomes such as patient survival and response to therapy. In studies that stratify tumors into subtypes using somatic mutation profiles, the use of Random Walk with Restart (RWR) method may cause ill-conditioning problem. In this thesis, our aim is to elude this ill-conditioning problem using our approach. Network embedding-based methods have classified tumor subtypes using supervised learning and training stage can take a long time. Also, our approach has simple and easy implementation compared to them.

Chapter 4

Materials and Methods

4.1 Input Data

The method proposed in this thesis requires two input data to cluster tumor mutation profiles into robust tumor subtypes: a molecular interaction network (reference molecular network) that contains gene-gene interactions and a tumor-by-gene binary matrix that represents somatic tumor mutation profile of cancer patients.

Gene-gene interaction networks describe the functional interactions between pairs of genes. Knowledge of these interactions could provide essential information about complex diseases. The causative drivers of tumor growth are thought to be contained in the mutations as stated in the section 2.1.2.4. So, somatic mutation profiles could be informative for tumor stratification.

For a better understanding, representation of somatic tumor mutation profiles was shown in Equation 4.1 as a tumor-by-gene binary matrix. 1 indicates the mutated genes and 0 indicates the wild types (non-mutated). Figure 4.1 is a visualized version of a small excerpt of the colon cancer somatic mutation data. The mutated genes are shown in blue.

$$\begin{array}{c} \\ \textbf{Tumors} \end{array} \begin{array}{c} \textbf{Genes} \\ \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \vdots & \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \end{bmatrix} \end{array} \tag{4.1}$$

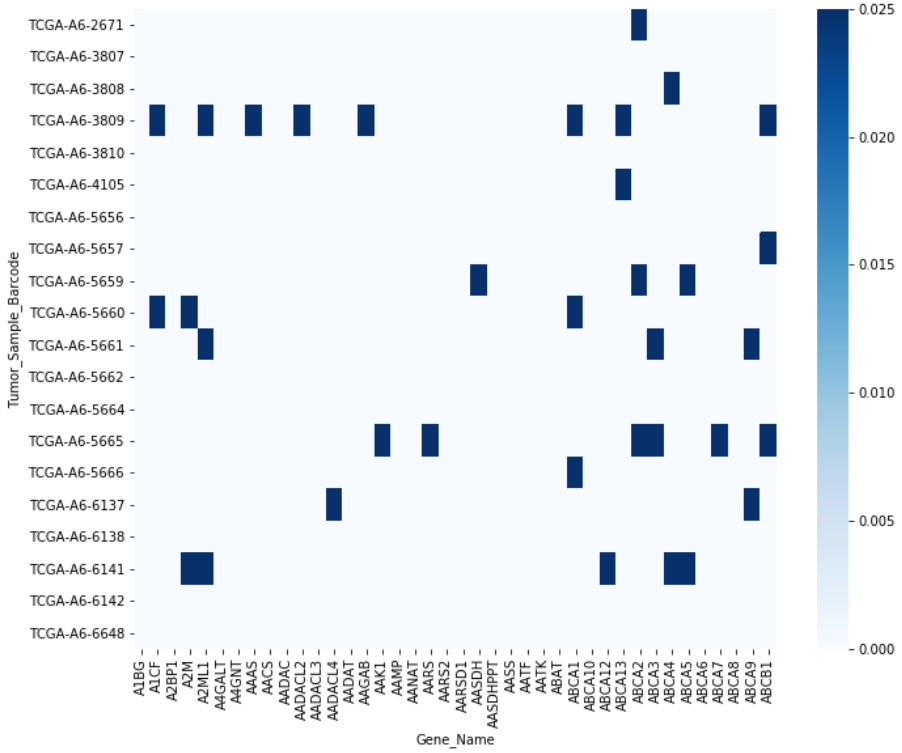

Figure 4.1 An example for somatic tumor mutation profiles

In this thesis, a reference molecular (gene-gene interaction) network and somatic tumor mutation profiles of different types of cancer were used for the stratification of tumors into subtypes. As a reference molecular network, a filtered network retaining only cancer genes was used. This filtered network has 2291 nodes [47]. The tumor mutation profiles derived from the TCGA (The Cancer Genome Atlas) belong to the following cancer types: Colon: colorectal adenocarcinoma, uterine: uterine corpus endometrial carcinoma, head and neck: head-neck squamous cell carcinoma, bladder: urothelial bladder carcinoma and breast cancer.

For five tumor mutation datasets, the number of tumors and genes are shown in Table 4.1.

Table 4.1 Descriptive statistics of the datasets

Dataset	Number of tumors	Number of genes
COAD (Colon)	315	17390
UCEC (Uterine)	248	17341
HNSC (Head and Neck)	510	16521
BLCA (Bladder)	395	17201
BRCA (Breast)	286	571

4.2 Label Propagation

By considering $\mathbf{F_0} \in \mathbb{R}^{n\times K}$, where K ≪ n, (tumor-by-gene binary matrix) as a label set matrix, we can define the tumor stratification problem as a well-known label propagation algorithm [75]. To be more specific, we will use the indicator values of each column of the $\mathbf{F_0}$ matrix as a label of that of tumor. Since $K \ll n$ we need to interpret the rest of the known label methodology. To this end, we will define known labels as $\mathbf{Y_L} = (\mathbf{y_1}, \dots \dots \mathbf{y_K})$ and unknown labels as $\mathbf{Y_U} = (\mathbf{y_{K+1}}, \dots \dots \mathbf{y_{K+u=n}})$. Now, our objective is to determine the set of $\mathbf{Y_U}$ by depending on $\mathbf{Y_L}$ and the graph's topological structure. To be more consistent with the terminology, we will call $\mathbf{Y_L} = \mathbf{F_0}$.

4.3 Proposed Method

By using the above defined $\mathbf{F_0} \in \mathbb{R}^{n\times K}$ (tumor-by-gene binary matrix) as known labels in our setting, we will define the tumor stratification problem as a label propagation approach. Let $\mathbf{f_0} \in \mathbb{R}^n$ denotes a prior known vector, where the location of 1s indicates that the belonging of the nodes to the clusters. In this thesis, the basic premise is that we assume the manifold smoothness of known labels and penalize the sparseness of unknown labels ($\widehat{\mathbf{f_0}}$). Mathematically, we define the smoothness as follows [76]:

$$\begin{aligned} \boldsymbol{Smoothness}\left(\widehat{\mathbf{f_0}}\right) &= \sum_{\mathbf{i,j=1}}^{\mathbf{n}} \mathbf{A_{i,j}}\,(\mathbf{f_{0_i}} - \widehat{\mathbf{f_{0_j}}})^2 \\ &= \textstyle\sum_{\mathbf{i,j=1}}^{\mathbf{n}} \mathbf{A_{i,j}}\,(\mathbf{f_{0_i}}^2 - 2\mathbf{f_{0_i}}\widehat{\mathbf{f_{0_j}}} + \widehat{\mathbf{f_{0_j}}}^2) \\ &= 2\widehat{\mathbf{f_{0_j}}}^{\boldsymbol{T}}(\mathbf{D} - \mathbf{A})\mathbf{f_{0_i}} \end{aligned} \qquad \textbf{(4.2)}$$

$$= 2\widehat{\mathbf{f_0}}\mathbf{L}\mathbf{f_0}$$

Here, $\boldsymbol{L}$ denotes the graph Laplacian, and defined as $\boldsymbol{L = D - A}$. $\boldsymbol{A}$ is the degree normalized adjacency matrix of $\mathcal{G}$ and $\boldsymbol{D}$ represents the diagonal degree matrix of $\boldsymbol{A}$. The sparsity of the assignments can be measured as follows:

$$\textbf{Sparsity}\left(\widehat{\mathbf{f_0}}\right) = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\widehat{\mathbf{f}_{0_i}})^2 = \left\|\widehat{\mathbf{f_0}}\right\|^2 \quad \textbf{(4.3)}$$

Formalization of the objective function is obtained by combining these two equations. Then we have following objective:

$$\mathbf{J}(\widehat{\mathbf{f_0}}) = \widehat{\mathbf{f_0}}\mathbf{L}\mathbf{f_0} + \boldsymbol{\sigma}\left\|\widehat{\mathbf{f_0}}\right\|^2 \quad \textbf{(4.4)}$$

Here, we aim to find the $\widehat{\mathbf{f_0}}$ that minimizes $J(\widehat{\mathbf{f_0}})$. The final term serves applying sparsity penalizing to solutions which are too far from zero. The parameter $\boldsymbol{\sigma}$ configures the effect of this penalization. By taking the partial derivative of $\boldsymbol{J}$ with respect to $\widehat{\mathbf{f_0}}$, the following equation is obtained:

$$\frac{\partial \mathbf{J}}{\partial \widehat{\mathbf{f_0}}} = \widehat{\mathbf{f_0}}\mathbf{L} + \boldsymbol{\sigma}\widehat{\mathbf{f_0}} \quad \textbf{(4.5)}$$

If the Equation 4.5 set to 0 then $\widehat{\mathbf{f_0}}$ that minimizes $J(\widehat{\mathbf{f_0}})$ can be calculated as:

$$\widehat{\mathbf{f_0}} = (\mathbf{L} + \boldsymbol{\sigma}\mathbf{I})^{-1}\mathbf{f_0} \quad \textbf{(4.6)}$$

Now, by using the above equation, we try to find the clusters of $\widehat{\mathbf{f_0}}$ by relying on graph Laplacian, we stratify the tumors. Our approach here is a reflection of Ridge regularization on tumor clustering.

Firstly, we rewrite the Random Walk with Restarts (RWR) equation as follows [77]:

$$\mathbf{F_{t+1}} = \boldsymbol{\alpha}(\mathbf{I} - (\mathbf{1} - \boldsymbol{\alpha})\mathbf{A})^{-\mathbf{1}}\mathbf{F_0} \quad \textbf{(4.7)}$$

Hofree et al. [3] uses random walk model to diffuse information about tumor mutations using molecular interaction knowledge in the network. Instead, here in this thesis, we propose the above label propagation approach by changing the part of $\alpha(\mathbf{I} - (1 - \alpha)\mathbf{A})$ with $\mathbf{L} + \sigma\mathbf{I}$, so that we can elude the ill-conditioning problem that might be introduced by RWR approach. Finally, we define the diffuse strategy of tumor mutations using knowledge of molecular interaction as:

$$\mathbf{F_{t+1}} = (\mathbf{L} + \boldsymbol{\sigma}\mathbf{I})^{-\mathbf{1}}\mathbf{F_0} \quad \textbf{(4.8)}$$

In the Equation 4.8 the parameter σ is set to 0.01, 0.1 and 0.2 to use a different value in each run. After obtaining propagated mutation profiles we applied non-negative matrix factorization to get patient clusters. 80% of somatic mutation matrix rows and columns was subsampled without replacement. Graph regularized NMF (GNMF) was performed 100 times on subsamples of the dataset to stratify the input cohort. To produce robust patient clusters consensus clustering is used and a final stratification of the patients into clusters is recovered. Aggregate GNMF results of 100 samples was converted into a co-clustering matrix as described in section 2.2.1.1. Each element in this matrix represents the frequency with each two tumors was discovered to belonging the same cluster among all clustering iterations.

Chapter 5

Performance Results

This section presents the performance results of our proposed method evaluated on various datasets. To evaluate the performance of our method, the results are compared against the state-of-the-art unsupervised and supervised methods. Our method was tested in colon: colorectal adenocarcinoma, uterine: uterine corpus endometrial carcinoma, head and neck: head-neck squamous cell carcinoma and bladder: urothelial bladder carcinoma datasets. And we present the comparative results with the NBS method (unsupervised) are given in the section 5.1. Our method also tested in breast cancer dataset. And the comparative results with the validation performance of NBS^2 method (supervised) are given in section 5.2.

5.1 Comparison of our proposed method with unsupervised tumor stratification

To make fair comparisons with NBS [3], we evaluate our proposed method on the datasets used in their python implementation paper [74]. We applied the NBS and our method to cluster somatic mutation profiles of four cancer: colon (COAD), uterine (UCEC), head and neck (HNSC), and bladder (BLCA). As a reference molecular network, we used a filtered network which has to preserve only cancer genes. This filtered network has 2291 nodes [74].

Our label propagation-based method was executed with 3 different $\boldsymbol{\sigma}$ values (0.01, 0.1, 0.2) for each data. Both these results and the results of the NBS method were evaluated and presented in the following section. In this thesis, to evaluate the clustering performances, following metrics mentioned in the section 2.2.2 are used: Silhouette

coefficient, Davies-Bouldin index, intra-cluster distance, t-Distributed Stochastic Neighbor Embedding (t-SNE) and Co-clustering heat map.

Each method was run five times, each time the Silhouette coefficient and Davies-Bouldin index were measured, and the results were shown by the boxplots. A higher average Silhouette coefficient and a lower Davies-Bouldin index indicates better clustering quality. Intra cluster distance values (complete, average, and centroid) are calculated separately for each extricated cluster. For a better assessment, the calculated values for each cluster were averaged and the results are shown by tables. Tables should be interpreted considering the information that intra-cluster distance should be minimum to obtain the best clustering result. Therefore, the lowest distance in each column is highlighted in bold.

Clustering of colon cancer: By applying the NBS and our method to colon cancer data, patient profiles are clustered into 3 predefined subtypes. Figure 5.1 shows the performances of clustering methods for colon cancer (COAD). The boxplot shows the Silhouette coefficient and Davies-Bouldin index scores obtained by running each method five times. The intra cluster distance results are given in the Table 5.1. Figure 5.2 shows visualization of the clusters of COAD found using different methods. (A) Co-clustering maps: In all maps except of NBS, blue blocks along the diagonal on a white background. And these maps represent well-defined 3-cluster structure. (B) t-SNE plots: When tumor mutation profiles are stratified using our method, it is seen that 3 clusters are well grouped among themselves.

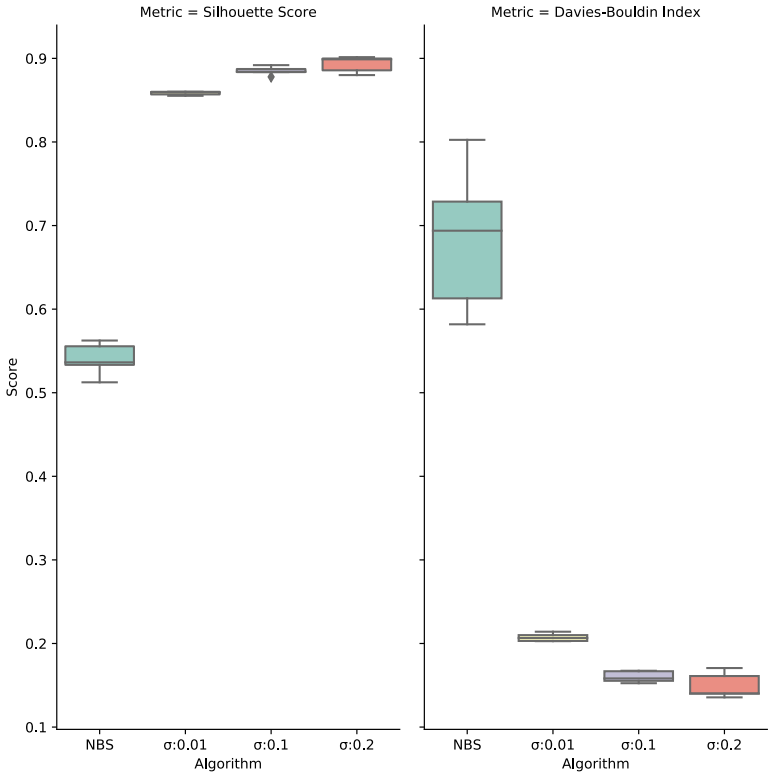

Figure 5.1 Performance evaluation for COAD

Table 5.1 Intra cluster distances for COAD

Distance / Algorithm	*Complete*	*Average*	*Centroid*
NBS	7,906	3,038	2,274
σ = 0.01	6,180	1,193	1,110
σ = 0.1	**6,118**	**1,107**	**1,101**
σ = 0.2	7,372	1,516	1,213

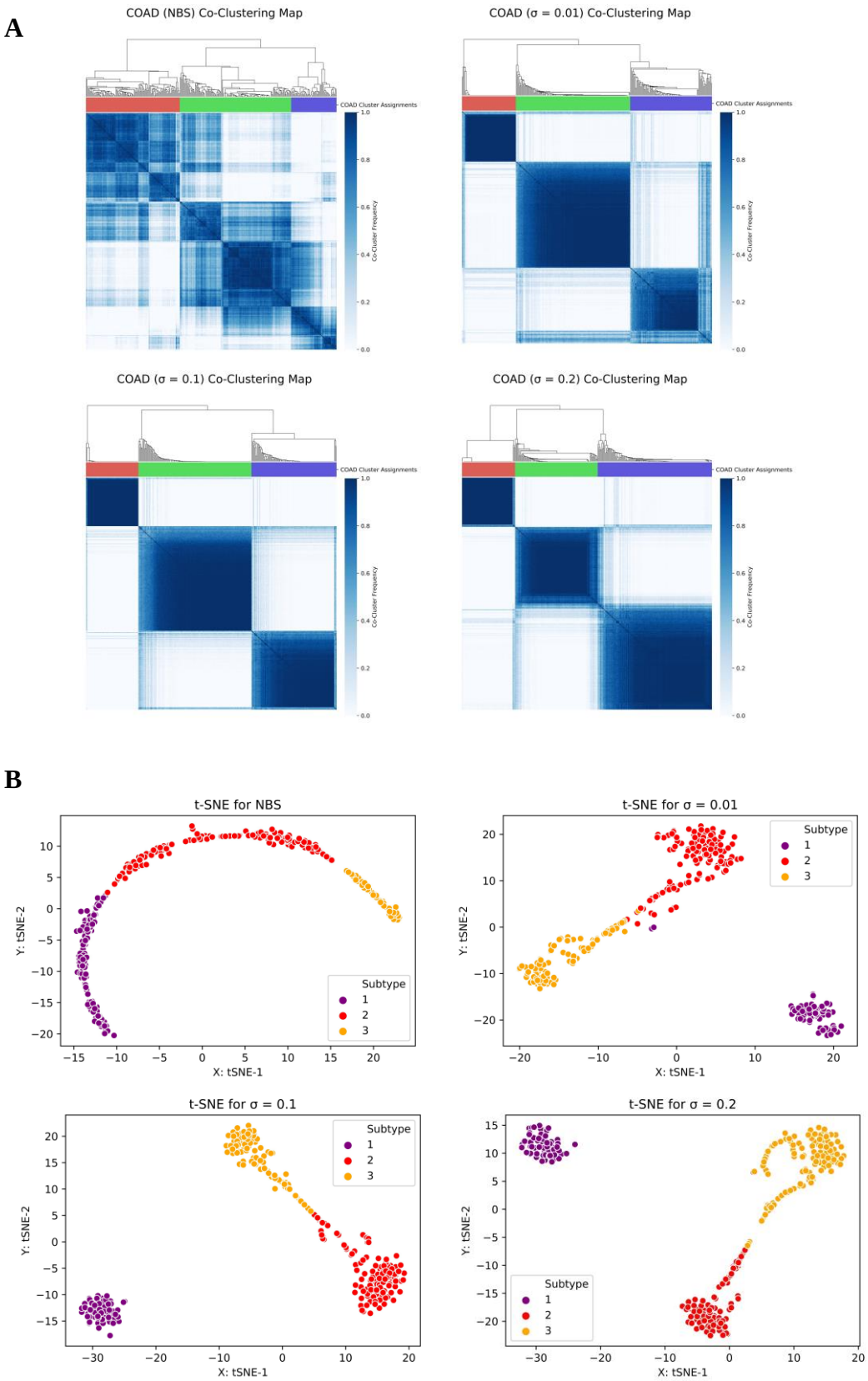

Figure 5.2 Visualizing clusters of COAD

Clustering of uterine cancer: Using uterine cancer data, patient profiles are clustered into 4 predefined subtypes. Figure 5.3 shows the performances of clustering methods for uterine cancer (UCEC). The boxplot shows the Silhouette coefficient and Davies-Bouldin index scores obtained by running each method five times. Intra cluster distance results of UCEC are shown in the Table 5.2. Figure 5.4 shows the visualization of clustering results. (A) Co-clustering maps: In all maps except of NBS, 4 blue blocks along the diagonal on a white background. And these maps represent well-defined 4-cluster structure. (B) t-SNE plots: When tumor mutation profiles are stratified using our method, it is seen that 3 clusters are well grouped among themselves.

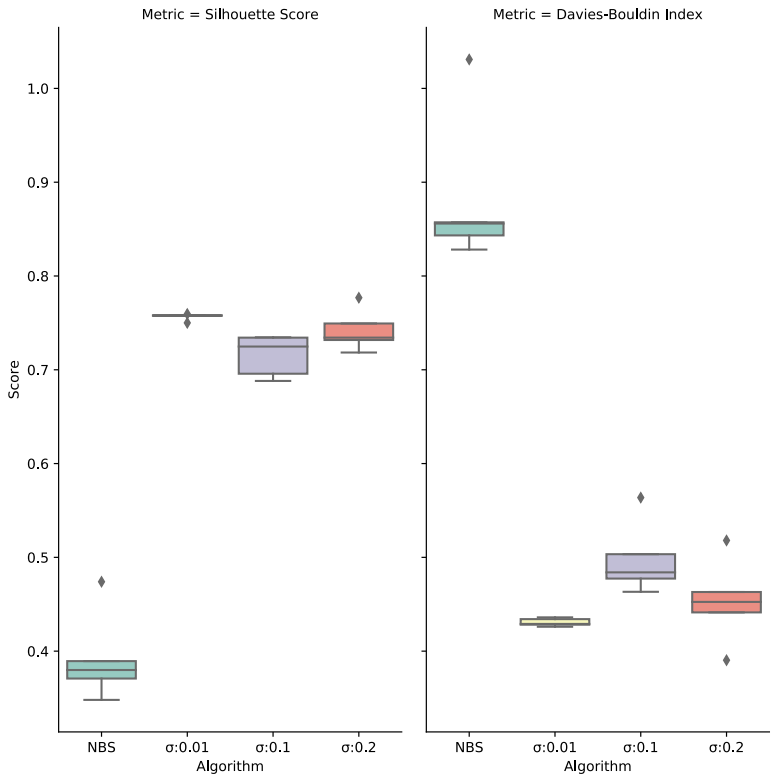

Figure 5.3 Performance evaluation for UCEC

Table 5.2 Intra cluster distances for UCEC

Distance / Algorithm	*Complete*	*Average*	*Centroid*
NBS	5,581	2,658	1,870
σ = 0.01	**4,195**	**1,704**	**1,141**
σ = 0.1	4,385	1,841	1,256
σ = 0.2	4,853	1,963	1,384

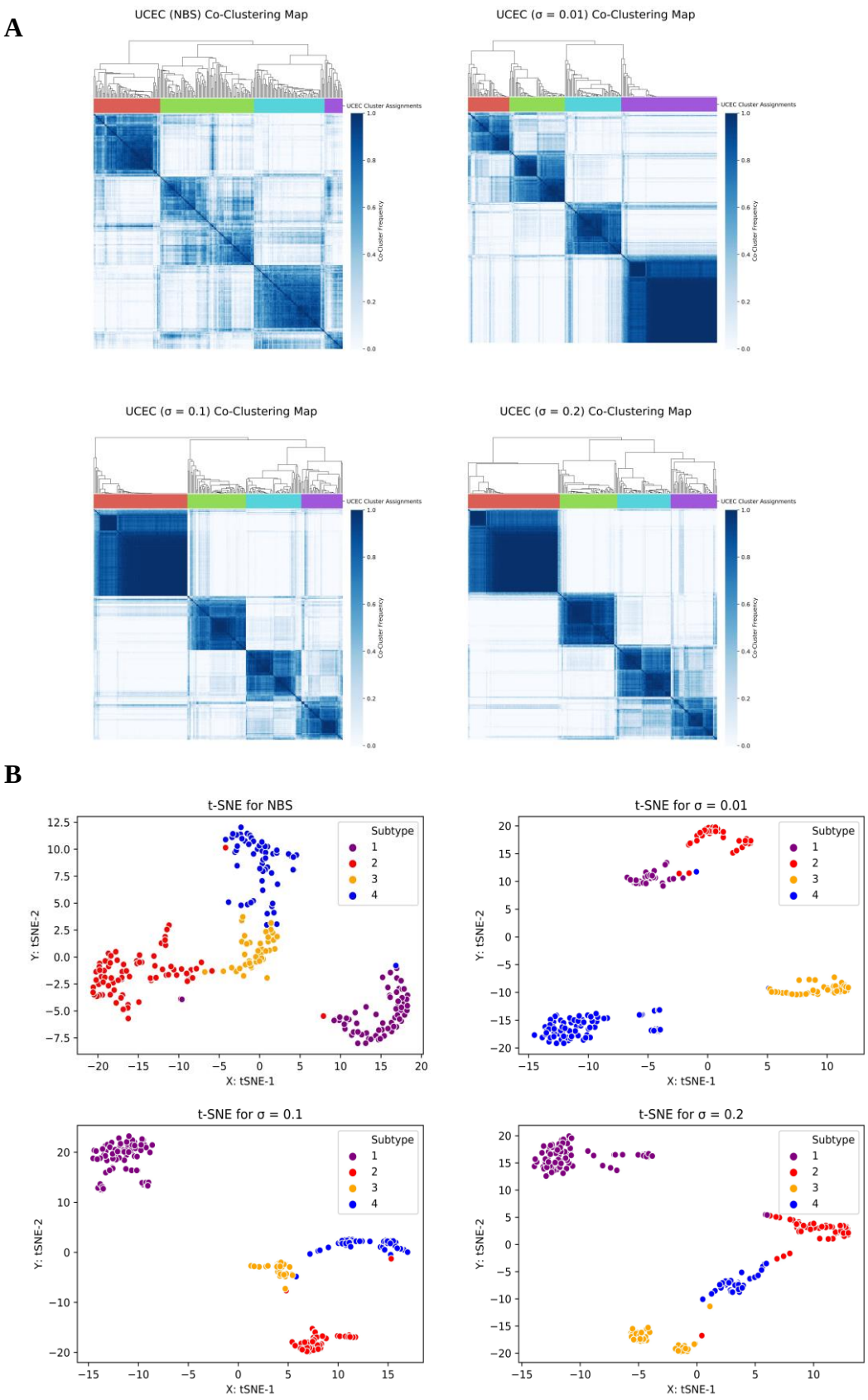

Figure 5.4 Visualizing clusters of UCEC

Clustering of head and neck cancer: Using head and neck cancer data, patient profiles are clustered into 4 predefined subtypes. Figure 5.5 shows the performances of clustering methods for head and neck cancer (HNSC). The boxplot shows the Silhouette coefficient and Davies-Bouldin index scores obtained by running each method five times. Intra cluster distance results of HNSC are shown in the Table 5.3. Figure 5.6 shows the visualization of clustering results for head and neck cancer (HNSC). (A) Co-clustering maps: Although there is no clear difference between the maps, 4 blue blocks on maps belonging to our method are defined better. (B) t-SNE plots: Visualization of clusters grouped among themselves is given.

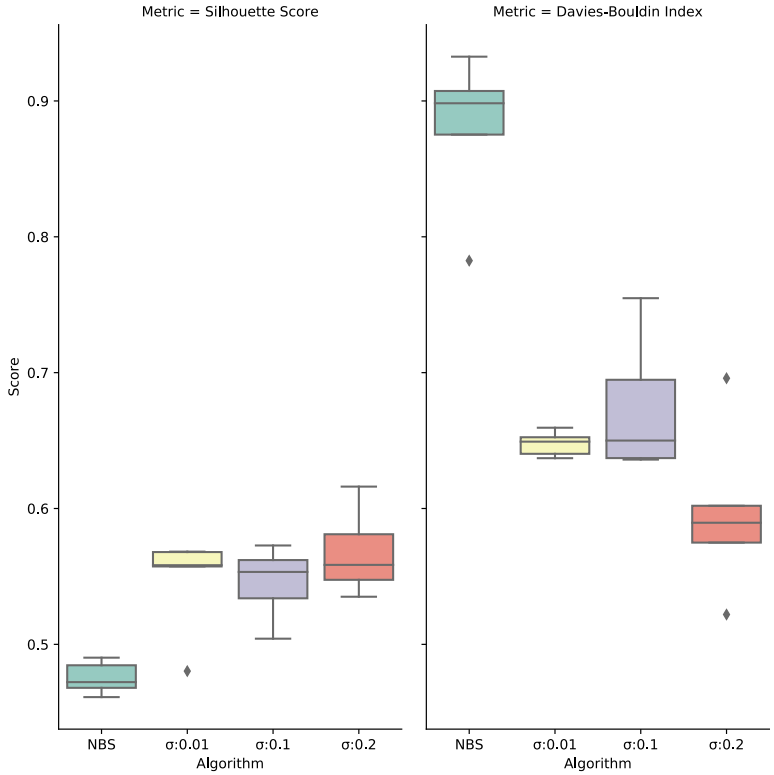

Figure 5.5 Performance evaluation for HNSC

Table 5.3 Intra cluster distances for HNSC

Distance / Algorithm	*Complete*	*Average*	*Centroid*
NBS	8,967	3,732	2,738
σ = 0.01	8,202	**3,130**	**2,147**
σ = 0.1	**8,142**	3,207	2,214
σ = 0.2	8,396	3,389	2,309

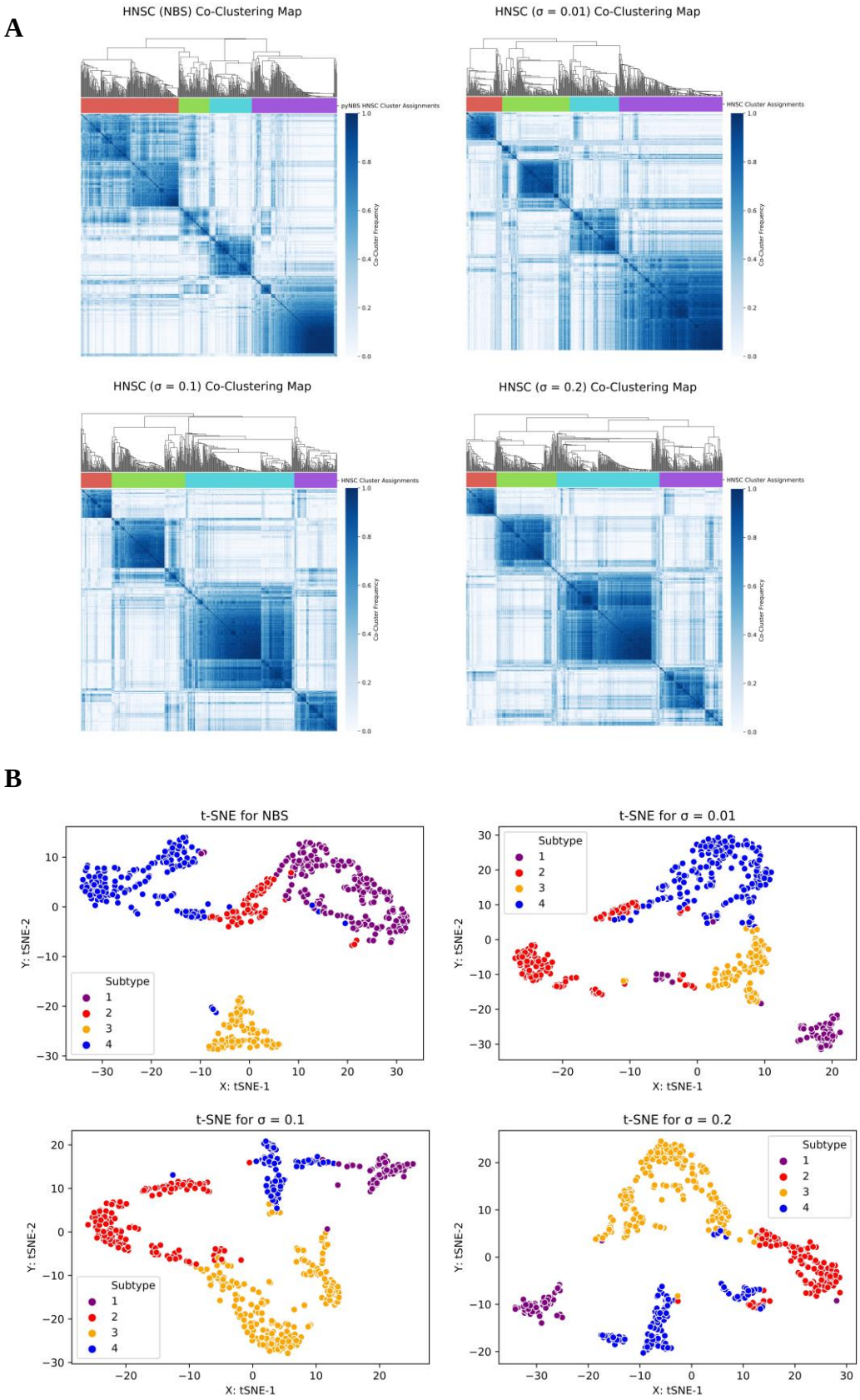

Figure 5.6 Visualizing clusters of HNSC

Clustering of bladder cancer: When the NBS and our method are applied to bladder cancer data, patient profiles are clustered into 4 predefined subtypes. Figure 5.7 shows the performances of clustering methods for bladder cancer (BLCA). The boxplot shows the Silhouette coefficient and Davies-Bouldin index scores obtained by running each method five times. Intra cluster distance results of BLCA are shown in the Table 5.4. Figure 5.8 shows the visualization of clustering results for bladder cancer (BLCA). (A) Co-clustering maps: There is no clear difference between the co-clustering maps. When examining the maps given here, the information that a darker blue color corresponds to higher clustering for tumor pairs should be considered. (B) Visualization of clusters grouped among themselves is given.

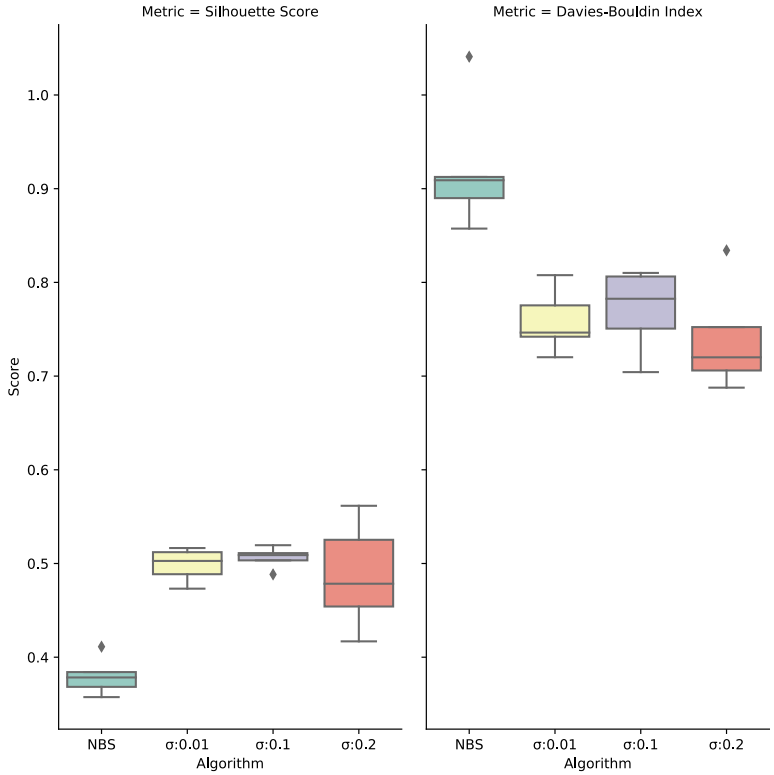

Figure 5.7 Performance evaluation for BLCA

Table 5.4 Intra cluster distances for BLCA

Distance / Algorithm	*Complete*	*Average*	*Centroid*
NBS	6,624	2,371	1,677
σ = 0.01	**6,153**	2,290	**1,420**
σ = 0.1	6,182	**2,264**	1,494
σ = 0.2	6,480	2,300	1,501

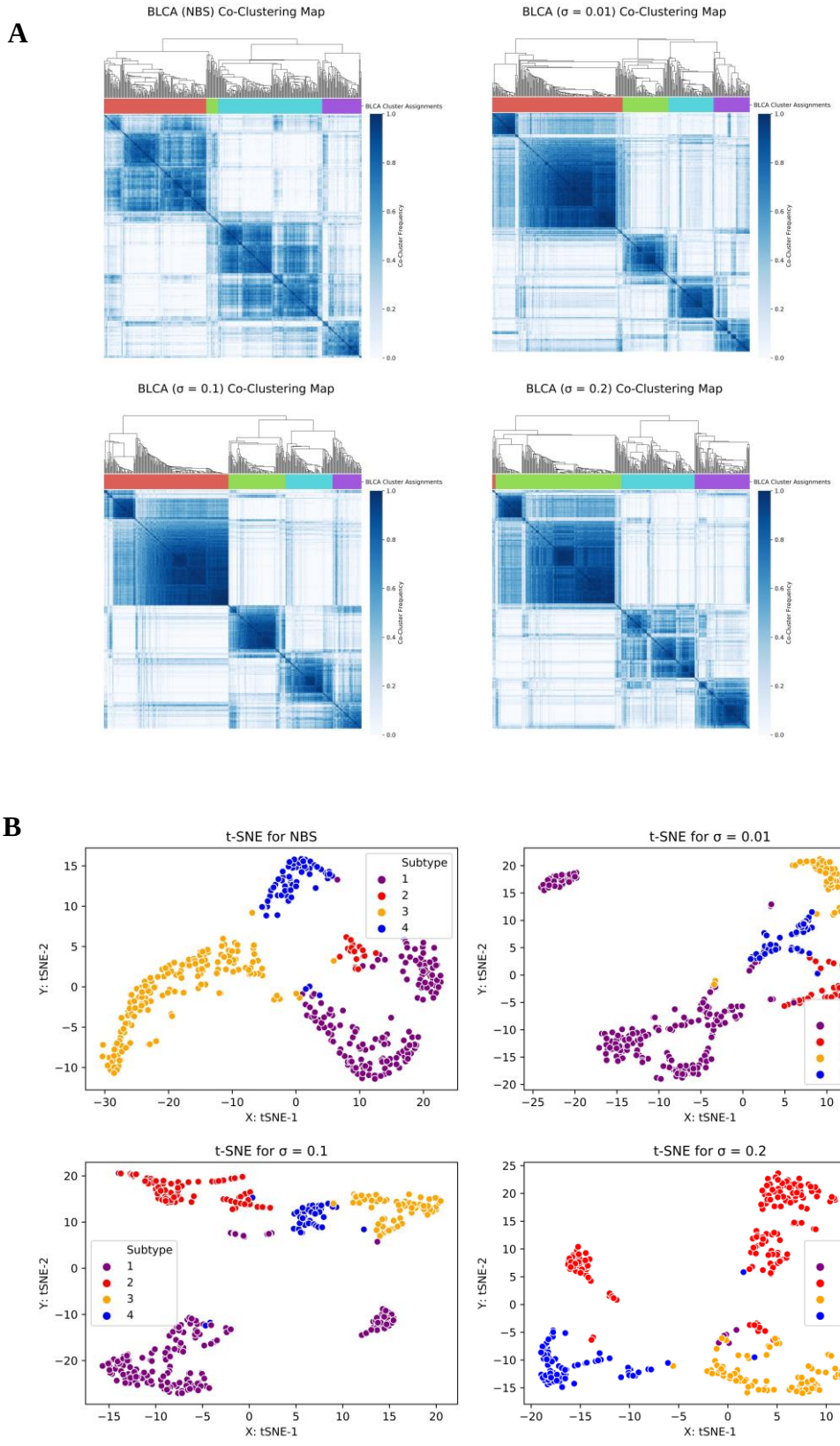

Figure 5.8 Visualizing clusters of BLCA

5.2 Comparison of our proposed method with supervised tumor classification

In this section we applied Network-based Supervised Stratification (NBS^2) which is the baseline supervised method for cancer subtype identification [7] on breast cancer data and compared validation performance of NBS^2 to performance of our proposed method. Training dataset used for the NBS^2 method contains 577 tumors and 571 genes while validation dataset contains 286 tumors and 571 genes. Also reference molecular network with 557 genes was used. The accuracy of 286-tumor validation set increased from 54 to 58% at the end of 316 iterations in total. It is not easy to compare a supervised method and an unsupervised method comprehensively. So, to compare our unsupervised method with the NBS^2 we used a comparison just as Zhang et al. did [7]. To make validation set predictions NBS^2 calculates the cluster centroids of the training set. We evaluated the performances of algorithms and our method achieved 60% accuracy as the best score, a performance 2% upper than that of NBS^2 as seen in the Figure 5.9. Classification accuracy is plotted against the number of NBS^2 iterations on the validation data. Accuracy of unsupervised methods are equal for each iteration.

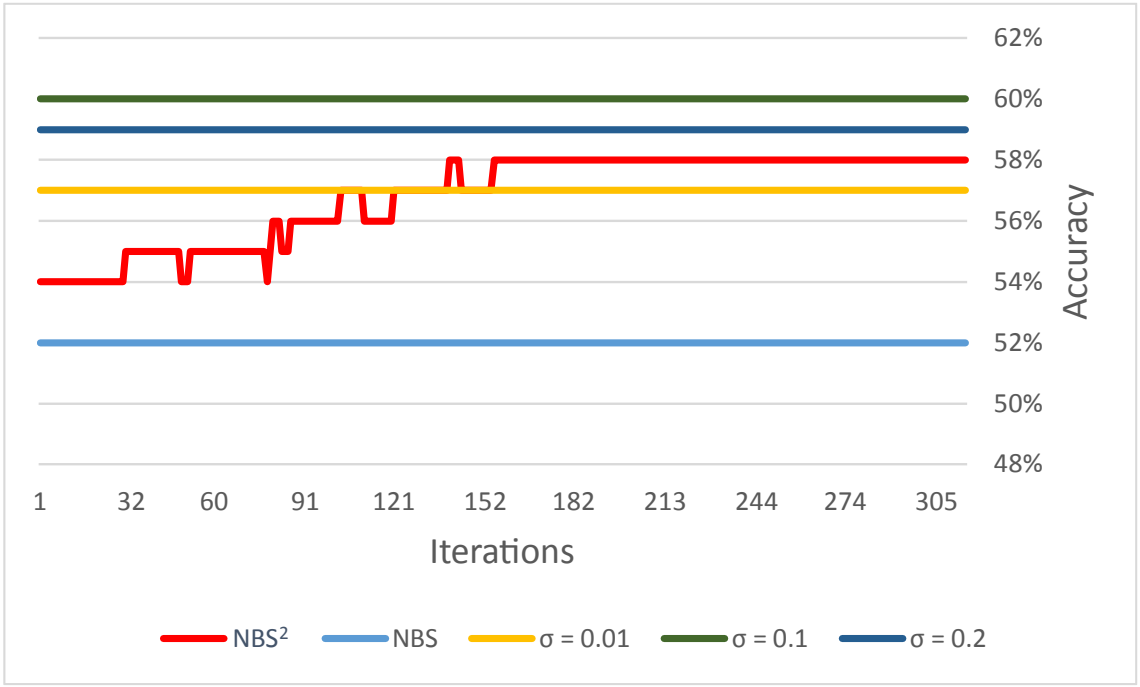

Figure 5.9 Performances of algorithms for breast cancer subtype identification

Chapter 6

Discussions

Experiments are performed on five different disease datasets: colon, head and neck, uterine, bladder and breast cancer in order to evaluate the performance of the proposed method in this thesis. These datasets consist of 315, 510, 248, 395 and 286 tumor samples respectively.

In order to interpret the outcomes, the proposed method is compared with network-based stratification method (NBS) [3] which is an unsupervised method. To evaluate the clustering performance of our method on colon, head and neck, uterine, bladder datasets; Silhouette coefficient, Davies-Bouldin index, intra-cluster distance, t-Distributed Stochastic Neighbor Embedding (t-SNE) and co-clustering heat map are used. According to the performance evaluation results given in section 5.1 our label propagation-based method (with all three $\boldsymbol{\sigma}$ values) drastically outperforms the NBS method in subtype stratification. We have seen that $\boldsymbol{\sigma}$ parameter can be tuned and changing it may have different conclusions for stratification results.

To make a comparison with a supervised method (NBS^2) [7], we use breast cancer dataset and calculated the cluster centroids of the training set. Although results given in section 5.2 does not seem to make any sense mathematically, considering that our method is an unsupervised model, it is a promising result that the performance of subtype identification without using any label obtains close results with the subtype classification using label.

Supervised methods use labeled datasets and node-labelling is often expensive and time consuming. And this can be disadvantageous in some cases. So, a method can show a competitive performance even without any label would be very useful to identification of cancer subtypes.

Chapter 7

Conclusions and Future Prospects

7.1 Conclusions

Response of many types of cancer treatment varies from patient to patient. This variability emerges the need to identify characteristics of patient tumor mutation profiles and cluster patients based on their genomic similarity. Discovering cancer subtypes is one of the research subject of cancer informatics. To date, various approaches have been developed for cancer subtype identification using many types of omics data. In this thesis, a new unsupervised method is presented for stratify tumor mutation profiles into meaningful subtypes. This method is based on label propagation, and it takes two input datasets: a reference molecular network and tumor mutation profile of a cohort. Using various datasets and extensive experimental configurations on these datasets, we show that our proposed approach outperforms the alternative methods in identifying cancer subtypes in large margin.

7.2 Societal Impact and Contribution to Global Sustainability

This study, which is based on the division of a heterogeneous tumor population into informative subtypes as determined by the similarity of molecular profiles, contributes to cancer subtype identification which has an important place in cancer informatics. These tumor subtypes are related with important clinical outcomes mentioned in Section 2.1.1. These relations show that knowing cancer subtype of patient provides a more suitable treatment for the patient. In this respect, this thesis contributes to society by reducing the side effects of drugs on the patient and increasing the effectiveness of the treatment.

Moreover, identifying the different genetic subtypes of patients demonstrates the potential applicability of this thesis to advance personalized medicine. Personalized medicine advancements provide a more integrated therapeutic approach that is specific to the genome of the individual. By providing more accurate diagnosis and early intervention, as well as tailored therapy, personalized medicine has the potential to reduce suffering and the cost of cancer treatment.

The main contribution of this thesis is that presenting a label propagation-based unsupervised method which outperforms the state-of-the-art methods for cancer subtype identification. As described in Section 4.3, our proposed method includes a label propagation approach based on assuming manifold smoothness of known labels and penalizing the sparseness of unknown labels. We can avoid the ill-conditioning problem that might be introduced by the Random Walk with Restart approach utilized in previous studies to stratify cancer into subtypes by using the methodology we suggest.

In summary, awareness of tumor heterogeneity has increased in recent years as cancer research has learned a lot about the genetic diversity of cancer types. Knowledge of cancer subtypes is crucial for understanding and interpreting tumor heterogeneity. The methodology proposed in this thesis can be applied on many cancer types to stratify them into informative subtypes. In this way, the treatment that will benefit the patient the most can be chosen.

7.3 Future Prospects

In the future work, our study can be expanded in several ways. First, our proposed method could be applied to other cancer types where somatic mutation information is available. With the discovery of meaningful subtypes of different cancer types, we can contribute to cancer research. Second, for high performance, network embedding approach can be implemented to our proposed method. This approach can help with the cancer subtype identification problem by its ability to convert the network into a low-dimensional space while preserving the network's structural information.

BIBLIOGRAPHY

[1] Cancer, https://www.who.int/news-room/fact-sheets/detail/cancer (October 14, 2020)
[2] Precision Medicine in Cancer Treatment, https://www.cancer.gov/about-cancer/treatment/types/precision-medicine (June 16, 2020)
[3] M. Hofree, J. P. Shen, H. Carter, A.Gross, T. Ideker, "Network-based stratification of tumor mutations," Nature methods, 10, 1108-1115 (2013).
[4] N. Jin, H. Wu, Z. Miao, Y. Huang, Y.Hu, X. Bi, D. Wu, K. Qian, L. Wang, C. Wang, H. Wang, K. Li, X. Li, and D. Wanga, "Network-based survival-associated module biomarker and its crosstalk with cell death genes in ovarian cancer," Scientific Reports, 5, 11566 (2015).
[5] Y. A Kim, D. Y. Cho, T. M. Przytycka, "Understanding Genotype-Phenotype Effects in Cancer via Network Approaches," PLoS Computational Biology, 12, e1004747 (2016).
[6] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, I. Lee, "MUFFINN: cancer gene discovery via network analysis of somatic mutation data," Genome Biology, 17, 129, (2016).
[7] W. Zhang, J. Ma, T. Ideker, "Classifying tumors by supervised network propagation," Bioinformatics, 34, i484–i493 (2018).
[8] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," Nature, 474, 609–615 (2011).
[9] C. M. Perou et al., "Molecular portraits of human breast tumours," Nature, 406, 747–52 (2000).
[10] P. Mischel et al., "Identification of molecular subtypes of glioblastoma by gene expression profiling," Oncogene, 22, 2361–2373 (2003).
[11] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," Nature, 474, 609–615 (2011).
[12] J. S. Reis-Filho and L. Pusztai "Gene expression profiling in breast cancer: classification, prognostication, and prediction," Lancet, 378, 1812–1823 (2011).
[13] M. R. Stratton, P. J. Campbell, P. A. Futreal, "The cancer genome," Nature, 458, 719-724 (2009).
[14] Histology Stains, https://dermnetnz.org/topics/histology-stains/ (October 15, 2020)
[15] M. Liang, Z. Li, T. Chen and J. Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12, 928-937 (2015).
[16] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," Nature Genetics, 45, 1113–1120 (2013).
[17] International Cancer Genome Consortium, "International network of cancer genome projects," Nature, 464, 993–998 (2010).
[18] What Is Cancer?, https://www.cancer.gov/about-cancer/understanding/what-is-cancer (August 8, 2020)
[19] Cancer Subtype, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer-subtype (August 8, 2020)
[20] C. J. Creighton, "Making Use of Cancer Genomic Databases," Current Protocols in Molecular Biology, 121, 19.14.1–19.14.13 (2018).

[21] M. L. Kuijjer, J. N. Paulson, P. Salzman, W. Ding and J. Quackenbush, "Cancer subtype identification using somatic mutation data," British Journal of Cancer, 118, 1492–1501 (2018).
[22] Gen ifadesi, https://tr.wikipedia.org/wiki/Gen_ifadesi (December 4, 2020)
[23] Y. Peng, C. M. Croce, "The role of MicroRNAs in human cancer," Signal Transduct Target Ther, 1, 15004 (2016).
[24] A. Prat and C. M. Perou, "Deconstructing the molecular portraits of breast cancer," Molecular Oncology, 5, 5–23 (2011).
[25] C. V. Brennan et al., "The somatic genomic landscape of glioblastoma," Cell, 155, 462–477 (2013).
[26] L. D. Moore, T. Le and G. Fan, "DNA Methylation and Its Basic Function," Neuropsychopharmacol, 38, 23–38 (2013).
[27] R. L. Momparler and V. Bovenzi, "DNA methylation and cancer," Journal of cellular physiology, 183, 145-54 (2000).
[28] A. R. Karpf and D. A. Jones, "Reactivating the expression of methylation silenced genes in human cancer," Oncogene, 21, 5496-5503 (2002).
[29] M. Kulis, M. Esteller, "DNA methylation and cancer," Advanced Genetics, 70, 27-56 (2010).
[30] R. Beroukhim, "The landscape of somatic copy-number alteration across human cancers," Nature, 463, 899–905 (2010).
[31] J. C. Smith and J. M. Sheltzer, "Systematic identification of mutations and copy number alterations associated with cancer patient prognosis," eLife, 7, e39217 (2018).
[32] Somatic mutation, https://www.cancer.gov/publications/dictionaries/cancer terms/def/somatic-mutation (October 22, 2020)
[33] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, 13, 8-17 (2014).
[34] E. Alpaydin, "Introduction to Machine Learning (2nd. ed.)," The MIT Press, (2010).
[35] S. Ray, "A Comparative Analysis and Testing of Supervised Machine Learning Algorithms," (2018)
[36] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences (2005).
[37] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, A. A. Moustafa, "The Application of Unsupervised Clustering Methods to Alzheimer's Disease," Frontiers in Computational Neuroscience, 13, 31 (2019).
[38] L. Cowen, T. Ideker, B. J. Raphael R. Sharan, "Network propagation: a universal amplifier of genetic associations," Nature Reviews Genetics, 18, 551–562 (2017).
[39] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, R. Sharan, "Associating genes and protein complexes with disease via network propagation," PLoS Computational Biology, 6, e1000641 (2010).
[40] K. Pearson, "The problem of the random walk," Nature, 72, 342 (1905).
[41] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, 401, 788–791 (1999).
[42] D. Cai, X. He, J. Han and T. S. Huang, "Graph Regularized Nonnegative Matrix Factorization for Data Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 33, 1548-1560, (2011).
[43] D. Cai, X. He, X. Wu and J. Han, "Non-negative Matrix Factorization on Manifold," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 63-72 (2008).

[44] J.P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” Proc. Natl. Acad. Sci. USA, 101, 4164–4169 (2004).
[45] F. Vandin, E. Upfal and BJ. Raphael, “Algorithms for detecting significantly mutated pathways in cancer,” Journal of Computational Biology, 18, 507-22 (2011).
[46] S. Monti, P. Tamayo, J. Mesirov and T. Golub, “Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data,” Machine Learning, 52, 91–118 (2003).
[47] J. K. Huang, T. Jia, D. E. Carlin, T. Ideker, “pyNBS: a Python implementation for network-based stratification of tumor mutations,” Bioinformatics, 34, 2859–2861 (2018).
[48] L. Backstrom and J. Leskovec, “Supervised random walks: predicting and recommending links in social networks,” In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), NY, USA, 635–644 (2011).
[49] H. Hijazi and C. Chan, “A classification framework applied to cancer gene expression profiles,” Journal of Healthcare Engineering, 4, 255-283 (2013).
[50] L. Mateen and G. Hinton, “Visualizing data using t-SNE Visualizing Data using t-SNE,” Journal of Machine Learning Research, 9, 2579- 2605 (2008).
[51] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, 20, 53–65 (1987).
[52] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, 224–27 (1979).
[53] N. Bolshakova and F. Azuaje, “Cluster validation techniques for genome expression data,” Signal Processing, 83, 825-833 (2003).
[54] H. W. Nies Z. Zakaria, M. S. Mohamad, W. H. Chan, N. Zaki, R. O. Sinnott, S. Napis, P. Chamoso, S. Omatu and J.M. Corchado, “A review of computational methods for clustering genes with similar biological functions,” Processes, 7, 550 (2019).
[55] L. Deng, Y. Chen, R. Wu, Q. Wang and Y. Xu, "A Two-Dimensional Clustering Method for High-Speed Railway Trains in China Based on Train Characteristics and Operational Performance," IEEE Access, 8, 81918-81931 (2020).
[56] P. A. Konstantinopoulos, D. Spentzos, and S. A. Cannistra, “Gene-expression profiling in epithelial ovarian cancer,” Nature Clinical Practice Oncology, 5, 577–587 (2008).
[57] P. A. Konstantinopoulos, D. Spentzos, B. Y. Karlan, T. Taniguchi, E. Fountzilas, N. Francoeur, D. A. Levine and S. A. Cannistra, “Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer,” Journal of Clinical Oncology, 28, 3555–3561 (2010).
[58] J. P. Hou, and J. Ma, “Dawnrank: discovering personalized driver genes in cancer,” Genome Medicine, 6, 56 (2014).
[59] F. Hu, Q. Wang, Z. Yang, Z. Zhang, X. Liu, “Network-based identification of biomarkers for colon adenocarcinoma” Research Square, 20, 668 (2020).
[60] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker and D. Lee, “Inferring pathway activity toward precise disease classification,” PLoS Computational Biology, 4, e1000217 (2008).
[61] Y. Yuan, R. S. Savage and F. Markowetz, “Patient-specific data fusion defines prognostic cancer subtypes,” PLoS Computational Biology, 7, e1002227 (2011).
[62] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi and C. Sander, “Integrative subtype discovery in glioblastoma using iCluster,” PLoS One, 7, e35236 (2012).
[63] D. Levine and The Cancer Genome Atlas Research Network, “Genome sequencing centres: Broad Institute. et al. Integrated genomic characterization of endometrial carcinoma,” Nature, 497, 67–73 (2013).

[64] R. G. Verhaak et al., “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,” Cancer Cell, 17, 98–110 (2010).
[65] D. Y. Cho and T.M. Przytycka, “Dissecting cancer heterogeneity with a probabilistic genotype–phenotype model,” Nucleic Acids Research, 41, 8011–8020 (2013).
[66] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, “Similarity network fusion for aggregating data types on a genomic scale,” Nature Methods, 11, 333-337 (2014).
[67] N. K. Speicher and N. Pfeifer, “Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery,” Bioinformatics, 31, i268–i275 (2015).
[68] L. Yang, S. Wang, M. Zhou, X. Chen, W. Jiang, Y. Zuo, and Y. Lv, “Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network,” Scientific Reports, 7, 738 (2017).
[69] P. Chalise and B. L. Fridley, “Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm,” PLoS One, 12, e0176278 (2017).
[70] Z. He, J. Zhang, X. Yuan, Z. Liu, B. Liu, S. Tuo and Y. Liu, “Network based stratification of major cancers by integrating somatic mutation and gene expression data,” PloS One, 12, e0177662 (2017).
[71] D.G. Mun et al., “Proteogenomic characterization of human early-onset gastric cancer,” Cancer Cell, 35, 111-124 e10 (2019).
[72] H. Xu, L. Gao, M. Huang and R. Duan, “A network embedding based method for partial multi-omics integration in cancer subtyping,” Methods, S1046-2023(20)30160-2 (2020).
[73] N. Rohani and C. Eslahchi, “Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach,” Frontiers in Genetics, 11, 553587 (2020).
[74] C. Liu, Z. Han, Z. K. Zhang, R. Nussinov and F. Cheng, “A network-based deep learning methodology for stratification of tumor mutations,” Bioinformatics, 37, 82–88 (2021).
[75] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” Carnegie Mellon University Tech Report (2002).
[76] M. Coskun, B. Bakir-Gungor and M. Koyuturk, “Expanding Label Sets for Graph Convolutional Networks,” arXiv preprint arXiv:1912.09575 (2019).
[77] M. Coskun, A. Grama and M. Koyuturk, “Indexed Fast Network Proximity Querying,” Proceedings of the VLDB Endowment, 11, 840-852 (2018).

CURRICULUM VITAE

20014 – 2018	B.Sc., Computer Engineering, Erciyes University, Kayseri, TURKEY
2018 – Present	M.Sc., Electrical and Computer Engineering, Abdullah Gül University, Kayseri, TURKEY
2019 – 2021	Research Assistant, Computer Engineering, Yozgat Bozok University, Yozgat, TURKEY
2021 – Present	Research Assistant, Computer Engineering, Abdullah Gül University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

C1) P.Guner, I.Sahan, B. Gorkemli, Maden işçilerinin Kaza SırasındaYönlendirilmesi Problemi ve Bu Problem İçin Alternatif Çözüm Yöntemlerinin İncelenmesi in Yöneylem Araştırması ve Endüstri Mühendisliği Kongresi (YA/EM) (June 2018).

C2) P.Guner, B. Bakir-Gungor, Protein-Protein Etkileşim Ağlarında Alt Ağ Arama Yöntemlerinin Performans Değerlendirmeleri in 27. Sinyal İşleme ve Uygulamaları Kurultayı (SIU) (April 2019).

C3) P.Guner, B. Bakir-Gungor, Performance Evaluations of Active Subnetwork Search Methods in Protein-Protein Interaction Networks in 4rd. International Conference on Computer Science and Engineering (September 2019)