VISUALIZATION AND INTEGRATIVE ANALYSIS OF CANCER MULTI-OMICS DATA

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Hao Ding, B.E.

Graduate Program in Computer Science and Engineering

The Ohio State University

2016

Dissertation Committee:

Dr. Raghu Machiraju, Advisor
Dr. Kun Huang, Co-Advisor
Dr. Ewy Mathe
Abstract

Understanding and characterizing cancer heterogeneity not only generates new mechanistic insights but can also lead to personalized treatments for patients. With advances in data generation technologies, ever-increasing amounts and types of multi-omics open great opportunities for researchers to gain extremely valuable information for cancer research and clinical biomarker discovery. However, the vast and complex nature of multi-omics data pose significant challenges regarding the extraction of useful information and the effective integration of multiple types of data.

This dissertation tackles the problem of multi-omics data analysis through both visual analytics and computational angles. First, we present GRAPh based Histology Image Explorer (GRAPHIE), a visual analytics tool designed to explore, annotate, and discover potential relationships in phenomics datasets (histology images). By taking a data-driven approach, we developed an unbiased way to visualize the entire dataset with node-link graphs. The intuitive visualization and rich set of interactive functions allow users to effectively explore the dataset. While (GRAPHIE) focusing on analysing the histological information, we present the second visual analytics tool, integrative Genomic Patient Stratification explorer (iGPSe) which leverages multiple types of molecular features to further characterize patients and tumors. iGPSe is designed to assist researchers in effectively performing integrative multi-omics analysis through interactive visualization components. The tool integrates unsupervised
clustering with graph and parallel sets visualization and allows a direct comparison of clinical outcomes via survival analysis. For both tools, we comprehensively analyzed the design requirements and carried out users’ case studies to demonstrated the usefulness.

Lastly, we developed a computational method that can jointly cluster cancer patient samples based on multi-omics data. The proposed method creates a patient-to-patient similarity graph for each data type as an intermediate representation of each omics data type and merges the graphs through subspace analysis on a Grassmann manifold. We applied our approach to a breast cancer dataset and showed that by integrating gene expression, microRNA, and DNA methylation data, the proposed method would produce potentially clinically useful subtypes of breast cancer.

The proposed visual analytics tools and computational method can be extended to more generalized applications in which exploration and integration of multi-omics data are needed. This dissertation also provides high-level design considerations for visual analytics tools to conceptual methodologies in integrative analysis to future researchers and practitioners for devising effective multi-omics data analysis.
Dedication

This document is dedicated to my family.
Acknowledgments

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

Firstly, I would like to express my deepest gratitude to my advisor, Dr. Raghu Machiraju, for his excellent guidance, caring, patience, and providing me with an extraordinary atmosphere for doing research. His intellectual and philosophical guidance always reminds me to keep thinking deep. He also spent many nights reviewing and revising my papers. I would also like to thank my co-adviser, Dr. Kun Huang. I am extremely lucky to have two highly respected academics as my mentors. I am also very grateful to my committee member Dr. Ewy Mathe for her invaluable comments and suggestion regarding my research.

I will always be thankful for people I work with in my lab: Dr. Chao Wang, Qihang Li, Dr. Zhi Han, Brian Arand, Arunima Srivastava, Travis Johnson and Ross Donatelli.

Most importantly, I am very grateful to my parents for their kindness and endless support. Without their support, I would not complete my education.
Vita

May 25, 1987 ............................... Born - Shanghai, China

2009 ............................. B.E. Software Engineering, Tongji University

2010-present ........................... Graduate Research Associate, The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Computer Science and Engineering
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Background and scope of research</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Thesis statement</td>
<td>10</td>
</tr>
<tr>
<td>1.4 Outline of achievements</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Organization of this dissertation</td>
<td>14</td>
</tr>
<tr>
<td>2. Previous works</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Visual analytics tools for omics data</td>
<td>15</td>
</tr>
<tr>
<td>2.1.1 Networks</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2 Heatmaps</td>
<td>17</td>
</tr>
<tr>
<td>2.1.3 Genome browser</td>
<td>20</td>
</tr>
<tr>
<td>2.2 Integrative analysis</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1 Concatenation-based integration</td>
<td>22</td>
</tr>
<tr>
<td>2.2.2 Transformation-based integration</td>
<td>24</td>
</tr>
<tr>
<td>2.2.3 Model-based integration</td>
<td>25</td>
</tr>
<tr>
<td>2.3 Summary</td>
<td>27</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>3.2</td>
<td>Related work</td>
</tr>
<tr>
<td>3.3</td>
<td>Task analysis and design</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Image representation</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Graph visualization of image collection</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Feature selection</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Interactive user interface</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Image annotation</td>
</tr>
<tr>
<td>3.4</td>
<td>GRAPHIE-Graph based Histology image Explorer</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Back-end image representation generation</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Front-end visualization interface</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Implementation details</td>
</tr>
<tr>
<td>3.5</td>
<td>Results</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Zebrafish retina histology images</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Epithelial-stromal tumor tissues</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Preliminary User Feedback</td>
</tr>
<tr>
<td>3.6</td>
<td>Summary</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>Previous work</td>
</tr>
<tr>
<td>4.3</td>
<td>Methods</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Requirements Analysis and Design</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Visualization Components</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Implementation</td>
</tr>
<tr>
<td>4.4</td>
<td>Results</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Use Case Studies</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Users’ Feedback</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>5.2</td>
<td>Methods</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Method overview</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Data set and Data preprocessing</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Integrative clustering</td>
</tr>
<tr>
<td>5.3</td>
<td>Results</td>
</tr>
</tbody>
</table>
5.3.1 Comparison with results from similarity network fusion (SNF) 100
5.3.2 Identification of integrative subtypes of Breast Cancer . . . 101
5.3.3 Molecular basis of integrative breast cancer subtypes . . . . 104
5.4 Summary ................................................................. 108

6. Conclusion and Future Works ............................................ 109

A. Supplementary Material for Chapter 2 ................................. 113

B. Supplementary Material for Chapter 5 ................................. 115

Bibliography .................................................................. 119
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 The Demographics of TCGA breast cancer subset.</td>
<td>95</td>
</tr>
<tr>
<td>5.2 Comparison of Cox survival p-values from integrative clustering on a grassman manifold with those from SNF</td>
<td>101</td>
</tr>
<tr>
<td>5.3 Clinical attributes of TCGA breast cancer subtypes</td>
<td>104</td>
</tr>
<tr>
<td>A.1 Visual analytics tools for omics data analysis</td>
<td>114</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Multi-omics data types generated from different biological levels including genome, epigenome, transcriptome, proteome, metabolome and phenome</td>
<td>7</td>
</tr>
<tr>
<td>1.2</td>
<td>Single level analysis v.s. integrative analysis</td>
<td>8</td>
</tr>
<tr>
<td>1.3</td>
<td>Outline of achievements: (a) GRAPHIE, (b) iGPSe, (c) Integrative cancer patient stratification</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>Screenshot of Cytoscape 3.4 Desktop application</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>Heatmaps visualization: (a) Gitools interactive heatmap, (b) Circular heatmap</td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>a) Integrative Genomics Viewer b) Circos visualization</td>
<td>21</td>
</tr>
<tr>
<td>2.4</td>
<td>a) Concatenation-based integration involves combining datasets from different data types at the raw or processed data level before modelling and analysis. b) Transformation-based integration involves performing mapping or data transformation of the underlying datasets before analysis, and the modelling approach is applied at the level of transformed matrices. c) Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest. Figure from [1]</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>Subset of histology image collection of zebrafish retinas arranged in a 2-D grid-layout.</td>
<td>30</td>
</tr>
</tbody>
</table>
3.2 Workflow to create image representations: (a) Preprocess each given histology image. (b) Extract feature set. (c) Build a visual codebook through unsupervised clustering. (d) Generate the bag-of-features image representation. 40

3.3 GRAPHIE interface: Graph visualization of the image collection from case study 1. 42

3.4 GRAPHIE interface: Individual image view. 43

3.5 GRAPHIE interface: Subgroups gallery. 44

3.6 GRAPHIE interface: Feature selection view. 45

3.7 Morphology zebrafish retina case study. (a) Initial graph visualization. (b) Example phenotypes of the zebrafish retina: (red) retina of a mutant zebrafish. The mutant possess a small lens and fissure closure defect. (blue) Eye of wild-type zebrafish. Normal retinal lamination and a fused ventral fissure can be seen. (c) Graph created with selected features: the resulting graph shows better separation between wild-type and mutant-type zebrafish eye images, also helps users to spot the mis-annotated images. This figure shows three mis-annotated samples, with the upper two images were mis-annotated as wild-type and lower image being mis-annotated as mutant. 50

3.8 Examples of heterogenous tissue compartments in the histopathological images of breast cancer in our adapted dataset. (a) Epithelial tissue. (b) Stromal tissue. 52

3.9 Epithelial-stromal case study. In graph visualizations, an orange-colored node represents a patch contains stromal; a blue-colored node represent a patch contains mainly epithelium. (a) Initial graph visualization. (b) Graph created with selected features. 54

3.10 (a) User annotation accuracy boxplot. (b) User satisfaction results from the post-study questionnaire. 56

xii
4.1 Screen shot of the interactive analysis page. On the top left, patient samples are clustered based on both mRNA expressions (left panel) and miRNA expressions (right panel). Heatmaps of the clustered data are shown aligned with the parallel sets. On the top right panel, the survival plot shows the patient outcome information and the power of statistical significance (p-value). On the bottom, force directed graph demonstrates the affinity between patient samples. The patient samples selected in the parallel sets are circled. The color of the nodes correspond to the ones in the parallel sets. ......................... 63

4.2 (a)Heatmaps (b)Silhouette Plot. Both (a) and (b) are from the same K-means clustering result(k=4) ................................. 70

4.3 Graph Visualization of the patient population ......................... 73

4.4 Parallel set visualization and survival analysis panels in iGPSe. **Left:** The clusters of patients using different feature sets are shown as bars. On the left side, the blue/yellow/purple/pink bars indicate the four patient groups separated using the PAM50 breast cancer genes [2]. The gray/green/light blue bars indicate three groups separated using the above discussed miRNAs. The gray bands linking matched patients. **Right:** The Kaplan-Meier curves of the survival times for the two groups of selected patients. The colors of lines matches the colors of the selected bands. In addition, the p-value of the difference in survival times between the two groups based on log-rank test is also listed (p-value = 3.8173e-06 in this example). ................................. 75

4.5 Feature selection and input panel in iGPSe. **Left:** The user can select list of genes (for mRNA data) from the left list (extracted from input files) by left clicking on the genes or input the gene list (e.g., copy and paste) into the window on the right side. The user can also load gene/miRNA list from text files. The selected genes will also show up in this window. **Right:** The selection and input for miRNA is similar to that for genes. ................................. 86

4.6 The clustering analysis page. ........................................... 87
4.7 Comparisons of different choices of subgroups. (a) Compare the blue and yellow/purple/pink groups based only on mRNA data. (b) Compare the grey/green and light blue groups based only on miRNA data. (c) Two groups with different mRNA profiles and different miRNA profiles. (d) Two groups with different mRNA profiles but similar miRNA profiles.

5.1 Workflow for integrating and merging cancer genomics datasets. A patient-to-patient similarity graph is constructed for each data type. The similarity matrices are converted to subspaces and embedded in a Grassman manifold, where they are integrated into a single, representative subspace. This subspace is then clustered to obtain the final integrative patient groups.

5.2 Survival plots of integrative clusters for BIC (a), KRCCC (b), LSCC (c), and GBM (d). P-values are computed from the log rank test.

5.3 Integrative clustering of breast tumors produces prognostically relevant and biologically significant groupings. (a-c) The adjacency matrices of patient-to-patient similarity graphs, produced from mRNA (a), microRNA (b), and methylation (c) datasets. (d) Integrative clustering of patients using all three datasets. Color bars at left show the clusters of patients, and the heatmap to the left of each colorbar shows the eigenvector clustering results. (e-h) Survival analysis of patient stratification results using integrative and single-data-type clustering methods. Kaplan-Meier survival curves of clusters produced by: (e) integrative clustering, (f) Gene expression alone, (g) miRNA expression alone, (e) DNA methylation alone, listed along with estimated p-values (log-rank test).

5.4 Density of differentially expressed genes on chromosome 19p13 are shown in (a). (b-e) Survival curves and clustered heatmaps produced by separating four datasets based on expression of genes on chromosome 19p13. Kaplan-Meier survival curves, clustered heatmaps, and accompanying p-values (log-rank test) generated from the training dataset from TCGA are shown in (b). Results generated from three validation sets are shown in (c-e): Netherlands Cancer Institute (NKI) in (c), GSE3143 in (d), and GSE1456 in (e).

B.1 Box plot of Gene expression level for MYBL2, CENPA, AURKB and KIF2C in the integrative clusters.
B.2  Box plot of copy number variations of chromosome 19p13 in the integrative clusters ........................................... 116

B.3  Plot of number of subtypes versus silhouette score values ........... 117

B.4  ROC curves on integrative subspace, mRNA subspace, miRNA subspace and DNA methylation subspace ......................... 118
Chapter 1: Introduction

Approximately 40 percent of the US population will be diagnosed with cancer at some point during their lifetime [3]. In just the year 2011, more than 575,000 people died of cancer, and more than 1.5 million people were diagnosed with cancer. The burden of cancer is reflected on not only the life lost but also in the form of economic cost to society. According to the National Institutes of Health (NIH), cancers cost the United States an estimated $263.8 billion in medical costs and lost productivity in 2010 alone [4].

Consequently, a tremendous amount of effort has been directed towards understanding disease mechanisms and developing novel treatments toward cancer. New technologies and platforms generate ever-increasing amounts and types of biomedical data on a daily basis. While the broad variety of data offers opportunities for researchers to enhance their understanding of diseases and accelerate treatment development, it also poses significant challenges to the extraction of the useful information. This dissertation makes a series of attempts to tackle the problem of multi-omics data analysis. The proposed visual analytics tools and applications can be used to explore high-dimensional datasets, identify integrative subtypes of cancers, and generate new biomedical hypotheses.
In this chapter, we first elaborate upon the motivation of this work (Section 1.1), followed by a general introduction to the research areas and necessary background related to this dissertation (Section 1.2). After framing the thesis statement, we describe this dissertation’s achievements at a higher level and summarize our contributions (Section 1.4). Finally, we present an overview of the structure of this dissertation (Section 1.5).

1.1 Motivation

Cancer is not a single disease, rather it is a spectrum of diseases. Many types of cancer are highly heterogeneous with different subtypes. These subtypes often possess distinct clinical outcomes, such as the response to a particular treatment or the likelihood of recurrence or metastasis. Therefore, precisely identifying these subtypes not only assists clinicians to better treat their patients, but also scientists to better understand tumor heterogeneity.

High-throughput omics technologies have flourished over the course of the last decade. In fact, since 2008 the growth of published genomics data has been outpacing Moore’s Law by a factor of 4 [5]. This increasing volume of diverse omics data has opened new opportunities for researchers to gain a better understanding of cancer’s biological mechanisms from multiple perspectives. This may lead to conceptual developments and discoveries of the new subtypes. For instance, a landmark study by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) analyzed approximately 2000 tumors and proposed a genomics-driven classification of breast cancer based on an integrative analysis of gene expression and genome-wide copy number alterations (CNAs) [6, 7]. However, the immensity and heterogeneity
of such omics data also bring with it new problems. One of the major concerns that frequently has been voiced is the growing disparity between the fast pace of data generation and the relative lack of in-depth analysis of those data. Kenneth Weiss, professor of biology at Pennsylvania State University, characterized undertaking massive omics data generation efforts as “easier than to think critically and ask deeper questions” [8].

What makes in-depth analyses of high-throughput omics data so hard? In general, there exist three fold challenges: (i) First, the size and dimensionality of the omics data make it very difficult to gain clear biological insight. For anyone to carry out the analysis at the genome scale, the method being used needs to be able to handle not only the volume of the data, but also separate useful signals from thousands of noisy measurements. (ii) Second, omics data analysis, which frequently comprises explorations or hypothesis-generating stages in biological research, requires both analytical skills and biological domain knowledge. Currently, most bioinformaticians engaged in the analysis of omics data are either “trained computer scientists or statisticians devoted to biology”, or “trained biologists that were required to learn the basics of programming to dig deeper into their data” [9]. The knowledge gap between disparate fields significantly slows down research progress. (iii) Last but not the least, the heterogeneity of existing data types and the multitude of competing formats further complicates the problem. It is well-accepted that integrating multi-omics data could help researchers study cancer comprehensively; however, integrating heterogeneous and large omics data constitutes not only a conceptual challenge but a practical hurdle as well.
Given the increased need for analyses of large multi-omics datasets, this dissertation aims to develop tools and applications to overcome the aforementioned challenges. Specifically, this dissertation achieved the following three aims:

**Aim. 1** Develop user-friendly tools that target efficient navigation of large omics datasets and identification of outcome-related features (*GRAPHIE*).

**Aim. 2** Develop tools that allow a user to quickly carry out integrative analysis on multi-omics data and identify potential combined markers (*iGPsE*).

**Aim. 3** Develop efficient computational methods to perform integrative analysis on multiple datasets.

In our effort to approach **Aim 1** and **Aim 2**, we presented two visual analytics tools: *GRAPHIE* and *iGPsE* to explore large histological image datasets and perform integrative patient stratification, respectively. To address **Aim 3**, we developed an integrative clustering application to identify cancer subtypes. While this dissertation mainly focused on breast cancer data to showcase the ability and efficacy of the proposed tools and application, our approaches are generalizable to many diseases.

### 1.2 Background and scope of research

**Cancer Heterogeneity**

Cancer is a heterogeneous group of diseases, each subtype with its own intricate genetic diversity and different morphological phenotypes. This heterogeneity can be observed both within a tumor (intra-tumor heterogeneity) and across patients (inter-tumor heterogeneity). This dissertation primarily focuses on inter-tumor heterogeneity, which is defined by variation presented between patients with the same cancer
type [10]. Specifically, tumors within the same type of cancer may belong to different subtypes, leading to differences in prognosis and therapeutic sensitivities. This introduces significant challenges in designing effective treatment strategies to optimize individual outcomes. Research aimed at characterizing tumor heterogeneity not only generates new mechanistic insights but can also lead to personalized treatments for patients.

Historically, clinicians and scientists have been using a variety of clinical and pathological factors to categorize patients into different subtypes. Take breast cancer for example; the common factors include axillary lymph node status, tumor size, histological grade, hormone receptor (ER/PR) status, and HER2 status [11, 12, 13]. While these factors have been useful for assessing prognosis and recurrence rate to an extent, they are not conclusive. Because of tumor heterogeneity, patients with similar pathological and clinical profiles may still have very different outcomes [14].

Due to advances in high-throughput sequencing technology, studies using molecular data to categorize patients have become highly prevalent. Molecular subtypes with distinct clinical outcomes and response to treatment have been found in almost all major cancer types [15, 16, 17, 18]. In breast cancer, studies using gene expression profiling have identified five major breast cancer subtypes beyond the traditional hormone receptor subtypes. Luminal A and luminal B groups are identified among the hormone receptor-positive cancers. The HER2 and basal-like groups are the major molecular subtypes identified among hormone receptor-negative cancers. Other molecular subtypes such as normal breast-like groups have also been identified in some studies. These breast cancer molecular subtypes differ with regard to their patterns of gene expression, clinical features, response to treatment, and prognosis.
It has become clear that cancer heterogeneity exists on all biological levels: genetic, epigenetic, genomic, proteomic, morphological, and clinical. Integrative analyses on multiple biological levels could more precisely categorize patients and eventually will greatly increase the accuracy of diagnosis, prognosis, and prediction of cancer. However, comprehensive research on integrative data analysis is still lacking. The following challenge remains: how to integrate multi-omics data effectively into integrative computational models that can be used to gain new insight into cancer biology, discover new prognosis biomarkers, and predict responses to treatments.

**Multi-omics data**

Multi-omics data generated from multiple levels of biological systems measures different aspects of biological mechanisms. Figure. 1.1 provides a schematic diagram of various biological levels and lists example data types which have been widely used by the research community. Multi-omics approaches aim at understanding interrelation and the functioning of different levels of biological systems, thereby developing more effective treatments for cancer patients.

With advances in data generation technologies, ever-increasing amounts and type of multi-omics data are being produced daily. Large-scale projects like The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) systematically profile a large number of patients samples from almost all common cancer types resulting in public available multi-omics databases [19, 20]. The rise of multi-omics data resources has made cancer one of the most well-characterized diseases at the molecular level. Consequently, it pushes the boundaries of personalized medicine [21].
Many efforts have been made to analyze multi-omics data and decipher the molecular heterogeneity of cancer. However, there is still a need for powerful and advanced analysis strategies to fully harness the utility of these comprehensive high-throughput data to identify true associations and reduce the number of false associations.

**Integrative analysis of multi-omic Data**
Integrative analysis is defined as the use of multiple sources data to provide a better understanding of a system/situation/association/etc. Although studies based on single-source omics data are still useful, much of the etiology of complex traits remains unexplained. Researchers have acknowledged that biological systems cannot be fully understood by the analysis of single-type datasets as the regulation of the system certainly occurs at many levels [22, 23, 24].
To gain a deeper understanding of the biology of cancer, we have to answer fundamental questions about the molecular behaviors and interactions between/within different biological levels. In recent years, with the multi-omics datasets becoming abundant, increasing efforts have been made to develop computational models and applications to integrate data from different levels [25]. For instance, [26] combined both different progression levels, as well as multiple levels of molecular data (mRNA, copy number alterations, microRNAs, and methylation), clinical data and pathway information to define groups of patients with distinct biological signatures and differing prognoses.

In foreseeable future, integrative analyze will play a more crucial role in computational biology. However, the challenges of developing effective integrative methods cannot be overlooked. The most fundamental challenge is the dimensionality of the multi-omics data, making it susceptible to overfitting. While integrating multiple types of data allows researchers to obtain more information about a biological system, the dimensionality of the input data is increased tremendously. This could potentially weaken the ability of the analysis to infer biological knowledge.

**Visual analytics**

In this dissertation, we adopted visual analytics as one of the approaches to conduct the exploration and integrative analysis of multi-omics data. Visual analytics is defined as the science of analytical reasoning supported by interactive visual interfaces [27]. Originating from information visualization and automatic data analysis, visual analytics quickly became an active field of research due to its ability to process large and complex datasets. Visual analytics integrates the interactive visual
representation, and analytical processes and more focuses on the analytical reasoning. Visual analytics approaches offer various features which make it an intriguing approach to the multi-omics data analysis task:

1. Users’ domain knowledge, flexibility, and creativity can be actively engaged in the analyzing process through interacting with the visual interface [28].

2. Visual representation empowers human perceptual capabilities which makes insights discovery within massive dataset much more efficient [29].

3. The integration of analytical models or algorithms also allows visual analytics systems to conduct complex analysis.

4. The interactivity makes visual analytics more capable of suggesting hypotheses, assessing assumptions, and collecting insights for the basis of further investigation [30].

This dissertation presents two visual analytics applications designed for exploring omics data and identifying integrative subtypes. Our work makes a stronger case for the advantage of using visual analytics approach for omics data analysis.

1.3 Thesis statement

In-depth analysis of multi-omics data can reveal novel biological insights which will lead to more effective personalized treatment development. The use of intuitive visual analytics tools and computational applications which use omics data from multiple biological levels can overcome the existing challenges in the multi-omics analysis, such as high volume, variety, and heterogeneity. The proposed tools and case studies will suggest new potential biomarkers and subtypes of cancer.
1.4 Outline of achievements

To achieve the aforementioned aims, this dissertation presents the following three topics (Figure 1.3):

- **GRAPHIE**: **GRAPh based Histology Image Explorer**:

  We developed GRAPHIE, a visual analytics tool, to explore, annotate and discover potential relationships in histology image collections within a biologically relevant context. By representing each image with informative features and then subsequently visualizing the image collection with a node-link graph, GRAPHIE allows users to effectively explore the image collection. Users can perform feature selection in an interactive way to improve the visualization of the image collection and the overall annotation process. We
demonstrated the usefulness of our visual analytics approach through two case studies.

- **iGPSe: integrative Genomic Patient Stratification explorer:**

  *iGPSe* is a visual analytic system designed to significantly reduce the computing burden for biomedical researchers in the process of exploring complicated integrative genomics data. *iGPSe* integrates unsupervised clustering with graph and parallel sets visualization and allows a direct comparison of clinical outcomes via survival analysis. We designed a case study using a breast cancer dataset obtained from the The Cancer Genome Atlas (TCGA) project. Using *iGPSe*, we are able to quickly explore different combinations of gene expression and microRNA features and identify potential integrative markers for survival prediction.

- **Integrative cancer patient stratification via subspace merging:**

  We presented an integrative clustering application which creates a patient-to-patient similarity graph for each data type as an intermediate representation of each omics data type and merges the graphs through subspace analysis on a Grassmann manifold. We hypothesize that this approach generates more informative clusters by preserving the complementary information from each level of 'omics data. We applied our approach to a TCGA breast cancer data set and showed that by integrating gene expression, microRNA, and DNA methylation data, our proposed method would produce clinically useful subtypes of breast cancer.
In \textit{GRAPHIE} and \textit{iGPSe}, we developed user-friendly interfaces which allow biomedical researchers to explore multi-omics data without knowledge of programming. Both proposed tools also support interactive feature selection which not only reduces the volume and dimension of the omics data but allows users to inject their domain knowledge into the analysis process. Most importantly, \textit{GRAPHIE} and \textit{iGPSe} could both help researchers forming new hypotheses in an interactive fashion. This is essential during analyzing process, as in biomedical research it is common for researchers to use multi-omics data as a hypothesis generation tool. Without a clear hypothesis and a prior knowledge on which models or tools to use, researchers could leverage the proposed visual analytics tools to explore their data visually and define hypothesis and molecular features of interest for further validation. During the development our tools, we comprehensively analyzed the design requirements and carried out users case studies. These experiences can lay the foundation for the developing of future multi-omics analysis tools.

This dissertation also explores different integration methodologies through \textit{iGPSe} and integrative clustering application. The \textit{iGPSe} adopts a model-based integration strategy to integrate clustering results from various types of data through interactive visual components. The tool grant users sufficient control over the integration process, including feature selection, model selection, sub-group selection and comparison.

In our proposed integrative clustering application, a transformationed-based integration method is used to incorporate multiple types of omics data for obtaining better prognostic subgroups of patients. Overcoming the limitations of scale and dimensionality present in most clustering method, our approach uses a subspace representation to preserve the data-type-specific properties and maximize the signal-to-noise ratio.
Compared to previous approaches, our method doesn’t require iterative optimization which leads to a more stable and computationally efficient solution. We also showed that by integrating gene expression, microRNA, and DNA methylation data, our method identified a large cluster of genes located on chromosome 19p13 has a strong correlation with prognosis. We further validated these findings on three additional independent breast cancer datasets.

1.5 Organization of this dissertation

The rest of this dissertation is organized as follows. In the next chapter, we review previous work in multi-omics data analysis. Chapter 3 and 4 describe GRAPHIE and iGPSe respectively. For each chapter, we detail the design rationales of the visual analytics tools followed by use case studies to illustrate the efficacy of the proposed tools from a target user’s perspective. Chapter 5 presents an integrative clustering application. Lastly, we conclude this dissertation with a summary and future research directions in Chapter 6.
Chapter 2: Previous works

Over the last two decades, the emerging fields of computational biology and bioinformatics have led to significant advances in cancer multi-omics data analysis. This chapter presents a review of existing visual analytics tools and methods deployed on the multi-omics data analysis. It opens with a discussion of common visualization techniques for omics data along with the reviews of popular visual analytics tools in Section 2.1. Subsequently, Section 2.2 summarizes the recent approaches to the integrative analysis of multi-omics data. Finally the chapter ends with a brief summary in Section 2.3.

2.1 Visual analytics tools for omics data

Parallel to advances in omics data generation technologies, visual analytics tools are continuously being developed. Recent studies have shown that the use of intuitive visual analytics tools can bring significant benefits to the field of cancer genomics. For instance, [31] demonstrated that by using Caleydo, a popular visual analytics platform for cancer genomics analysis, researchers can efficiently discover genes that are distinctive for specific cancer subtypes. In [32], Cytoscape, a network visual analytics software, was used to perform the analysis on genetic interaction and gene regulatory events. These results motivate researchers to keep creating intuitive, meaningful and
integrated visual analytics tools that can facilitate biological discovery in the face of complex and high volume omics data that would, otherwise, overwhelm. Here, we organized our review by discussing three effective visualization techniques which are commonly adopted by existing tools. These approaches complement each other, and each is best suited to different specific tasks.

2.1.1 Networks

Networks are routinely used in biology to capture and model the complexity of relationships between different entities such as genes, proteins, or patients [33, 34, 35]. The key feature of network visualization is that it highlights functional relationships in large unstructured datasets and build bridges between the computation analysis and biological insights [36].

Numerous solutions have been recently developed for the visualization as well as the analysis of complex biological networks of multi-omics datasets [37](Table. A.1). Cytoscape [38] is one of the most widely used network visualization and analysis tools in genomics research. The user-interface of the tool is intuitive and requires no special bioinformatics knowledge for users (Figure. 2.1). The properties of nodes, edges and network layout are customizable. Moreover, Cytoscape provides a wide selection of plugins allowing users to achieve specific tasks in their exploration. A web version, Cytoscape-web [39], is compatible with common internet browsers and facilitates interaction with the networks displayed. One popular use case of Cytoscape-web is the eBio Cancer Genomics Portal [40], which uses the web plugin to visualize given coexpression networks. In this dissertation, we adopted network visualization in both GRAPHIE and IGPSe to provide an overview of the dataset in high dimensional
feature space. In our tools, the network visualizations are fully interactive at both construction stage and exploration stage, which allows users to actively engaged in the analyzing process.

One of the main challenges of network visualization in bioinformatics is choosing the appropriate layout to provide meaningful visualization for the increasingly larger and more complex networks. The traditional force-directed based layout algorithms, although wildly used, do not perform well for large graphs [41]. To address this challenge, novel layout algorithms continue to be developed and refined to make network visualizations more approachable and informative [42]. In a recent work, Krzywinski et al. [43] presented the hive-plot, a novel network layout technique, to generate informative and reproducible network visualizations. Hive-plot places nodes on radially oriented linear axes according to nodes’ connectivity and uses curved edges to connect nodes from different axis. This layout focuses on the densely connected patterns existing in networks which often leads to the discovery of potential biological insight. For instance, [44] used hive-plot to summarize proximity among patients and identified the correlation between immunologic mutations and patient survival status.

2.1.2 Heatmaps

Heatmaps are the most commonly used graphical representations of omics data stored in the form of matrices [45]. Most of the existing tools support the generation of the heatmap visualization (Table. A.1). The columns in a heatmap usually correspond to samples (e.g. patients or tumor), whereas the rows correspond to features (e.g. genes, transcripts, or other genomic elements). The color of each cell indicates its expression level or mutational status.
A key operation on heatmaps is the reordering of the columns and rows. Typically the reordering is done using hierarchical clustering, and a dendrogram showing the hierarchy is usually arranged next to the heatmap. This ensures that similar samples are placed near each other, which makes it easier to spot the salient and important functional relationships within the data matrix. Some tools support ordering the columns based on the clinical attributes of samples to make the comparison between different groups [46, 47].

Gitools [47] is one of the most popular applications designed for the analysis and visualization of omics data using interactive heatmaps (Figure 2.2a)). The key feature of Gitools’ heatmaps visualization is that each cell can contain multiple elements which make it well suited to the exploration of multi-omics data. In addition, Gitools provides interactive functions to allow users to manipulate the heatmap visualization,
such as zoom in/out, reorder the columns with custom labels. Since the tool also supports exporting heatmaps to high-resolution images, many researchers have been using Gitools to produce figures in their publications [48, 49, 50]. Heatmaps play an important role in IGPSε to help users to examine the feature patterns in clusters and identify the potential cancer subtypes related features.

A general limitation of the heatmap visualization is that relationships between features are difficult to grasp. Particularly, the locations of genes on the chromosomes are ignored in the heatmap visualization, as the orders of columns and rows only depend on the similarity between entries. Gitools tries to solve this problem by offering the possibility of adding genomic annotations to the rows that can encode functional or structural information. Caleydo StratomeX [51] solves this problem by incorporating pathway diagrams displaying functional relationships between the genes, and CircleMap [52] plots can also be used as nodes to construct a network diagram for this purpose.

Figure 2.2: Heatmaps visualization: (a) Gitools interactive heatmap, (b) Circular heatmap

19
2.1.3 Genome browser

The typical genome browser aligns data to their genomic loci (Figure. 2.3 a)) and shows the annotations or alterations at corresponding locations. This allows users to visually inspect genomics topography of alterations in their data. Many applications also provide interactive functions which enable users to investigate a particular genomic locus in detail.

Integrative Genomics Viewer (IGV) [53], the UCSC Cancer Genomics Browser [54], and the Savant Genome Browser [55] are three of the most popular tools that use genome browser to visualize molecular features such as chromatin accessibility, gene expression, and genomic alterations, genome-wide. All three tools share similar features on their genomic coordinates visualization: (i) support multiple data formats that are used to represent various types of genomic alterations, (ii) display the alterations in each sample as genomic tracks, (iii) provide convenient zooming and scrolling interaction for users to navigate particular genomic regions.

Circos [56] is another widely adopted tool that provides a circular layout to visualize the genomic coordinates which enables genomic coordinates of all chromosomes to be represented in the same visualization. This tool aptly illustrates relationships between distinct alterations, represented as data tracks outside the ideogram, which take place at different locations within the genome. These relationships between regions are normally depicted as ribbons (Figure. 2.3 b)). Intra- and inter-chromosomal translocation (cause of many diseases such as cancer, infertility, and Down syndrome [57, 58, 59]) can be particularly well represented in Circos.

The biggest limitation of genomic coordinates tools is the lack of capacity to display relationships between genomic features that are independent of location, such
as the coordinated expression of genes. The IGV addresses this limitation through using the split-screen, which allows multiple loci to be displayed next to each other.

Figure 2.3: a) Integrative Genomics Viewer b) Circos visualization

2.2 Integrative analysis

Despite the advantages of visual analytics approaches on analyzing the cancer multi-omics data, more sophisticated methods are necessary to discover novel biological insights. In recent years, there have been multiple efforts on essential methodologies in the integrative analysis of multi-omics datasets. This section focuses on methods that combine multiple omics data types in a simultaneous analysis to reveal biological models. In general, the existing methods can be categorized into three classes: concatenation-based integration, transformation-based integration, and model-based integration (Figure. 2.4).
Figure 2.4: a) Concatenation-based integration involves combining datasets from different data types at the raw or processed data level before modelling and analysis. b) Transformation-based integration involves performing mapping or data transformation of the underlying datasets before analysis, and the modelling approach is applied at the level of transformed matrices. c) Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest. Figure from [1]

### 2.2.1 Concatenation-based integration

Concatenation-based integration approaches straightforwardly concatenate multiple types of omics data and then carry out the analysis with the combined matrix. One of the main advantages of these approaches is that many existing single omics analysis methods can be directly applied to the combined matrix if combing features
appropriately. For instance, [60] and [61] map identifiers from different platform to a common set of identifiers to generate a single concatenated matrix for subsequent analysis. Fridley et al. [62] concatenate SNPs and mRNA gene expression into a single data matrix, and the joint relationship of mRNA gene expression and SNP genotypes was then modeled using a Bayesian integrative model to predict a quantitative phenotype. Mankoo et al. [63] predicted patient survival status of ovarian cancer patients through applying a multivariate Cox LASSO model to the matrix which is concatenated with copy number alteration (CNV), DNA methylation, miRNA and gene expression data. Another advantage of the concatenation-based integration approach is that it accounts for interactions between different types of genomic data. For example, microRNA-mRNA interactions were reported as involved in tumor development [64, 65]. Cascione et al. [66] identified four molecular subclasses in triple negative breast cancer through clustering gene expression and microRNA concatenated data matrix.

Despite advantages of concatenation-based integration mentioned above, it is challenging to identify an appropriate approach for combining multiple types of omics data [1]. Since platforms are not universal, different types of omics data have different scales or belong to different data categories (discrete or continuous). Simply combining these data may cause bias in the model. Additionally, the concatenation of these high-dimensional matrices will largely inflate the dimension of the model, which could introduce noises and increase computational burden [67]. Therefore, the concatenation-based approach is suitable when the appropriate way to aggregate the multi-omics data is determined and the dimension of data is moderate.
2.2.2 Transformation-based integration

Transformation-based integration approaches often consist of two main steps: (i) transforming each data type into an intermediate form, such as a graph or a kernel matrix and (ii) merging the intermediate forms into an integrative representation. Compared to concatenation-based approaches, transformation-based integration is more robust to the different data measurement scales, as the integration is conducted in a unified space. This approach can be used to integrate many types of data, including continuous or categorical values and sequence data. Moreover, the intermediate representation has the advantage of preserving data-type-specific properties from each dataset when each type of data is transformed into an appropriate intermediate representation.

Lanckriet et al. [67] introduced a computational framework for genomic data fusion, including amino acid sequences, in which each type of data is represented by a kernel function that defines similarities between pairs of entities, such as genes or proteins. Speicher et al. [68] applied multiple kernel learning for data integration and subsequently perform cancer subtype identification. A multiple kernel learning for dimensionality reduction (MKL-DR) framework was used to perform dimension reduction and data integration at the same time. By contrast, Kim et al. [69] proposed a graph-based integration framework for the predicting cancer clinical outcomes in brain cancer and ovarian cancer by integrating copy number alteration, methylation, microRNA and gene expression data. In this work, a graph-based semi-supervised learning technique was used to classify samples. Integration of multi-level genomic data sources was achieved by finding an optimum value of the linear combination coefficient for the individual graphs derived from each type of data. In a more recent
work, Wang et al. [70] introduced a network-based approach, Similarity Network Fusion (SNF), for aggregating data types on a genomic scale. The integration is done by constructing a patient similarity network for each available omics data and fusing all networks into a single similarity network with a nonlinear combination method to represent the full spectrum of underlying data. The approach has been applied to combine 3 omics data, i.e. gene expression, DNA methylation, and microRNA expression for five cancer datasets. The results indicate that SNF outperforms single data type analysis and established integrative approaches.

However, transformation-based integration is not without its disadvantages. First, the interpretability of transformation-based integration is weak. Since different types of data are transformed and merged in the same feature space, it is difficult to recognize each data type’s contribution of to the final results. Moreover, the interactions between different types of data can be difficult if the separate transformation of the original feature space changes the ability to detect the interaction effect. Therefore, transformation-based integration is suitable if there is a relevant intermediate representation, such as a kernel or graph, for each omics data type, and the goal is to preserve data-type-specific properties while integrating them.

2.2.3 Model-based integration

Model-based integration approaches first separately analyze each data type, subsequently combining the results. The main advantage of model-based integration approaches is the flexibility as different models can be applied to different data types. The model-based integration approaches have been extensively used in the field of
bioinformatics. Based on the methods applied on the individual data type, model-based integration approaches can be categorized as supervised and unsupervised.

Under the supervised setting, multiple models are generated using the different types of data as training sets and then combine these model through bagging or voting. For instance, in [71] a set of $k$-nearest neighbor classifiers were trained on multiple proteomics datasets collected from different platforms. Then weighted voting was used in classifiers combination for protein fold recognition. In [72], a majority voting approach was taken for prediction of drug resistance of HIV protease mutants. [73] proposed a bagging approach to ensemble multiple classifiers and used the integrative model for predicting tumor classes.

Under the unsupervised setting, the integration is conducted by aggregating clustering results from different data types into a set of “consensus clusters based on certain optimization criteria. Strehl et al. [74] seek a consensus clustering by maximizing the mutual information. Topchy et al. [75, 76] consider a representation of multiple clusterings as a set of new attributes characterizing the data distributions, and then a mixture model (MM) offers a probabilistic model of consensus using a finite mixture of multinomial distributions in the space of base clusterings. In a more recent work [77], an integrative clustering algorithm called Regularized Patient Stratification (RPS) is developed, which utilizes the clusters from molecular features to regularize the clustering from clinical classifications. The key feature of RPS is that it allows one to integrate numerical and categorical data. For example, it is possible to integrate gene expression levels (numerical) with clinical staging information (categorical).
2.3 Summary

This chapter reviews the previous work in visual analytics and integrative analysis of cancer multi-omics data. We listed the advantages and limitations of three widely used visual techniques and three integrative analysis approaches. Furthermore, multiple previously proposed methods and tools are reviewed. In the next two chapters, we present two visual analytics tools and carefully discuss the choices of visual techniques adopted in our tools. Beyond the visual analytics approaches, Chapter 5 presents a novel transformation based integrative method to leverage multiple types of omics data.
Chapter 3: GRAPHIE: Graph based histology image explorer

In this chapter\(^1\), we present a visual analytic tool designed to explore the heterogeneity within a single type omics data. Specifically, we focus on exploring the heterogeneity of morphological phenotypes that manifest in histology images. The design of the graph based histology image explorer (GRAPHIE) is driven by a real-world problem brought up by our pathologist collaborator: Can we develop an application to assist researchers in efficiently and accurately annotating large histology image collections?

We took a visual analytics approach to tackle this problem. By describing each image’s morphological properties with local texture features, GRAPHIE subsequently visualizes the image collection in a semantic fashion. Its interactive capabilities allow the user to navigate, annotate and discover potential relationships in histology image collections within a biologically relevant context. We demonstrated the usefulness of our visual analytics approach through two case studies. Both of the cases showed efficient annotation and analysis of histology image collection.

The chapter is structured as follows. Section 3.1 provides background and motivations. Section 3.2 describes related work. Section 3.3 describes the overall design

\(^1\text{This chapter is published in BMC Bioinformatics 2015 16(Suppl 11):S10}\)
rationale for our work motivated by the need for large histology image collection studies, while we describe our workflow and methods in Section 3.4. Comprehensive case studies are reported in Section 3.5. We close the chapter by summarizing the contributions of our work and point to the future.

3.1 Introduction

Large-scale phenotyping studies have recently gained considerable attention in biomedicine. Phenotypes occur in various forms and at different levels as enumerated in [78]. Histology images offer an important source of knowledge for phenotyping studies at the cellular and tissue levels. Pathologists have been traditionally cataloging and studying the morphology of cellular phenotypes arising from genomic alterations and adverse conditions [79].

Since the morphological phenotypes that manifest in histology images are highly heterogeneous and variable, the analyses and annotation of these images require well-trained experts (e.g. pathologists) and rely on specific abstract patterns that they glean through extensive experience. Still, there is much variability in annotations obtained from various experts, especially given the uncertainty that exists in the underlying biological mechanisms. With the development of high-content and high-throughput acquisition methods, there is an ever increasing availability of high-resolution digital histology images [80], requiring the use of effective tools to explore and annotate the entire dataset efficiently. It is, therefore, necessary to gain a global overview of the morphology while being able to discern differences between images from individual samples.
Consider the zebrafish histology image data subset presented in Figure 3.1. In biomedical research, the use of animals such as zebrafish as a genetic model is useful for understanding the basic mechanisms that underlie human disease [81]. The heterogeneity expressed by the morphology in the eye of the zebrafish also makes the eye a perfect object for phenotypical study. Like most classes of extant vertebrates, the retina of a zebrafish is composed of seven major cell types derived from the neural ectoderm, six types of neurons and one type of glial cell [82]. In Figure 3.1, each image depicts the retina of a zebrafish of a specific genotype, which is the result of systematic gene knock-downs. A typical phenotypical study involves differential comparison of these images to identify common and distinctive characteristics across
the genotypes and to identify the various subtypes in the collection [79]. A common approach towards the study and the annotation of images is realized by placing them side-by-side in a 2-D grid-layout (Figure. 3.1). However, as the sample size increases, it becomes too difficult to discern differences with the naked eye. In addition, a 2-D grid-layout provides no information about how the images are similar or different, making it very difficult to learn the correlations across the image collection and to further annotate large repositories of data.

We believe that an interactive visualization tool that allows for the tangible organization of image collections can lead to a better overall understanding of the studied datasets as well as facilitate the image annotation processes. We present a visual analytics tool, GRAPh based Histology Image Explorer or GRAPHIE, designed to assist pathologists and systems biology researchers to explore, annotate and reveal potential relationships of phenotypical properties within a biologically relevant context. GRAPHIE employs the bag-of-features (BoFs) approach [83] to capture visual patterns from a given collection of histology images, thus allowing a semantic organization of unstructured image collections. By further using a proximity graph and a flexible graph layout, GRAPHIE provides a visual representation of the entire image collection, which in turn enables the intuitive exploration of the underlying structure of the dataset as well as the capability of drilling down to subgroups of interest. Our tool provides a rich set of interactive functions which make exploring and annotating on the graph more efficient and consistent. The tool also allows interactive feature selection, permitting users to examine subsets of image features and iteratively refine the eventual graph visualization.

In summary, the main strengths of GRAPHIE include:
1. An interactive visual analytics tool to efficiently explore and annotate histology image datasets.

2. A framework which supports interactive comparison of subgroups and selection of features in histology images.

We conducted two case studies to evaluate GRAPHIE. In the first study, an histological images dataset of 168 mutant specimens of the model organism *Danio rerio*, commonly known as the zebrafish, were acquired. The mutant fish were the results of direct genetic alterations. GRAPHIE offers users a flexible way to explore distinctive morphological features and thus glean the structural changes wrought by genetic alterations. The second case study is concerned with histology images from a human breast cancer study. The task here is to explore image collections and annotate regions of the images if they predominantly contain specific tissue types (e.g., epithelium or stromal).

### 3.2 Related work

Efforts have been expended to employ visual analytics approaches to explore image datasets. For instance, [84] visualizes an image collection using a multidimensional scaling layout based on semantic similarities between images. In [85] a method called clustered album thumbnails (CAT) was presented towards the hierarchical browsing of large image collections allowing users to interactively explore different levels of details. Both 2D grid and spiral layouts were used in [86] to present search results of images resembling an example image. While these methods support data exploration, they lack the ability to interactively assist users to spot differences among images,
let alone glean subtle patterns in the more content-rich and complex histology image collection.

On the other hand, since manual annotation and analyzes of histology images requires well-honed expertise, the study of quantitative image-based assessment has attracted attention [87]. There is a considerable amount of work leveraging advances in the semantic content analysis to study histology images [88, 80]. The SHIRAZ (System of Histological Image Retrieval and Annotation for Zoomorphology) project [79] proposed a content-based image retrieval system designed to rapidly annotate both histological phenotypes and identify potentially confounding imaging artifacts. In spite of their high performance, most of these automated systems do not allow users to incorporate domain knowledge in the annotation process. Our approach adopts semantic content analysis to facilitate exploration and annotation of the histology image collection. Additionally, GRAPHIE also allows the user to interactively evaluate the significance of image features.

Little work has been reported on the use of interactive visualization techniques to aid domain experts explore histology image collections. A web application for remote visualization and collaborative annotation of histology images was proposed in [89]. Jeong et al. introduced a visualization framework that targets interactive examination of histology image stacks [90]. While facilitating the annotation and analysis of histology images, these approaches still put the emphasis on individual images. In contrast, our approach systematically organizes image data and helps users discover implicit and latent relationships among phenotypes manifest as images.

There are many approaches to visualize the internal relationship of high-dimensional data. Techniques like the classic multidimensional scaling (MDS) and the heatmap
have been used to visualize the similarity between images [91, 84]. These tech-
niques share the advantage of compactly displaying a large amount of data in an
intuitive format. Our approach uses graphs to visualize the similarity structure of
image collections. Graph representations are widely used in large population stud-
ies [92, 93, 94, 95]. Interactive methods to explore and analyse network topology as
well as the multivariate data are presented in [96]. In [97], Palmer argues that node-
link representations are powerful to display the internal relationship. For this purpose,
a graph has a more perceptual impact especially on gleaning proximal and similitude
relationships [98]. Compared against the MDS technique, graph-based approaches
maintain the topology and relationships among data entries, while, this information
is difficult to glean after embedding in lower dimensional 2D planes or even 3D spaces.
Additionally, GRAPHIE provides a much richer set of interactive analysis functions
than the above approaches, allowing a more effective and efficient browsing of images.
To the best of our knowledge, our prototype is the first system that allows users to
focus on the entire histology image collection instead of an individual image.

3.3 Task analysis and design

To better understand the needs in exploring histology data collections, we worked
closely with domain experts. To reiterate, the domain experts expressed an interest
in a visual analytics system to aid them in exploring and annotating histology images.
Through our interactions with the analysts, we derived three main tasks that guided
the design and implementation of GRAPHIE:

Task 1 Overview of the histology image collection. With a large histology im-
age collection, it is important to gain an expedient overview, so that users can
glean the similarities inherent in an image collection at a glance. Before analyzing an individual image, researchers often would like to have an overview of the entire image collection to learn the basic contents of a collection, their distributions, and their relationships at a glance.

**Task 2 Selection and comparison of images subgroups.** Selection is extremely important in an image-based phenotypical study. Our collaborators mentioned that during analysis that searching for images sharing the same morphological features and viewing images back-and-forth is time-consuming and tedious. Researchers seek more efficient ways to select desired images and scrutinize the morphological difference between subgroups of images.

**Task 3 Efficient annotation of images.** The annotation set is one of the key outputs of the histology image collection analysis. Our collaborators expect that the tool allows users to record the annotation while exploring the image collection.

*GRAPHIE* was designed to support these three tasks by providing a visualization overview of the image collection and an interactive user interface to examine individual images and the entire collection. Without such a visualization and an interface, the researchers’ ability to explore and annotate the image collections is limited. After several iterations, we agreed upon using graph visualization to represent the image collection: that is each node in the graph represents an image and the layout of the graph reflects the multivariate distribution and depicts the similarity relationships present in the image collection. In addition, the selection and annotation of individual
images can also be achieved by interacting with the displayed graph. We further justify the tasks and our approaches below.

### 3.3.1 Image representation

In order to create a visualization of the histology image collection (Task 1), one first needs to define a way of measuring the similarity between images. Thus, it is necessary to use an image representation to summarize the images’ content quantitatively. However, analyzing histology images is particularly challenging, since visual patterns are generally complex combinations of fundamental visual features associated with texture, color and shape [99]. Even with experienced pathologists, the visual inspection process often suffers from the disadvantages of being subjective, laborious, and insufficient when complex information is needed or is simply unknown.

In this work, we employed the *bag-of-features* (BoFs) [83] approach to represent a collection of histology images. This approach is an evolution of texton-based representations and is also influenced by the *bag-of-words* representation for text classification and retrieval. The key property of the BoFs approach is that it represents images as orderless collections of local features.

The procedure for generating a Bag of Features image representation is summarized as follows: first sample images from the entire image collection are collected. Then a visual codebook is constructed by clustering features extracted from the sample images. The image features aim to capture the texture patterns from the local areas of the image and Clustering is required so that a discrete codebook can be generated from large amount of local features. Each feature cluster is a visual word in the visual codebook. Finally, given a novel image, features are extracted and assigned.
to their nearest matching visual word from the visual codebook. Subsequently, the given image is represented by the frequency of each visual word that it contains.

An important advantage of the BoFs approach is that it models image content in a robust way. The BoFs approach examines small characteristic image regions, allowing the representation of complex image contents without explicitly modeling objects and their relationships. In doing so, we simultaneously obviate the need for segmenting images—which is often in itself a formidable challenge. It should be noted that the BoFs approach has found much success when deployed on images and towards general computer vision tasks [83]. Further, the BoFs representation has been successfully applied to some problems in medical imaging [100].

3.3.2 Graph visualization of image collection

The overview should display the image collection in a visually meaningful manner using sensible layouts (Task 1). In this work, we employ interactive graph visualization to provide an overview of the dataset and allow users to explore the entire image collection. In our case, the similarity between images is defined by the distance between corresponding image representations. The graph is constructed such that each node in the graph represents an image and nodes are only connected by an unweighted edge when two images are similar. Therefore, images with similar contents tend to be closer in the graph thus providing a convenient way to compare and analyze groups of subjects. There exist many similarity graph construction algorithms. In GRAPHIE, we choose the $k$-Nearest Neighbor which has been widely used in many visualization and machine learning applications [101].
3.3.3 Feature selection

Our collaborators observed that data-driven graphs may not capture the entire set of relationships of the target phenotype population. In other words, image representations generated in an unsupervised manner often contain a large amount of irrelevant and redundant features. This makes it very difficult to depict the true relationship between nodes in a given graph. For meaningful exploration, it is useful to find features that are highly correlated with distinct semantic classes. We have thus implemented a functionality via which a user can select arbitrary regions of the graph and examine which features can separate images into different classes. Users then can select the feature subset with higher distinctive power and use them to refine the graph visualization.

3.3.4 Interactive user interface

Interactivity is an essential requirement given the complex nature of heterogeneity of histology image data. As stated in Task 2, the tool should allow users to select arbitrary subgroups in the graph visualization, and display them to help users to analyze differences or similarities among groups. We designed and implemented a web-based application with four interactive components, enabling flexible, efficient and adequate analytical functionalities for histology image data exploration.

3.3.5 Image annotation

Scoring and annotating histology images plays an essential role in phenotype study (Task 3). The annotations provided by the pathologists could guide researchers to discover phenotype-relevant biomarkers [102]. Nevertheless, the annotated histology image collections are an important source of information and knowledge, which may
support educational activities and various research studies. From a machine learning point of view, accurate annotations are considered to be valuable labels which directly influence the quality of automated histology image annotation systems. Given the complexity of visual patterns in histology images and the lack of rigorous phenotypical definitions, the traditional side-by-side manual annotation is often subjective, and sometimes an error-prone process [103]. Thus, the proposed tool allows users to interact with the constructed graph and enable the annotation of the image collection in a consistent fashion.

3.4 GRAPHIE-Graph based Histology image Explorer

GRAPHIE is an interactive visual analytics tool designed for the exploratory analysis of histology image collections. We first describe the workflow that GRAPHIE realizes. Our workflow consists of two main parts:

1. **Back-end generation of image representation:** generate visual codebook and encode images as BoFs representations.

2. **Front-end visualization interface:** enable users to efficiently explore and annotate images collection by creating and manipulating a graph that accentuates similarities and distinctions across images.

3.4.1 Back-end image representation generation

As we stated earlier, we choose the bag-of-features (BoFs) approach to represent a collection of histology images. This workflow of BoFs is summarized by the schematic in Figure. 3.2. Now, we describe the specific steps of our image representation process:
Figure 3.2: Workflow to create image representations: (a) Preprocess each given histology image. (b) Extract feature set. (c) Build a visual codebook through unsupervised clustering. (d) Generate the bag-of-features image representation.

**Preprocessing:**

Generally, histology image data is replete with noise, artifacts and non-informative regions. Appropriate preprocessing of data is necessary for robust analysis. The choice of appropriate preprocessing methods will depend on the nature of the given image collection. In this work, the images were converted into gray scale without loosing the
image texture properties and processed for noise removal using anisotropic diffusion filter [104].

**Feature extraction:**

A scale-invariant feature transform (SIFT) [105] is employed towards the histology images in our study. SIFT descriptors are local and based on the appearance of objects or artifacts at particular interest points (e.g., scale-space extrema) and are invariant to image scale and rotation. They are also robust to changes in illumination and noise. Additionally, they are highly distinctive, relatively easy to extract and allow for robust object identification with a remarkably low probability of mismatch. Mikolajczyk and Schmid [106] asserted that SIFT-based descriptors outperform the other image descriptors (including Gabor filter banks, image moments, etc.) in many situations. Here we used the following parameter configurations to compute SIFT features: 8 orientations and $4 \times 4$ blocks of cells, resulting in a descriptor of 128 dimensions. This configuration of SIFT feature has been widely used by the computer vision application, as it is capable of capturing most texture patterns in the image collection [107].

**Visual codebook training:**

Given the feature descriptors extracted above, a visual codebook characterizing extant visual patterns is generated in an unsupervised manner. We use a clustering approach to prune down the data features to a core set of representative features (cluster centres) that constitute the visual codebook. The $k$-means algorithm is used in this work to find a set of cluster centroids that correspond to visual words. Although it has been reported that learning large numbers of $k$ can improve supervised
classification results [108, 109], we observed that finding the appropriate number of 
\( k \) depends on the size of the image collection and the specific phenotypes present in 
the image collection.

**Image representation:**

Given the visual codebook, a candidate image is represented as follows. Features 
are extracted from candidate image and then assigned to bins, where the bins are 
obtained by quantifying the feature space using the words in the codebook. The 
resulting histogram—where the count for each bin gives the frequency of occurrence 
of the words—is now the representation of the histology images.

Figure 3.3: *GRAPHIE* interface: Graph visualization of the image collection from case 
study 1.
3.4.2 Front-end visualization interface

The design of GRAPHIE visualization interface follows the principles of creating efficient visualization systems suggested by well-known InfoVis mantras [110, 111]. We further deploy Prof. Shneiderman’s general abstract tasks [110] (overview, zoom, filter, details-on-demand, and relate) to the interactive functions design of essential components. Specifically, we implemented four main visualization components (Figure 3.4) to achieve these general abstract tasks: (a) Graph visualization of the image collection allows users to gain an overview of the entire image collection and view relationship (relate) among images; (b) Individual image view lets users zoom in the selected images; (c) Subgroups gallery allows users to filter out uninterested images;
(d) Feature selection view lets users improve the graph visualization on-demand. The layout of our interface uses multiple coordinated views [112], therefore users can see all four components within one page. We also aim to keep our user interface and interaction design as simple as possible so that users with little training can start using our tool immediately.

Graph visualization of the image collection:

With the BoFs histogram summarizing the content of images, all pairwise distances between images are computed using the BoFs histograms. Euclidean distance is used in this work. The resulting similarity matrix reveals the inner relationships of the entire image collection. The similarity matrix provides a fully connected graph, which is pruned down for effective visualization. The goal of pruning the graph is to extract and summarize the topology of the underlying feature space. As mentioned previously, GRAPHIE adopts the k-Nearest Neighbour (k-NN) to construct the unweighted graph. In a k-NN graph, each node only keeps edges which connect to the k closest nodes, where the euclidean distance between corresponding BoFs features...
defines closeness. Given the graph, we visualize it using force-directed graph layout—a method that has been shown to be effective in creating uncluttered visualizations[41]. The layout is fully interactive; users can adjust the graph layout by clicking and dragging the node. Users can also interactively browse a single image by clicking nodes to examine them in more detail in the individual image view. In order to avoid clutter, users can adjust the size of the nodes and the parameters pertaining to the metaphorical charges and forces that constitute the graph’s layout. Nodes can be further colored with categorical annotations, thus allowing one to explore the structure
of the data more easily. We provide more specific examples when we later discuss the two case studies.

**Individual image view:**

The individual image view allows users to inspect individual images and annotation of them. Users can either select images by clicking nodes in the graph or by searching for them by name. This view also contains the preview and annotation of the selected image. Users can inspect the full resolution of selected images in a pop-up window by click the **zoom-in** button. There are two options to annotate the selected image: users can either select the corresponding score or qualifier by interacting with a drop down button or input notes in a text box at the bottom. Once, the entire image collection is annotated users could export the annotation by clicking the **save annotation** button.

**Subgroups gallery:**

The subgroups gallery component enables browsing and comparing subgroups when needed. Each row in the subgroups gallery represents a subgroup of images. Users can select/update subgroups of the image collection by interacting with the derived graph visualization. Further, users can interactively browse the images by moving the scroll bar and clicking a thumbnail to examine each on the individual image view. The subgroups gallery also enables users to batch annotate all images in the subset by selecting the corresponding score with the drop down button on the left side.
Feature selection view:

Although BoFs representation could effectively characterize the histology image, it is generated in an unsupervised manner and can inadvertently generate irrelevant phenotypical features. We implement the feature selection view which enables users to examine the distinctiveness of visual words for selected groups of images, therefore enhancing the visualization. Users can define the groups by selecting nodes from the graph visualization according to target phenotypes, or by using existing annotations. In order to select visual words that are most different between two groups of images, we conduct the Student’s $t$-test for features to test hypotheses for group discrimination. We list features in ascending order using their $p$-values, computed according to Student’s $t$-distribution. The features with more significant mean differences between selected groups are more likely to distinguish the images in those groups. Besides the significant statistic, we also provide boxplots for each feature to display summary information of feature distribution in selected groups. These plots help in better estimating the separability of the groups by the selected feature. In order to further investigate the quality of features, GRAPHIE enables users to select a feature subset from the given BoFs features. Then, users can regenerate the graph with the selected feature subset for the entire image collection to further inspect the changes in the new feature space. In section 3.5 we discuss the use of this component.

3.4.3 Implementation details

The visualization interface is developed in Javascript and $R$ using the $R$/Apache module running on an Apache server. The data processing is implemented with $R$
script which is triggered by Javascript. The interactive visualization is created using \textit{D3.js} [113], a visualization JavaScript library.

3.5 Results

In this section, we demonstrate the efficacy of our methods using two case studies. We described two repositories of data in our possession in the introductory section. The first case study serves the basic science community where the variety in histology of a specific animal model is examined. We especially focused on a portion of the zebrafish retina given that the structure is well understood and that we have labeled descriptions of the morphology. The second case study serves clinical practice where clinicians and pathologists classify patient biopsies to grade and diagnose diseases such as cancers. In this case study, epithelial-stromal tissue slides were examined.

3.5.1 Zebrafish retina histology images

In biomedical research, the use of animals such as zebrafish as a genetic model is useful for understanding the basic mechanisms that underlie human disease [81]. The heterogeneity expressed by the morphology in the eye of the zebrafish also makes the eye a perfect object for our phenotypical study. Like most classes of extant vertebrates, the retina of a zebrafish is composed of seven major cell types derived from the neural ectoderm, six types of neurons and one type of glial cell [82].

The histology images in this case study are obtained from the Zebrafish repository [79]. A total of 168 images of larval zebrafish eyes were manually extracted from 20× magnification virtual slides acquired by the Zebrafish Functional Imaging Core facility at the Pennsylvania State College of Medicine. Image dimensions measure \(768 \times 768\) pixels. Each slide is a well-stained with hematoxylin and eosin
(H&E stains); the nuclei are the targets of hematoxylin while eosin stains the cytoplasmic and stromal regions. Each image is marked with a score (ranging from 0 to 3) which represent the level of phenotypical abnormality. A score of 0 indicates that the image has no visible abnormality while a score of three indicates an extremity of occurrence of corresponding abnormality. This ground truth scoring is manually recorded by pathologists with expert knowledge of the vertebrate anatomy. With the traditional slide-by-slide inspection, it is extremely difficult for a user to capture the subtle differences and annotate these images consistently and objectively. Even with the abnormality annotation made by a domain expert, it is still hard to comprehend the abstract visual pattern characteristics.

We now demonstrate how one can explore the zebrafish dataset with GRAPHIE. First the BoFs-based image representation was computed. In order to cover most of the visual patterns, three images were manually selected from each score subset (total twelve images) as representative images. We built the visual dictionary with features extracted from these twelve representative images. We tested with several visual dictionaries using $k$-means clustering, varying the value of $k$. The resulting graph visualization did not improve significantly when $k$ given larger than 25, thus we chose $k = 25$ as the dictionary’s size in this study. With the BoFs histograms, the similarity distance between images was calculated using the Euclidean distance metric.

Figure 3.7(a) illustrates the initial graph visualization obtained in the form of a $k$-NN graph($k = 3$) with 168 image nodes. The color for each node represents the abnormality score assigned to each image by our collaborator. The color blue indicates abnormality score of 0, which essentially implies that the genotype of the zebrafish
Figure 3.7: Morphology zebrafish retina case study. (a) Initial graph visualization. (b) Example phenotypes of the zebrafish retina: (red) retina of a mutant zebrafish. The mutant possess a small lens and fissure closure defect. (blue) Eye of wild-type zebrafish. Normal retinal lamination and a fused ventral fissure can be seen. (c) Graph created with selected features: the resulting graph shows better separation between wild-type and mutant-type zebrafish eye images, also helps users to spot the mis-annotated images. This figure shows three mis-annotated samples, with the upper two images were mis-annotated as wild-type and lower image being mis-annotated as mutant.

as manifest in the image is wild-type. Light blue indicates abnormality score 1, a mutant zebrafish, with a subtle alteration from the wild-type phenotype. Light and dark orange respectively indicate that the mutant zebrafish has abnormality scores of 2 and 3, signifying a high level of abnormality in the specimen. By interactively examining the images in individual image view, we can observe that in the initial graph (Figure. 3.7(a)), nodes close to each others tend to have similar properties and visual patterns. For instance, consider the nodes shown in Figure. 3.7(b), the images lying
in the red rectangle (nodes selected from bottom of the initial graph in Figure. 3.7(a)) all have abnormality score 3. And in fact, the cells in these retinas are highly disorganized. On the other hand, the images in the blue box (images selected from upper right of the initial graph in Figure. 3.7(a)) are resembling organizations often found in non-mutant, wild-type species. This distribution of nodes indicates that the proposed framework is able to capture different levels of phenotypical abnormalities present in the data.

However, the absolute separation between wild-type and mutant-type images remains unclear in the initial graph visualization. The small cluster in the upper left corner is connected to the main cluster of mutant-type images but is distinct from the main cluster of wild-type. Therefore, we performed feature selection in order to remove visual words that are irrelevant to the target groups. In this particular case, we used the abnormality annotations as the group labels. In order to pick the visual words that are most different between two groups of images, we conducted a Student’s $t$-test on each visual word, computed per group to indicate the ability of separating the two groups. As the initial visualization is not optimal, a new graph is regenerated with the top 10 ranked visual words. As noted in Figure. 3.7(c), the newly created graph has a much more clear separation between wild and mutant-types. We took a closer look at these images interspersed between distinct wild-type clusters. Interestingly, the abnormalities of these eyes are relatively subtle. When confirmed with our collaborators, some of these images were mis-annotated. With the graph visualization and subgroups gallery components, we could easily spot potential mis-annotated images which is a challenging task with the traditional 2-D grid-layout.
Through this case study, we can note that graph visualization is superior to a 2-D grid-layout for exploring image collections due to its ability to organize the images based on its content and reveal relationships between groups. By interacting with subgroups gallery and individual image view, one could efficiently compare and spot the differences between subgroups of images, therefore, form hypotheses regarding phenotypes. In addition, graph visualization can be improved by interactively selecting features with feature selection component.

3.5.2 Epithelial-stromal tumor tissues

![Figure 3.8: Examples of heterogenous tissue compartments in the histopathological images of breast cancer in our adapted dataset. (a) Epithelial tissue. (b) Stromal tissue.](image)

In a second case study, we deployed our methods on digitized tumor tissue slides. An important application of histology images is the examination of cancer tissue slides, wherein pathologists evaluate the composition of epithelial and stromal tissues and their interactions. Tissue slide analysis is the standard approach for the analysis of diagnostic, prognostic and predictive morphology biomarkers [114]. A majority of solid tumors are composed of epithelial cells. Classifying the tissue as epithelial or
stromal is an important part of the automated cancer diagnosis [115]. In this case study, we demonstrate that through GRAPHIE researchers can gain a better perspective of the underlying structure of the tumor microenvironment that is composed of both epithelial and stromal tissues.

Here, we present our visualization of the epithelial-stromal distributions. The histological images we used in this case study are collected from The Ohio State University (OSU) Pathology Core Facility. In order to view the datasets in more detail, we use the superpixel method to divide each image into 50 patches [116] where each patch contains an approximately homogeneous visual pattern and homogeneous tissue (Figure. 3.8a stromal / Figure. 3.8b epithelium).

The visualization workflow adopted here is similar to the one used in the previous case study. We randomly sample and select superpixel patches, and train a codebook. A superpixel patch represents a region of homogenous tissue. For each patch, we again generate a BoFs histogram to summarize the image. Here, each node represents a patch. Due to the large amount of nodes, clutter and costly computation are inevitable with the use of a straightforward force-directed layout scheme. To better interact with the visualization and to allow for the examination of the quality of visual words, we randomly sample 400 patches from the collection. In this case study, we carried out the following two explorations:

**Exploration 1:** Figure. 3.9(a) is the resulting graph which consists of 400 nodes. The color of each node represents the type of tissue in that patch. A blue-colored node indicates a patch containing mainly epithelium; an orange-colored node indicates a patch that mainly contains stroma. We observe that the blue and orange nodes form two clusters. However, the graph visualization shows that clusters of mixed tissue
Figure 3.9: Epithelial-stromal case study. In graph visualizations, an orange-colored node represents a patch contains stromal; a blue-colored node represent a patch contains mainly epithelium. (a) Initial graph visualization. (b) Graph created with selected features.

prevail. Investigating the content of patches, we find that these patches contain both stromal and epithelial tissue. This artifact occurs because of the lack of quantitative control and subjective bias in our own manual labeling process.

Exploration 2: In another exploration, we perform the feature selection over the samples and create a new graph (Figure. 3.9(b)) by using only the top 6 visual words. Comparing with the initial graph, the new graph shows an equally good separation between the two types of tissue patches with less visual words. Thus, the classification task can be efficiently achieved with a much lower feature dimensional space. It should be further noted that distinct groups are delineated. For instance, subgroup i (Figure. 3.9(b)) includes purely stromal patches while subgroup iii includes only epithelium patches. In addition, subgroup ii includes patches with a mixture of stromal and epithelium.
3.5.3 Preliminary User Feedback

We conducted a preliminary study where we scrutinized the efficacy of GRAPHIE. Ten participants were recruited from our university setting. All participants were either graduate students or medical students who had computing or bioinformatics background. In order to evaluate GRAPHIE, participants were asked to annotate the unlabelled zebrafish data collection with both GRAPHIE and a grid lay-out of image collection. For comparison, we created a web-page and showed all images in 2-D grid-layout (shown in Figure. 3.1).

The study began with a brief introduction and demonstration (circa 10 minutes) of the various features of GRAPHIE. The participants were not given access to the manuscript or relevant supplementary material. Then, participants were presented the unlabelled zebrafish dataset. After a brief instruction pertaining to the labelling and the presentation of a few example images, participants were asked to annotate histology images according to perceived impact of mutation. During the study, the time needed for a user to complete annotation task was also recorded. Participants were asked to think aloud, in order to record any comments. At the end of study, we provided questionnaires and later conducted a short interviews to collect feedback.

After annotation with one specific tool, the users would naturally gain sufficient knowledge of the underlying data, which may affect their annotation performance with the other tool. In order to minimize the effect, participants annotated the dataset with GRAPHIE first and were asked to return after two weeks and complete the test with the second tool.

The results are reported in Figure 3.10(a). The annotation accuracy with GRAPHIE is significantly higher than using grid-layout. Interestingly, we observed that
among all participants the medical students achieved a higher accuracy with both GRAPHIE and the 2-D grid-layout, while others achieved better accuracy with GRAPHIE. We also observed that most participants spent less time to finish the annotation task with GRAPHIE. Majority of the participants suggested that the batch labelling function significantly enhances the labelling process and the subgroups gallery mechanism helped them compare the existing differences between images and identify the images which are vastly different from the rest of images in subgroup. We also received comments that annotating with the 2-D grid-layout is much more tedious than with GRAPHIE.

Figure 3.10: (a) User annotation accuracy boxplot. (b) User satisfaction results from the post-study questionnaire.

In a post-hoc survey, participants completed a questionnaire of six questions using a 1-5 Likert scale (from strongly-disagree to strongly-agree). Figure 3.10(b) lists the questions reports the average user ratings and also lists the standard error. Based on
the results pertaining to question Q1, participants agreed that \textit{GRAPHIE} is easy to learn and use, which was fairly encouraging as novice users were able to conduct labelling in \textit{GRAPHIE} with a very brief introduction and minimal assistance during the experiment. Regarding the exploration and annotating capabilities (Q2-Q5), participants accorded high scores and the small reported variances reflected the consensus users had on the effectiveness and usefulness of the graph visualization. Overall, participants ranked their desire of using \textit{GRAPHIE} to explore histology data as 4.7 out of 5.

### 3.6 Summary

This chapter presents \textit{GRAPHIE}, a visual analytics application designed to explore the histology image collection. By taking a data-driven approach, We developed an unbiased way for visualizing the entire collection. \textit{GRAPHIE} not only provides an intuitive overview of the data but also enables users to use domain knowledge to improve the visualization through interactive feature selection. The visualizations and interactions of \textit{GRAPHIE} are seamlessly integrated to allow users to effectively explore and annotate images, with a rich set of interactive functions. The use of \textit{GRAPHIE} was evaluated with two datasets.

The current prototype implementation suffers from a lack-of-scalability, the size of the image collection that can be meaningfully explored is limited. We believed this problem can be tackled by applying different techniques pertaining to graph layout including semantic zooming, focus+context exploration techniques and sparse sampling strategies.
In the future, we plan to test GRAPHIE with more histology image datasets and improve the image representation with more options of feature descriptors. We also aim to integrate with other types of data (e.g. genetic and epigenetic data) to enable an integrative phenotypical study.
In the last chapter, we present GRAPHIE for exploring the morphological heterogeneity within a single type omics data (histology image). However, in order to comprehensively understand the diseases, the study needs to include multiple types of omics data ranging from genotype to multiple levels of phenotypes. Since different biological levels are not independent, appropriate integrative approach is required to take into account the relationship between various types of data.

In this chapter we present another visual analytic tool, namely Integrative genomic based cancer patient stratification (iGPSe). iGPSe is designed for integrative cancer patient stratification using multi-omics data. The tool allows users to explore the feature spaces in different types of omics data with various visualization components and subsequently carry out an integrative analysis through a model-based integration method. The key feature of iGPSe is that allows users to directly verify the results of integration via an interactive parallel sets visualization.

We conducted a use case study using a breast cancer dataset consist with gene expression and microRNA profiles. We demonstrated that users can quickly explore

\[2\text{This chapter is published in BMC Bioinformatics 2014 15:203}\]
different combinations of gene expression and microRNA features and identify potential integrative markers for survival prediction.

This chapter is structured as follows. Section 4.2 describes previous work. Section 4.3.1 describes the overall requirements of our system and the underlying design rationale, while we describe the workflows of $iGPsE$ in Section 4.3.3. Comprehensive case studies and user evaluation are reported in Section 4.4.1 and Section 4.4.2 respectively. We end this chapter by summarizing the contributions of this chapter in Section 4.5.

4.1 Introduction

During the past fifteen years, high-throughput genomic experiments, which involve the usage of micro-arrays or next generation sequencing technologies, have significantly changed biomedical research and clinical practice. These technologies have expedited the process of discovering genes implicated in important biological phenomena and molecular markers for disease. However, for a better understanding of large high-dimensional data associated with high-throughput experiments, intuitive visualization tools are needed in order to effectively interpret the data and extract new biological knowledge and insights. Specifically, there is an important and unmet need to integrate collected data with previous biological and clinical knowledge. It is thus important to access large repositories of often structured biological knowledge and further to have access to interactive tools that facilitate the required integration and analysis of all data, especially against the backdrop of prior knowledge. The specific need of integrative visual analytics is of particular importance given the growing trend
of using integrative genomics (or panomics) approaches for personalized treatment of diseases, including cancer.

Most types of cancers are highly heterogeneous with different identifiable subtypes. These subtypes often possess different genetic variants, present different pathological phenotypes, and most importantly, confer different clinical outcomes such as varied prognosis and response to treatment as well as different likelihood for recurrence and metastasis. Patient stratification is necessary for prescribing viable regimens of treatment and also towards the discovery of prognostic and/or predictive biomarkers. In order to robustly characterize patient subtype demographics to achieve precision medicine, panomics approaches are being increasingly used especially in tandem with the development of large collaborative projects such as TCGA (The Cancer Genome Atlas). In these projects, large cohorts of patients were recruited and many different types of “omics” data including genotypes (e.g., single nucleotide polymorphism, copy number variance, and somatic mutations), gene/microRNA expression, epigenomics (e.g., DNA methylation), proteomics, pathology images and clinical records as well as outcome information were collected from these patients. It is conceivable that by integrating the data ranging from genotype to multiple levels of phenotypes, more precise and robust stratification of the patients with clinical outcome difference can be achieved. Very least, conflicting stratifications arising from the consideration of each data collection separately are avoided, leading to identification of potentially more robust biomarkers.

However, integrative analysis is a challenging task. Most of such analysis requires extensive analysis and development of complicated algorithms such as network integration [117], statistical association and regression [118], and partial least square
analysis [119]. Since many of the algorithms are still in the development and testing stage, users are often required to have extensive preparation in quantitative analysis, algorithm development and computational methods. These requirements severely hinder the wide utilization of such data by clinicians and biomedical researchers who are often only trained in clinical and biological sciences. Therefore, there is an urgent need for effective tools that allow biomedical researchers achieve tangible exploration of patient population using multiple types of “omics” and patient outcome data.

In this chapter, we present a visual analytic system called interactive Genomics Patient Stratification explorer (iGPSe), designed to help biomedical researchers to perform patient stratification on-the-fly and visually explore disease subtypes in heterogeneous genomic data. In iGPSe, we employ machine learning algorithms that identify sub-populations based on molecular feature sets chosen by users. To be effective, iGPSe relies on an interactive realization of parallel sets and patient survival plots between selected patient groups. Thus, investigators are able to quickly examine how a selected list of molecular features of interest can separate patients into different groups and whether these groups show a difference in clinical outcomes. In addition, iGPSe offers several visualization techniques to help users evaluate the quality of the stratification results and thus assess the effects of clustering algorithms. In this work, we demonstrate the use of iGPSe on the stratification of breast cancer patients using both mRNA and microRNA (miRNA) expression data. The most distinctive feature of iGPSe is a novel visualization scheme to explore combinations of different data types and identify combined markers for robust survival prediction of a given population. Additionally, iGPSe integrates standard patient stratification workflows into an intuitive, user friendly interactive platform.
Figure 4.1: Screen shot of the interactive analysis page. On the top left, patient samples are clustered based on both mRNA expressions (left panel) and miRNA expressions (right panel). Heatmaps of the clustered data are shown aligned with the parallel sets. On the top right panel, the survival plot shows the patient outcome information and the power of statistical significance (p-value). On the bottom, force directed graph demonstrates the affinity between patient samples. The patient samples selected in the parallel sets are circled. The color of the nodes correspond to the ones in the parallel sets.
4.2 Previous work

Cancer tumors that seem very similar when examined through conventional diagnostic methods might look different at the molecular level leading to different and effective outcomes and/or treatment responses. Therefore, molecular features are being increasingly used to stratify patients to support more accurate and robust clinical and therapeutic decisions. Over the past decade, molecular stratification of tumors using gene expression microarrays has been an important area of cancer research [120, 121, 122, 123, 124, 125]. A typical stratification study often includes the application of statistical techniques to population groups including supervised learning and unsupervised cluster analysis. Heatmaps have been widely used in many tools [126] to visualize molecular signature patterns manifesting in various subgroups.

One of the popular genome tools is the UCSC Cancer Genomics Browser [127]. It allows researchers to identify and assess genomic signatures in cancer subtypes, to compare and contrast subtypes, and to assess their role in stratifying patients into different groups. However, this ability is restricted to one dataset at a time and offers no integrative capabilities. Many methods have been recently reported to discover characteristics from multiple classes of measurements [128, 129, 130, 131]. These computational methods either build statistical models [129, 130] or construct multiple networks or patient samples [131]. In bioinformatics, integrative analysis is becoming more prevalent with the increased adoption of the integrative genomics or the panomics [19, 132, 66, 133, 134]. A major challenge in any such integrative study is that the patient population is very heterogeneous for any given type of measurement data and thus the ensuing stratifications are often very different. Integrated visualization of heterogeneous data has not received much attention [135, 136].
A salient integrative visualization tool, StratomeX [137], was proposed towards the visualization and exploration of subtypes in a population afflicted with cancer using the TCGA data. StratomeX relies on the visual acuity of heatmaps and the discriminative efficacy of clustering of collected genomics data. It is however difficult to discover the underlying structure that exists in a population and pose hypotheses that compares and contrasts characteristics and outcomes across sub-populations. Additionally, the high dimensionality of the data makes it difficult to display all of the relationships in a meaningful manner.

The concept of integrative analysis of biomedical data, but it is a recurring one that only recently been unfolded as a major challenge in part driven by technology development producing increasing amounts and type of data.

4.3 Methods

We developed the visual analytics system $iGPsE$, a web based tool facilitating patient stratification and exploration of disease subtypes in heterogeneous genomic data. We first elaborate the overall requirements of our system and the underlying design rationale. We continue by presenting visualization techniques adopted in $iGPsE$. We then describe the $iGPsE$ workflow design and implementation.

4.3.1 Requirements Analysis and Design

In this subsection we outline various design decisions that ultimately guided our choices for computational methods and platforms. Our methods share commonalities with genome wide association studies (GWAS). To begin with, our studies will require a large population of patients. In addition, our collections of datasets are of multiple modalities (imaging, molecular expression, etc.), high-dimensional, large, and highly
heterogeneous. In this section, we first describe input data under consideration and then enumerate various functionalities that will be required in a typical patient subtyping study. Eventually, we will make a strong case for the use of iGPSe to perform and evaluate patient stratification.

Data Characteristics

The population data used in iGPSe is extracted from the TCGA. Each patient in the population is associated with gene expression, microRNA expression data and clinical information such as time until death, gender and age. For our current prototype, we limit our collection to include gene expression (mRNA) and micro-RNAs (miRNA) expression. The expression data were subject to log transformation and normalization following standard bioinformatics practice. Clinical information includes the age, tumor grade, survival time and survival status for each patient.

User Wish List for iGPSe

In order to identify the system requirements, we worked closely with domain experts. The suggestions and advise collected from our collaborators helped us to design, implement and improve our system. Accordingly, we identified the following functions, which are directly based on observations of the domain experts’ standard data analysis workflows. The implementation aspects will be discussed in detail in the next section.

1. Interactive feature selection: Molecular features such as genetic variance and expression levels for selected genomic regions and genes are often used as potential biomarkers for identifying cancer subtypes, which helps in early
diagnosis and effective treatment of cancer patients. Given the nature of cancer microarray data, which usually consists of a few hundred samples with thousands of genes as features, the selection of molecular features is important for effective gene expression data analysis. The feature selection step is a commonly addressed problem in machine learning especially in the context of supervised learning where different subtypes are labelled with prior domain knowledge [138]. Thus, one can effectively eliminate the potentially irrelevant or obvious features. As in the past decade numerous papers were published claiming successful application of gene expression analysis to patients subtyping and prediction of survival [123, 124, 125]. Users, therefore, should be able to add/drop/modify features interactively. Then by examining effects of eventual stratification, researchers can verify the quality of the selected feature set and refine it accordingly.

2. **Clustering:** It is now widely acknowledged that cancer is biologically heterogeneous. This complexity accounts partly for the variation in clinical outcome [19]. Given the large variations in genetic and environmental factors, there is a need to detect sub-populations and examine them under different annotations such as clinical outcome or histological types. One of the common methods is to cluster the patient population into subgroups of patients who share similar expression patterns. However, there is no one-size-fits-all solution to clustering. Each algorithm and similarity measurement impose certain assumptions on the data set. Different clustering algorithms can give widely differing stratifications, especially when deployed on gene expression data. Thus, the application should offer multiple choices of clustering algorithms for the user to choose. Moreover,
the application should allow users to run clustering on-the-fly, rather than use precomputed results. In this way, users can choose the algorithm that works the best on their input data set and have more control over the stratification process. The details of our implementation is described in Section 4.3.3.

3. **Cluster refinement:** In many cases, the notion of a patient subgroups is not well defined in the selected feature space. The stratification of the population is, in most cases, not unique. Moreover, the number of the viable subgroups is also mostly unknown. It is important to do an evaluation of how well an algorithm performs on expression data sets. The application should provide some clues for the user to evaluate the quality of the clustering results, e.g., whether the chosen number of clusters is appropriate, the high-similarity within a cluster and the low-similarity between clusters. Visualization techniques can provide the evaluations. We integrated three visualization techniques into our system to help users analyze and improve the clustering results.

4. **Integrated analysis for patient stratification:** Our main goal is to provide various ways to form combinations of data types interactively. The visualization of these combinations should be intuitive and easy to navigate. The user must be able to detect patterns in the data through rearrangement and grouping. Further, he/she must be able to drill down to interesting details and obtain biological insight. The stratification based on each type of data provides a candidate subtypes. It is also important to include all collected types of expression data [137]. There have been efforts on developing combined predictive models that include at least two or three modalities [139]. Clustering results
heavily depend on chosen parameters and different types of expression data will give rise to different clusters. Consequently the application should be able to compare stratification results from different data types, which could help researchers answer more complicated questions including whether there exist dependencies between stratification results from different data types. It should also offer interactive functions which help the user explore various combinations of subtypes. In order to perform the integrative analysis we adopt an interactive visualization method, parallel sets, which give an intuitive evaluation of the cluster coherence across different types of expression data and enable the user to interactively explore the various combinations of clustering. Details of our integrated analysis are described in Section 4.3.2.

5. **Interactive user interface:** Interactivity is an essential requirement given the complex nature of variation inherent in the datasets and the multitude of measurements and outcomes. The user needs to be able interactively to select sub-populations and clusters and compare outcomes and intrinsic labels between them. The user interface should be clear and intuitive, so that users with no computer science background are still able to operate the application. Moreover, it should be feasible to customize certain aspects of the visualization.

### 4.3.2 Visualization Components

In this section we describe the visualization components that are designed to address the aforementioned requirements. We adopted four visualizations techniques in our system. The heatmap helps users to examine the expression patterns in clusters. The silhouette plot provides an assessment of the relative quality of patient clustering.
Graph visualization facilitates the exploration of the overall population in the selected feature space. The parallel sets display enables users to evaluate cluster coherence between two data modalities. Moreover, representing sub-populations of patients in parallel sets allows the user to interactively visualize the data, as well as to perform statistical tests to ascertain differences in clinical outcomes.

Figure 4.2: (a) Heatmaps (b) Silhouette Plot. Both (a) and (b) are from the same K-means clustering result (k=4)
Heatmap

The heatmap is a popular and intuitive visualization technique for encoding quantitative measurements. In the context of gene/microRNA expression data, the color assigned to a cell in the heatmap grid indicates expression value of a particular gene in a given patient sample. The heatmap provides an overview summary of the data. An example is shown in Figure 4.2(a); the given patient population is stratified into four subgroups. Each subgroup corresponds to one block of the heatmap. The height of each block indicates the number of patients in the corresponding cluster. The main goal of heatmap in our system is to help the user to examine gene expression patterns related to the stratification. In our example, one can observe that the genes show distinct patterns between the clusters while the patterns are consistent within each cluster.

Silhouette plot

Clustering results depends heavily on the parameters (e.g., the number of clusters) and distribution of the data points. In order to evaluate the results (and hence the choice of parameters), we incorporated the silhouette plot [140] into our system. The silhouette plot displays the degree of certainty of each sample belonging to its cluster. For our application, this is measured by the difference of a patient’s average dissimilarity to other patients of its cluster and the patient’s average dissimilarity to all patients to the next closest cluster. The dissimilarity is measured by the squared Euclidean distance between patients in the selected feature space.

Assume the data have been clustered into $k$ clusters. For each patient $i$, let $a(i)$ be the average dissimilarity of $i$ with all other data within the same cluster. We can
interpret $a(i)$ as how well $i$ is assigned to its cluster. We then define the average dissimilarity of point $i$ to a cluster $c$ as the average of the distance from $i$ to all points in $c$. Let $b(i)$ be the lowest average dissimilarity of $i$ to any other cluster, of which $i$ is not a member. The cluster with this lowest average dissimilarity is said to be the “neighboring cluster” of $i$ since it is the next best fit cluster for patient $i$. The silhouette score for each patient $i$ can be defined as:

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

(4.1)

These silhouette scores are standardized between $-1$ and $1$ and a horizontal bar chart of differences is plotted for each cluster. Thus the silhouette plot allows the user to assess the relative quality of the clusters and provides cues to determine the appropriate number of clusters. Figure 4.2(b) shows an example of silhouette plot where the number of clusters is 4. The blue and purple clusters are relatively tight and also close to its neighbouring cluster.

**Graph visualization**

Graphs have been widely used to capture relationships among data points. In this tool we employ interactive graph visualization of similarity matrices to provide a perspective of the population structure in the feature space and allow users explore patient population. The graph is constructed such that each patient is represented by a vertex, and similarities between patients are represented by edges connecting these vertices. Graph visualizations allows one to browse complex relationships and determine community structures. Generating the graph visualization consists of the following steps:
1. **Similarity Matrix**: A similarity matrix represents a complete weighted graph where each node corresponds to a single sample (as measured by gene expression profiles or other features) and the weight of an edge between two nodes measures the similarity between the connected two patients. Similarity between two vertices (i.e., two patients) can be computed in various ways depending on the property of the measurement data. In our application, the user can choose between Pearson’s correlation and the Euclidean distance, two of the most commonly used similarity metrics in bioinformatics.
2. **Graph Sparsification:** The similarity matrix provides a fully connected graph, which needs to be pruned down for effective visualization. The goal of pruning the graph is to extract and summarize the topology of the underlying feature space. We apply a global threshold, keeping only those edges with similarity values exceeding the chosen threshold. With this approach, the resulting graph captures the population structure by connecting only highly similar vertices.

3. **Graph Layout and Vertex Labelling:** Graph layout algorithms project \(N\) dimensional data onto two dimensional space for visualization, with the aim of best preserving actual pairwise distances between vertices in the original high dimensional space. We use the force-directed graph layout method of [141] which was shown to be effective in creating uncluttered visualizations.

An example of the graph visualization based on the gene expression features is shown in Figure 4.3. Since we use a global threshold to sparsify the graph, there are a few isolated nodes. The color labeling of the nodes is consistent with the heatmap and silhouettes plot (Figure 4.2). It should be observed that several purple and blue nodes are densely connected with nodes of the same color. This confirms our previous observation made with the use of the silhouette plot (see above).

**Parallel Sets**

In order to perform the integrative analysis of different types expression data we employ the parallel sets method. Parallel sets enables users to interactively explore various subgroups obtained as clusters from the clustering step. For each data type, user-selected features (mRNA or microRNAs) are used to separate the patient population into a number of mutually exclusive subgroups by a clustering algorithm. Our
Figure 4.4: Parallel set visualization and survival analysis panels in iGPS. **Left:** The clusters of patients using different feature sets are shown as bars. On the left side, the blue/yellow/purple/pink bars indicate the four patient groups separated using the PAM50 breast cancer genes [2]. The gray/green/light blue bars indicate three groups separated using the above discussed miRNAs. The gray bands linking matched patients. **Right:** The Kaplan-Meier curves of the survival times for the two groups of selected patients. The colors of lines matches the colors of the selected bands. In addition, the p-value of the difference in survival times between the two groups based on log-rank test is also listed (p-value = 3.8173e-06 in this example).

Application offers $k$-means, spectral clustering and community detection as choices for the clustering algorithm.

Similar to our setting for the heatmap, in parallel sets, patient subgroups are represented as columns. Different types of data are placed independently side-by-side. In each column, subgroups are encoded by boxes whose height is proportional to the number of patients within that subgroup. The color of the boxes indicates the corresponding subgroups, which is consistent with the node color in the graph visualization for the same measurement. The ribbons connecting boxes represent the matching patients in different types of measurements. The width of the ribbons is proportional to the number of patients. The main goal of the ribbons is to offer an
intuitive view of the consistency of the clustering result between different measurements. A user can combine multiple subgroups into a larger subgroup by selecting the corresponding regions.

A user can interactively generate a Kaplan-Meier plots [142] by selecting subgroups from the parallel sets visualization. This allows users to evaluate the difference in clinical outcome between selected subgroups. In contrast to the purely data-driven approaches, our application enables the user to treat ribbons as subgroups and can interactively combine subgroups from the clustering algorithm. Thus, a new stratification of the population based on multiple measurements is now obtained. An example is shown in Figure 4.4.

4.3.3 Implementation

In this section, we describe the workflow design of \emph{iGPSe} and discuss how users interact with the system. Our system workflow has three phases: feature selection, clustering analysis, and integrative patient stratification.

Our system, \emph{iGPSe}, is developed in Javascript and R using the R/Apache module running on an Apache server. The data processing is implemented with R script which is triggered by Javascript. The interactive visualization part is created using d3.js[113], a visualization JavaScript library.

Feature selection

Within the high-dimensional expression data space, each patient is defined by thousands of genes and hundreds of miRNAs. Inevitably, the data sets contain noisy or irrelevant genes/miRNAs which makes subtyping more complicated. Therefore we need to filter out irrelevant genes and only focus on the genes in which the user is
interested. *iGPSe* offers an interactive gene/miRNA list selection panel (Fig. 4.5). There are three ways to input the feature list: 1) users can select genes/miRNAs of their interest from the table, or 2) type or copy-and-paste the gene names into the input areas or 3) upload a text file (comma delimited format) of the gene list. Once the lists for both genes and miRNAs features are determined, the following patient clustering and stratification will only be based on these selected features. Users can cluster patients in the next step by clicking the ‘Next’ button.

As shown in the Figure 4.5, the gene list is extracted from a recent *Nature* manuscript describing the TCGA BRCA (breast cancer) project [143]. It has been shown that somatic mutations in these three genes occurred at 10% incidence across all breast cancers, which are the highest among all genes. For the miRNA, we picked hsa-mir-130a, hsa-mir-222, hsa-mir-29a, hsa-mir-23a, hsa-mir-24-1, hsa-mir-24-2, hsa-mir-30a, hsa-mir-27a, hsa-mir-22, and hsa-mir-100 as suggested in [144]. This framework can be easily extended to accommodate all kinds of commonly used high throughput molecular data types and signatures.

**Clustering analysis**

Once the molecular features are selected, *iGPSe* offers a Clustering section which performs the stratification and help users evaluate and refine the clustering results. The Clustering section provides an interactive interface for tuning parameters of clustering algorithms. We provide users three clustering methods: K-means, spectral clustering, and community detection.

Following the stratification, *iGPSe* generates the heatmap, silhouette plots and population graph visualization to help user evaluate the validity and quality of the stratification result. *iGPSe* also provides visualization for an overview of demographic
and clinical information such as patients’ age distribution and tumorous grades. A Clustering analysis example for the TCGA BRCA dataset using afore-mentioned features is shown in Figure 4.6.

**Integrative patient stratification**

Integrative patient stratification section allows users to review the clustering results and provides an interactive interface to compare clinical outcomes, such as patient survival and patients ages/tumour grades. Users can select subgroups from the parallel sets to carry out survival analysis including Kaplan-Meier curves and a log-rank test. An example is shown in Figure 4.1

- **Parallel sets view** shows clustering results of patients using different feature sets which are arranged as columns side-by-side. The columns are split up into disjoint blocks representing clusters. Ribbons connect blocks of two columns, whose width represents how many patients shared between the two clusters.

- **Survival plot view** shows the Kaplan-Meier plots (survival curves) of selected subgroups. If more than one groups is selected, the p-value which tests the null hypothesis that the survival curves are identical in the overall populations using the log-rank test, will be given. Users can select the subgroups by clicking the blocks or ribbons in the parallel sets view.

- **Population graph view** provides an interactive similarity graph visualization. User can drag the graph to adjust the layout, and select a node to acquire correspondence patient’s clinical information.
The application also allows users to export the visualization into high quality figures. The users can save a figure to selected format using the ‘Save Figures’ button. Currently we support PDF, SVG, JPG, and PNG formats.

4.4 Results

4.4.1 Use Case Studies

We demonstrate the use of iGPSe on a patient cohort obtained from the TCGA invasive breast carcinoma (BRCA) project. The TCGA BRCA dataset contains 623 patients’ mRNA and miRNA expression profiles as well as clinical outcome information. One of the major advantages of iGPSe is that it enables the user to select genes/miRNAs for use as features to study the patient population.

We are especially interested in the relationship between a group of cell cycle/genome stability genes and the well-known mir-17-92 cluster of miRNAs. In a previous study, we identified a group of more than 400 genes which are frequently co-expressed in multiple types of cancers but not in normal tissues [145]. This gene group is highly enriched with genes involved in cell cycle, mitosis, and genome stability maintenance. It also includes the well-known breast cancer gene panel, PAM 50 [2], which purports to possess prognostic capabilities. The mir-17-92 cluster contains a group of six miRNAs which are proximal on human chromosome 13 and often co-express. They have been shown to be involved in the progression of many types of cancers including lymphoma [146]. Therefore it is of interest to explore if there is any relationship between the frequently co-expressed gene and miRNAs groups in the patient populations and if a combination of them can lead to better prognosis of the breast cancer patients.
Since the data has been log transformed we set Euclidean distance as the similarity function in the preview section. By selecting an appropriate threshold value we get the population graph as in Fig. 4.1 (left graph is constructed from mRNA, right graph is constructed from miRNA). We chose K-means as the clustering method to separate patients mRNA and miRNA metric into four and three subgroups, respectively.

Figure 4.1 shows the parallel view and the network view after the K-means clustering of the patients. In the graph visualization, one can note that the population stratifications suggested by the mRNA and miRNA display very different subtyping of patients. In the mRNA graph, patients in the green and brown groups are strongly separated from the orange group, while in the miRNA graph, members in the yellow and grey groups are relatively mixed.

With iGPSe using the parallel sets in the interactive visualization stage, we associate the clustering results with clinical outcomes for the two selected subgroups. We also compare other subgroups from one or more data modalities. Figure 4.7(a) shows the survival analysis for different choice of subgroups from mRNA only and miRNA only (Figure 4.7(b)). We can observe that in both mRNA and miRNA exist one subgroup has very different survival times than the others. A similar observation was reported by others [147, 2]. Interestingly, the two patient stratifications are not independent. The blue cluster in mRNA stratification is dominated by samples from grey and green clusters in the miRNA stratification. This observation suggests that combination of mRNA and miRNA features can improve the patient stratification. As shown in Figure 4.7(c), we can quickly test the choice of other combinations and the result shows a much more significant difference in survival times than just using any single data type to stratify the patient population.
4.4.2 Users’ Feedback

We collected comments from three domain experts (Biostatistician, clinician and basic scientist) on the use of \( iGPSe \), providing some evidence that our visualization and interaction design, grounded in a characterization of the domain requirements, supported efficient explorations of subtypes. We prepared datasets for the TCGA breast invasive carcinoma (BRCA) cohorts and invited our domain experts to perform population stratification using \( iGPSe \). The initial user feedbacks on the utility of \( iGPSe \) are positive.

Before we present the feedbacks and evaluation details, we outline typical analysis workflow in genomics based cancer patient stratification. As we stated earlier, a standard genomic based cancer patient stratification workflow includes steps of data curation and pre-processing, clustering, subtype characterization and validation. Traditionally, these steps involve scripts written in different platforms which requires researchers possess at least a basic level of programming skills. Using separate static scripts makes the analysis of large populations with multiple modalities laborious, in particular when studying the interactions between different data types. The design of \( iGPSe \) was directly motivated by these shortcomings and the following observations reflect how domain scientists used \( iGPSe \) to analyze the model described in previous section.

- **Feature Selection** – The identification of discriminant genes is of fundamental and practical interest. A small subset of highly discriminant genes can be extracted to build more reliable cancer classifiers. For users who do not have experience in writing scripts to process the input data, \( iGPSe \) interactive GUI offers a more convenient way than writing scripts to select the target genes list.
A collaborator mentioned that scroll down table of genes and input genes can help users adding and dropping genes from the selected list effectively, especially for people who do not have a programming background.

- **Clustering section** – The general aim of the clustering section is clustering patient population into subgroups and examine the results’ quality for each data type. Our collaborators agreed that the interactive clustering parameters control panel allows users who have no prior knowledge of machine learning to guide the stratification process. In addition, the graph visualization of the patient population gives an overview of the population density in the selected feature space which helps users to verify the clustering results are consistent with the proximity graph. The silhouette plot of the stratification gives a rapid overview of the quality of the clustering result. This immediate visual access makes the judgment of simulation output faster and more intuitive. Moreover, all views can be compared across runs simply by launching the visualization multiple times. The traditional workflow would have forced scientists to manually run scripts.

- **Integrative analysis** – our collaborators noted that parallel set visualization helps them to explore the relationship between stratification results from two different types of data. They also told us that the interactive combination of two types of data stratification results with the parallel sets is very useful, especially interactively generating the survival plot. Eventually experts are able to answer more challenging questions such as whether patient groups which share similar
gene expression have very different clinical outcomes when they are expressed differently with the miRNA.

In general, very positive outcome of the evaluation sessions with our collaborators was that in all cases they asked us to load further data to explore with iGPSe. This is convenient especially for testing or replicating reported gene lists.

4.5 Summary

Visual analytics is an emerging discipline that combines visualization methods with data analysis and human-computer interaction. As shown in this chapter, application of visual analytics methods in integrative genomics can enable quick integration of different types of data and significantly facilitate the discovery of integrated molecular markers for cancer subtyping and outcome prediction.

In our case study, it is observed that the two groups of patients with different gene (mRNA) expression profiles for the previously identified genome stability genes show differences in survival times only when they have similar specific expression profiles for the miRNAs in the mir-17-92 cluster. This observation suggests that the genome stability genes and mir-17-92 cluster may influence breast cancer development and progression in different pathways even though the genes such as PAM50 and mir-17-92 cluster interact with each other. Thus it is of great interest to study the genes and pathways targeted by the mir-17-92 clusters in order to elucidate the different mechanisms.

While we have demonstrated the functionalities of iGPSe using only mRNA and miRNA signatures, it nevertheless can accommodate other types of patient information such as genome-specific information, DNA methylation, and even morphological
features extracted from pathological images. In addition, the system can accommodate comparison among more than two subgroups as well as more than two types of data, which makes \textit{iGPS}e highly versatile for biomedical researchers to use and generate highly interpretable results much more promptly than the cumbersome script-based approach.

Our evaluation with domain experts shows that major strengths of \textit{iGPS}e is that it eliminated the needs of programming and scripting from users while still grant users sufficient control during the steps including feature selection, clustering, subgroup selection and comparison. The automatic comparison of clinical outcome (i.e., survival) is of particular interests to the users.

Currently a we are developing a publicly available web-based online platform, OSU Multi-Omics analysis (OSUMO) which integrates all functions from \textit{iGPS}e. With OSUMO, users will be able to upload and analyze their own data. In addition, more choices on the clustering algorithms will be implemented. From the machine learning point of view, our approach provides an alternative way of carrying out consensus clustering in order to reconcile different clustering results from different features. This system can be combined with existing consensus clustering approaches to further streamline the subgroup selection process. Since graph visualization allows interactive manipulation of the data points, a feedback mechanism for interactively assigning clustering membership based on visualization can be deployed to enable iterative feature selection.

Overall, we have demonstrated that by combining graph visualization with parallel views and bioinformatics analysis, we can significantly reduce the computing burden for biomedical researchers in order to explore the complicated integrated genomics
data. This approach is generalizable to enable more sophisticated analysis for cancer biomarker discovery and subtyping in order to achieve precision medicine.
Figure 4.5: Feature selection and input panel in *iGPSe*. Left: The user can select a list of genes (for mRNA data) from the left list (extracted from input files) by left clicking on the genes or input the gene list (e.g., copy and paste) into the window on the right side. The user can also load gene/miRNA list from text files. The selected genes will also show up in this window. Right: The selection and input for miRNA is similar to that for genes.
Figure 4.6: The clustering analysis page.
Figure 4.7: Comparisons of different choices of subgroups. (a) Compare the blue and yellow/purple/pink groups based only on mRNA data. (b) Compare the grey/green and light blue groups based only on miRNA data. (c) Two groups with different mRNA profiles and different miRNA profiles. (d) Two groups with different mRNA profiles but similar miRNA profiles.
Chapter 5: Integrative cancer patient stratification through subspace merging

In the previous chapter, we present \textit{iGPSe}, a visual analytics tool for integrative cancer patient stratification using multi-omics data. Despite many advantages of \textit{iGPSe}, it is challenging to use \textit{iGPSe} integrate more than two types of omics data due to the limitations of parallel set visualization. Moreover, the clustering analysis before the integration may also cause loosing data type specific information. In this chapter, we try to overcome these challenges by proposing a novel transformed-based integration methods to perform patient stratification with multi-omics data. Our proposed approach creates a patient-to-patient similarity graph for each data type as an intermediate representation of each omics data type and merges the graphs through subspace analysis on a Grassmann manifold. We applied our approach to a TCGA breast cancer data set and show that by integrating gene expression, microRNA, and DNA methylation data, our proposed method can produce potentially clinically useful subtypes of breast cancer. Further, we then investigate the molecular characteristics underlying these subtypes. We discover a highly expressed cluster of genes on chromosome \textbf{19p13} that strongly correlates with survival in TCGA breast cancer patients and confirm these results in three additional breast cancer datasets.
We also compare our approach with previous integrative clustering approaches and obtain comparable or superior results.

5.1 Introduction

As discussed in the earlier chapters, the boundaries of personalized medicine in cancers has been pushed by the large consortium efforts of TCGA and the ICGC. The prior understanding of tumor subtypes based on histology and immunohistochemical markers has been complemented by the vast amount of available high-throughput molecular data. While many of the previous classifications based on clinical attributes are still widely adopted, methods that leverage the ‘omics data produced by studies such as TCGA and ICGC enable the development new clinically useful biomarkers and investigation in tumor biology simultaneously. Different types of ‘omics data can be used to stratify tumors biologically and clinically; however, doing so is not without its hurdles. A major challenge is how to effectively integrate all of the diverse data types, such as microRNA, mRNA, and DNA methylation, available for each tumor, to achieve clinically meaningful patient stratification and tumor subtyping.

Methods used to leverage multiple sources of high dimensional biological data fall into three broad categories: concatenation-based integration, model-based integration, and transformation-based integration [1]. Methods belonging to the concatenation-based integration class combine all data types into one large single matrix and perform analytical methods directly on the combine matrix [148, 62]. In model-based integration, separate models are applied independently to each data type and the integration is conducted by assembling the results at the end [15, 149]. Transformation-based integration combines multiple data types after transforming each data type into an
intermediate representation (e.g. a graph or a kernel matrix) [129, 68, 70, 150]. For more literature reviews of integrative analysis, see Section 2.

Here, we present a subspace merging algorithm for integrative cancer patient stratification from multiple types of data. Our proposed approach falls into the transformation-based integration paradigm. Given multiple types of ’omics data for the same set of patients, we create a patient-to-patient similarity graph for each data type as an intermediate representation and integrate the graphs through subspace merging on a Grassmann manifold. This approach avoids problems created by using data types with different numbers of measurements with different scales, as is the case in multi-modal genomic data. The resulting combination can then be viewed as a lower dimensional representation of the original data. Finally, we cluster the patients on this lower-dimensional subspace to identify potential tumor subtypes. Subspace merging on a Grassmann manifold has been previously studied in the context of computer vision and wireless communication [151, 152] but has not been adopted for integrating genomic data.

Our approach has several advantages over the concatenation-based and model-based integration: (i) intermediate representation preserves data-type-specific properties; (ii) the transformation-based integration approach is robust to different data measurement scales; (iii) this approach can be used to integrate many types of data, including both continuous and categorical values, as long as different data types measure the same population; (iv) unlike previous approaches, such as Similarity Network Fusion (SNF) [70] and affinity aggregation for spectral clustering [150], our method does not require iterative optimization. This leads to a more stable and computationally efficient solution. We have extensively tested our method on multiple datasets
and compare with the recently published SNF algorithm [70] with better or comparable results in separating patients into groups with significantly different prognosis.

We further carried out an in-depth study on breast cancer by applying our method to the TCGA breast cancer tumor samples with matched RNA, microRNA and DNA methylation, and clinical follow-up data. There is a rich history of dividing breast cancer into clinically useful subtypes. RNA expression panels such as PAM50 [153] and OncotypeDX [154] use selected genes to separate patients into groups with distinct prognosis. The success of the in subtyping breast cancers, in part, likely reflects the fact that luminal type breast cancers and triple negative breast cancers are molecularly distinct cancers that happen to develop in the same anatomical location. OncotypeDX, on the other hand, uses measurements of 21 genes to calculate a recurrence score, which helps physicians and patients decide whether or not to use adjuvant chemotherapy. We show that by integrating gene expression, microRNA, and DNA methylation data, we produce clinically useful subtypes of breast cancer. We further investigated the molecular basis underlying these subtypes. Specifically, we find that a group of patients with excellent prognosis shows overexpression of a large group of genes located on chromosome 19p13. We further validated these findings on three additional independent breast cancer datasets. While many previous studies focused on identifying marker genes for poor prognosis for breast cancers, our finding with an exceptionally strong signal on chromosome 19p13 suggests a new marker to identify patients with good prognosis using integrative analysis.
Figure 5.1: Workflow for integrating and merging cancer genomics datasets. A patient-to-patient similarity graph is constructed for each data type. The similarity matrices are converted to subspaces and embedded in a Grassman manifold, where they are integrated into a single, representative subspace. This subspace is then clustered to obtain the final integrative patient groups.
5.2 Methods

5.2.1 Method overview

Given multiple types of 'omics data for the same set of patients, we first create a patient-to-patient similarity graph for each data type (Fig. 5.1). Here, the similarity is defined by heat kernel to measure the distance between patients. In our breast cancer case study, this means that three unique patient-to-patient graphs are created, one each for microRNA, methylation, and mRNA expression data. Prior to any further analysis, we remove edges with low similarity measures, which represent uncertain relationships between patients. To perform the integrative analysis, the differences between these graphs must be reconciled. Prior studies employ an iterative convergence approach [70]; in our method, we merge the graphs in two steps through a mathematical construct, the Grassmann manifold.

In the first step, the structure of each patient-to-patient similarity graph is captured by a subspace representation through spectral embedding [155]. Since each graph is reduced to its low-dimensional subspace representation, computation is minimal, and noise is reduced [156]. In the second step, each subspace representation is considered as a point on a Grassmann manifold [151, 152]. Therefore, we can find a new representative subspace on Grassmann manifold where the overall distance between new representative subspace and the individual subspaces is minimized. The resulting of this analysis is an integrative subspace that summarizes the desired merged graph, containing information from all data types.

Finally, we clustered the patients in the integrative subspace and conducted a post-hoc analysis to evaluate the clustering results. We hypothesize that such a method will generate more informative clusters by preserving the complementary
information from each level of ’omics data. The details of our methods is described in Section 5.2.3.

5.2.2 Data set and Data preprocessing

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n=441</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26-45 years</td>
<td>54</td>
<td>12.24%</td>
</tr>
<tr>
<td>46-55 years</td>
<td>127</td>
<td>28.80%</td>
</tr>
<tr>
<td>56-65 years</td>
<td>103</td>
<td>23.36%</td>
</tr>
<tr>
<td>older than 66 years</td>
<td>123</td>
<td>27.89%</td>
</tr>
<tr>
<td>Disease Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>133</td>
<td>30.16%</td>
</tr>
<tr>
<td>T2</td>
<td>232</td>
<td>52.61%</td>
</tr>
<tr>
<td>T3</td>
<td>63</td>
<td>14.29%</td>
</tr>
<tr>
<td>T4</td>
<td>12</td>
<td>2.72%</td>
</tr>
<tr>
<td>ER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive(+)</td>
<td>335</td>
<td>75.96%</td>
</tr>
<tr>
<td>Negative(-)</td>
<td>100</td>
<td>22.68%</td>
</tr>
<tr>
<td>PR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive(+)</td>
<td>299</td>
<td>67.80%</td>
</tr>
<tr>
<td>Negative(-)</td>
<td>134</td>
<td>30.39%</td>
</tr>
<tr>
<td>HER2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive(+)</td>
<td>43</td>
<td>9.75%</td>
</tr>
<tr>
<td>Negative(-)</td>
<td>239</td>
<td>54.20%</td>
</tr>
</tbody>
</table>

The data used in this chapter is TCGA level 3 data containing gene expression, miRNA expression and DNA methylation expression profiles from 441 primary tumors of breast cancer patients. Multiple platforms for each data type are available in TCGA. We chose UNC-Illumina-Hiseq-RNASeq platform for gene expression, BCGSC-Illumina-Hiseq-miRNAseq platform for miRNA expression and JHU-USC-Human-Methylation-450k platform for DNA methylation expression profiles. Pertinent clinical data was also available for all of these patients. The minimum follow-up duration is 3 months (91 days), and the median follow-up duration is 16 months (492
days). The demographic information of patient samples is listed in Table 5.1. We performed the following normalization for each data type.

\[
\hat{f} = \frac{f - E(f)}{\sqrt{Var(f)}}
\]  

(5.1)

Where \(f\) is a feature any data type, \(\hat{f}\) is the corresponding feature after normalization, and \(E(f)\) and \(Var(f)\) represent the sample mean and sample variance of \(f\), respectively.

5.2.3 Integrative clustering

Construction of the patient-to-patient similarity graph

Consider \(M\) types of omics data measurements \(\{X^m\}_{m=1}^M\) (each with dimension of \(p_m\)) collected from \(N\) patient samples, such that \(X^m\) is a \(p_m \times N\) matrix. For each data type \(X^m\), we construct a patient-to-patient similarity graph \(G^m\) to model the local neighborhood relationships between the samples. Let \(G^m = (V^m, E^m, W^m)\) denote a patient similarity graph for data type \(m\), where \(V^m\) represents the vertex set, \(E^m\) represents the edge set, and \(W^m\) represents the adjacency matrix.

The adjacency matrix \(W^m\) of the graph \(G^m\) is a symmetric matrix whose entry \(w^m_{ij}\) represents the edge weight if there is an edge between vertex \(v_i\) and \(v_j\), or 0 otherwise. To construct this similarity graph, we first compute a similarity matrix to measure the pairwise similarity between each sample pair. Here, we use a heat kernel as the similarity metric:

\[
S^m_{ij} = exp \left( -\frac{\|x^m_i - x^m_j\|^2}{2t^2} \right), \ i = 1, ..., N, \ j = 1, ..., N
\]  

(5.2)
We then extract $k$-nearest neighbors ($k$-NN) graph from the similarity matrix $S^m$:

We denote $N_i$ as a set of $v_i$’s neighbors including $v_i$ and size of $N_i$ is $k$. The number of $k$ normally depends on the sample size. We then connect $v_i$ and $v_j$ with an undirected edge with edge weight as $S_{ij}$ if $v_j \in N_i$. (Eq. 5.3).

$$W^m_{ij} = \begin{cases} 
S^m_{ij}, & \text{if } v_j \in N_i, \\
0, & \text{otherwise.}
\end{cases} \quad (5.3)$$

Essentially we make the assumption that local similarities are more reliable than remote ones. This is a mild assumption widely adopted by other manifold learning algorithms.

**Construction of subspace representation**

For each similarity graph $G^m$ constructed from data type $m$, we first compute the normalized graph Laplacian matrix $L^m$, which is defined as $L^m = D^{-\frac{1}{2}}(D^m - W^m)D^{-\frac{1}{2}}$, where $W^m$ and $D^m$ denote the adjacency matrix and degree matrix of $G^m$ respectively.

Once obtain the normalized graph Laplacian matrix from each data type, we conduct the graph embedding to construct the subspace representation for each graph. The main purpose of graph embedding is to find a low-dimensional subspace that preserves similarities between the vertex pairs. In other words, the resulting subspace best captures the characteristics of each data types. We use $U^m$ to denote the subspace representation of $G^m$. The optimal graph embedding in $k$ dimension is derived by minimizing the following objective function:

$$\min_{U^m \in R^{N \times k}} \text{tr}(U^m'L^mU^m), \quad \text{s.t. } U^m'U^m = I \quad (5.4)$$
By the Rayleigh-Ritz theorem [155], the solution to the problem of Eq. 5.4 is given by the first $k$ eigenvectors of the Laplacian $L^m$, which can be computed using efficient algorithms for eigenvalue problems.

Subspace merging on Grassmann manifold

With the subspace representations $U^M_{m=1}$ construct from each data type, we then merge them on a Grassmann manifold. The Grassmann manifold is defined as a set of linear subspaces of a Euclidean space, therefore each subspace representation. We use $G(k, n)$ to denote a Grassmann manifold with $k$ dimensional linear subspaces in $n$ dimensional Euclidean space $R^N$. An element of $G(k, n)$ can be represented by an orthonormal matrix $Y \in R^{n \times k}$ whose columns span the corresponding $k$ dimensional subspace in $R^n$; it is thus denoted as $\text{span}(Y)$. A distance between two subspaces is defined as the length of the shortest geodesic connecting the two corresponding points on the Grassmann manifold. However, there is a more convenient and efficient way of defining distances using the projection distance [157]. For instance, the projection distance between $\text{span}(Y_1)$ and $\text{span}(Y_2)$ ($Y_1, Y_2 \in n \times k$) is defined as follow:

$$d^2_{\text{proj}}(Y_1, Y_2) = \sum_{i=1}^{k} \sin^2 \theta_i = k - \sum_{i=1}^{k} \cos^2 \theta_i = k - \text{tr}(Y_1 Y_1' Y_2 Y_2')$$

With the distance measurement defined above, we can capture the similarity between the subspaces on Grassmann manifold and subsequently enable us to merge the information from multiple graphs in a meaningful manner. In the last section, we constructed subspace representations $\{U^m\}_{m=1}^M$ for $M$ data type by the conduct

98
the spectral embedding on the patient-to-patient similarity graphs \( \{G^m\}_{m=1}^M \). Each subspace representation \( \{U^m\} \) defines a \( k \)-dimensional subspace in \( R^n \), where \( n \) is the number of patients and \( k \) is the target number of clusters. To merge these subspaces, we want to find an integrative subspace \( \text{span}(U) \), which is close to all the individual subspaces \( \text{span}(U^m) \), and at the same time the representation \( U \) preserves the vertex connectivity in each \( G^m \).

With the distance measurement defined in Eq. 5.5, we can define a summation of projection distance between the integrative subspace \( U \) and \( M \) subspaces \( \{U^m\}_{m=1}^M \) as follow:

\[
d_{\text{proj}}^2(U, \{U^m\}_{m=1}^M) = \sum_{m=1}^M d_{\text{proj}}^2(U, U^m)
= \sum_{m=1}^M [k - \text{tr}(UU^mU^m')] = kM - \sum_{i=1}^M \text{tr}(UU^mU^m')
\]

(5.6)

The subspace \( U \) which minimizes Eq. 5.6 is close to all the individual subspaces \( \{U^m\}_{i=1}^M \) in terms of the projection distance on the Grassmann manifold. Since we also want \( U \) to preserve the vertex connectivity in graphs from each data type. Therefore, we finally propose to merge multiple subspaces by solving the following optimization problem that integrates Eq. 5.4 and Eq. 5.6:

\[
\min_{U \in R^{n \times k}} \sum_{m=1}^M \text{tr}(UU^mU^m') + \alpha[kM - \sum_{m=1}^M \text{tr}(UU^mU^m')], \text{ s.t. } U'U = I
\]

(5.7)
where $L^m$ and $U^m$ are the graph Laplacian and the subspace representation for $G^m$, respectively. Such a representation not only preserves the structural information contained in the individual graph, which is encouraged by the first term of the objective function in Eq. 5.7, but also keeps a minimum distance between itself and the multiple subspaces, which is enforced by the second term. The regularization parameter $\alpha$ balances the trade-off between the two terms in the objective function.

By ignoring constant terms and rearranging the trace form in the second term of the objective function, Eq. 5.7 can be rewritten as:

$$\min_{U \in \mathbb{R}^{n \times k}} \text{tr}[U'(\sum_{i=1}^{M} L^m - \alpha \sum_{m=1}^{M} U^m U^m')U], \text{ s.t. } U'U = I \quad (5.8)$$

The Eq. 5.8 shares the same form with Eq. 5.4, and the solution to the problem of Eq. 5.8 is then the first $k$ eigenvectors of the modified Laplacian $L_{mod}$ in Eq. 5.9, followed by the Rayleigh-Ritz theorem [155].

$$L_{mod} = \sum_{m=1}^{M} L^m - \alpha \sum_{m=1}^{M} U^m U^m' \quad (5.9)$$

Finally, we can cluster resulting integrative subspace $U$ using the k-means algorithm.

5.3 Results

5.3.1 Comparison with results from similarity network fusion (SNF)

To demonstrate the effectiveness of our methods, we first applied our method on the data set analyzed by Wang et al. [70] and compared our results to those generated by their similarity network fusion (SNF) method. The cancer types include
Table 5.2: Comparison of Cox survival p-values from integrative clustering on a grassman manifold with those from SNF

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>SNF (nature.2014)</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM (3 clusters)</td>
<td>$2.0 \times 10^{-4}$</td>
<td>$4.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>BIC (5 clusters)</td>
<td>$1.1 \times 10^{-3}$</td>
<td>$2.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>KRCCC (3 clusters)</td>
<td>$2.9 \times 10^{-2}$</td>
<td>$2.8 \times 10^{-3}$</td>
</tr>
<tr>
<td>LSCC (4 clusters)</td>
<td>$2.0 \times 10^{-2}$</td>
<td>$1.6 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

glioblastoma multiforme (GBM, N = 213), breast invasive carcinoma (BIC, N = 105), kidney renal clear cell carcinoma (KRCCC, N = 122) and lung squamous cell carcinoma (LSCC, N = 106). The Cox survival $P$ values reported by [70] and our method are listed in Table 5.2. In order to ensure comparability of the results, we inherit [70]'s choice of cluster numbers for each cancer type. As can be seen in the table, in three out of the four types of cancer our method provide more significant differences between the survival times. In GBM, the P-value is comparable to the p-value generated by SNF. Survival plots for GBM, BIC, KRCCC, and LSCC tumors are shown in Fig. 5.2.

5.3.2 Identification of integrative subtypes of Breast Cancer

For decades, researchers have been developing methods to subtype breast cancer. Gene signatures, such as PAM50 [153] and the 70-gene signature [120] have been developed to subtype breast cancer patients. However, subtyping based on other data sources, for example microRNA and DNA methylation lead to incoherent results when compared to subtyping done using mRNA data [158, 159].

In this study, we obtained DNA methylation, mRNA expression and miRNA expression data from 441 primary tumors of breast cancer patients from TCGA [132].
Figure 5.2: Survival plots of integrative clusters for BIC (a), KRCCC (b), LSCC (c), and GBM (d). P-values are computed from the log rank test.

Detailed clinical information for this cohort is listed in Table 5.3. We applied the proposed method to this dataset and partitioned the patient population into five integrative subtypes. The choice of number of cluster is determined by two fold of reasons: (i) the silhouette scores reached peak when the number of cluster is five (Figure. B.3); (ii) number of well-accepted breast cancer subtypes is also five.

Fig. 5.3(a-c) show the adjacency matrices and their corresponding representative subspace for the patient-to-patient similarity graphs from each data type. As shown in the plot, the connectivity of the graphs vary considerably from each other. For
instance, **subtype 2(red)** has high inner connectivity with the graph generated by miRNA and DNA methylation, while the graph constructed with mRNA data better supports the connectivity in **subtype 4(blue)**. Fig. 5.3(e-f) compares clusters obtained from integrative vs. single-data-type analyses. Fig. 5.3a shows the overall survival plot produced by integrative clustering, with an estimated p-value < 0.0001 (log rank test). While analyses of mRNA (Fig. 5.3b) and miRNA (Fig. 5.3c) data independently yielded significant clustering results (log rank test p = 0.0231 and 0.0286, respectively), overall survival of the clusters is much more clearly separated using integrative clustering. Supporting this finding, the subspace representations shown in Fig. 5.3(a-d) are most highly separated in the integrative cluster, shown in Fig. 5.3d. The clustering produced from methylation data (Fig. 5.3d) was not significantly prognostic (p = 0.2461).

**Subtype 1 (black)** largely contains patients negative for progesterone(PR), estrogen(ER), and epidermal growth factor 2 (HER2) receptors, also known as triple negative (see Table 5.3). The vast majority of tumors in **subtype 3 (green)** and **subtype 4 (blue)** are positive for ER and PR, but patients in **subtype 3** have a clear survival advantage over patients in **subtype 4**. Similarly for **subtype 2** and **subtype 5**, they have similar ER, PR, and HER2 statuses, yet different prognoses. These differences may represent useful prognostic and, more importantly, therapeutic opportunities. To investigate the molecular basis of our subtyping, we further analyzed the differences in mRNA, microRNA, and methylation abundances.
### Table 5.3: Clinical attributes of TCGA breast cancer subtypes

<table>
<thead>
<tr>
<th>Subtype</th>
<th>N</th>
<th>ER, (+)</th>
<th>ER, (-)</th>
<th>PR, (+)</th>
<th>PR, (-)</th>
<th>HER2 (+)</th>
<th>HER2 (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtype 1</td>
<td>83</td>
<td>23 (27.7%)</td>
<td>59 (71.1%)</td>
<td>19 (22.9%)</td>
<td>63 (75.9%)</td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>Subtype 2</td>
<td>76</td>
<td>55 (72.4%)</td>
<td>19 (25%)</td>
<td>51 (67.1%)</td>
<td>23 (30.3%)</td>
<td>9</td>
<td>41</td>
</tr>
<tr>
<td>Subtype 3</td>
<td>103</td>
<td>100 (97.1%)</td>
<td>3 (2.9%)</td>
<td>88 (85.4%)</td>
<td>15 (14.6%)</td>
<td>11</td>
<td>50</td>
</tr>
<tr>
<td>Subtype 4</td>
<td>111</td>
<td>103 (92.8%)</td>
<td>6 (5.4%)</td>
<td>95 (85%)</td>
<td>13 (11.7%)</td>
<td>12</td>
<td>62</td>
</tr>
<tr>
<td>Subtype 5</td>
<td>68</td>
<td>54 (79.4%)</td>
<td>13 (19.1%)</td>
<td>46 (67.6%)</td>
<td>20 (29.4%)</td>
<td>7</td>
<td>40</td>
</tr>
</tbody>
</table>

### 5.3.3 Molecular basis of integrative breast cancer subtypes

We have shown that our method produces subtypes with clinically relevant prognoses, yet of equal importance are the molecular alterations associated with these differences in prognoses. To assess the biological significance of our clustering results, we look for differentially activated groups of nodes in an integrative co-expression network, where each node is an mRNA gene, methylation site, or microRNA species (See Online Methods). Our first observation was that **subtype 1**, clinically identified as the TNBC subtype, indeed contains the molecular hallmarks previously associated with TNBC. For example, we observed that MYBL2, CENPA, AURKB and KIF2C are all overexpressed in **subtype 1** (p-value < 0.001, FDR-corrected) which is consistent with previous results on the TNBC subtype [160](Figure B.1). Also consistent with prior studies, patients in **subtype 1** have relatively poor overall survival.

Of particular interest are the molecular alterations that might cause tumors with similar receptor status to have differing prognoses. For example, **subtype 3** and **subtype 4** have nearly identical receptor statuses(Table 5.3), yet in contrast to **subtype 4**, no patients in **subtype 3** die during the follow-up period. In our coexpression
network we noticed that there was a large, highly coexpressed module of genes located on chromosome 19p13 (Fig. 5.4a). Remarkably, a large subset of these genes were overexpressed in subtype 3 compared to other groups. We found that chromosome 19p13 gene expression can clearly separate patients in the TCGA dataset into good and poor prognoses (Fig. 5.4b). To validate our findings, we further tested our 19p13 signature on the TCGA BRCA dataset as well as three additional publicly available validation datasets (NKI [120], Fig. 5.4c, GSE3141 [161], Fig. 5.4d, and GSE1456 [162], Fig. 5.4e). Of note, all four datasets were created using different expression platforms, which shows that our signature is platform independent. In addition, it has been reported that low SAFB levels are associated with worse outcome in breast cancer patients [163].

We further sought to investigate possible molecular alterations associated with chromosome 19p13 overexpression. We first hypothesized that copy number changes in chromosome 19p13 could explain the expression differences; however, an analysis of the corresponding CNV data showed that only a minority of the variation is explained by deletions or amplifications on chromosome 19p13 (Figure. B.2). Next, we searched the methylation data for an explanation, but there was no clear upregulation or downregulation of methyl sites located on chromosome 19p13. Further experiments are necessary to identify a single gene, microRNA, or methylation site responsible for the observed changes in chromosome 19p13 expression. This search is complicated by the fact that numerous known tumor suppressors and oncogenes are located in this region. Another possible explanation is a large chromosomal event that is not easily detected through high throughput sequencing events.
Figure 5.3: Integrative clustering of breast tumors produces prognostically relevant and biologically significant groupings. (a-c) The adjacency matrices of patient-to-patient similarity graphs, produced from mRNA (a), microRNA (b), and methylation (c) datasets. (d) Integrative clustering of patients using all three datasets. Color bars at left show the clusters of patients, and the heatmap to the left of each colorbar shows the eigenvector clustering results. (e-h) Survival analysis of patient stratification results using integrative and single-data-type clustering methods. Kaplan-Meier survival curves of clusters produced by: (e) integrative clustering, (f) Gene expression alone, (g) miRNA expression alone, (e) DNA methylation alone, listed along with estimated p-values (log-rank test).
Figure 5.4: Density of differentially expressed genes on chromosome **19p13** are shown in (a). (b-e) Survival curves and clustered heatmaps produced by separating four datasets based on expression of genes on chromosome **19p13**. Kaplan-Meier survival curves, clustered heatmaps, and accompanying p-values (log-rank test) generated from the training dataset from TCGA are shown in (b). Results generated from three validation sets are shown in (c-e): Netherlands Cancer Institute (NKI) in (c), GSE3143 in (d), and GSE1456 in (e).
5.4 Summary

In this chapter, we propose a novel method to perform efficient integrative patient stratification. Our approach aggregates information from multiple molecular expression data through a subspace analysis on the Grassmann manifold. We applied our method to stratify the breast cancer patient cohort datasets collected from The TCGA by integrating gene expression, DNA methylation and miRNA expression data. The result demonstrates our method can leverage information from different omics data into clinically relevant subtypes. Also, through our subtyping results, we can uncover a group of genes located on chromosome 19p13 with strong prognostic power. We further validate our finding on three independent datasets. Further follow-up studies on this gene set are necessary to reveal their biological implications.

Since the integration step in our method is independent of the properties of data source, the input data types are not limited to genomic data. With appropriate similarity measurements, our method can be extended to applications in which integration of clinical categorical information and image datasets for which clustering is needed. Nonetheless, our approach can also be applied to other tasks that require integration of multiple types of features.
Chapter 6: Conclusion and Future Works

Multi-omics data analysis plays an essential role in both basic research and clinical practices in cancers. Through the in-depth analysis of omics data generated from different biological levels, researchers can better comprehend the heterogeneous nature of cancer. Furthermore, the integrative analysis which utilizes multiple types of omics data can shed insights on the etiology of complex traits. There is a need for new tools and methods that allow researchers to efficiently discover deeper knowledge from multi-omics data. Nonetheless, the increasing heterogeneity and complexity of multi-omics data pose significant challenges to developing effective integrative analysis methods to identify novel integrative biomarkers and subtypes.

This dissertation aims to develop tools and applications to overcome the aforementioned challenges. Specifically, this dissertation achieved the following three tasks:

**Aim. 1** Develop user-friendly tools that target efficient navigation of large omics datasets and identification of outcome-related features.

**Aim. 2** Develop tools that allow a user to quickly carry out integrative analysis on multi-omics data and identify potential combined markers.

**Aim. 3** Develop efficient computational methods to perform integrative analysis on multi-omics datasets.
Chapter 3 presents GRAPHIE, a visual analytics application designed to explore the histology image collection. By taking a data-driven approach, we developed an unbiased way for visualizing the entire collection. GRAPHIE not only provides an intuitive overview of the data but also enables users to use domain knowledge to improve the visualization through interactive feature selection. The visualizations and interactions of GRAPHIE are seamlessly integrated to allow users to effectively explore and annotate images, with a rich set of interactive functions. The use of GRAPHIE was evaluated with two datasets.

Chapter 4 presents iGPSe, a visual analytic tools, which enables quick integration of multi-omics data and significantly facilitate the discovery of integrated molecular markers for cancer subtyping and outcome prediction. iGPSe integrates unsupervised clustering with graph and parallel sets visualization and allows a direct comparison of clinical outcomes via survival analysis. Our evaluation shows that the major strengths of iGPSe is that it give users sufficient control during feature selection, clustering, subgroup selection and comparison.

Chapter 5 proposes a novel method to perform efficient integrative patient stratification. Our approach aggregates information from multiple molecular expression data through a subspace analysis on the Grassmann manifold. We applied our method to stratify the breast cancer patient cohort datasets collected from The TCGA by integrating gene expression, DNA methylation and miRNA expression data. The result demonstrates our method can leverage information from different omics data into clinically relevant subtypes. Also, through our subtyping results, we can uncover a group of genes located on chromosome 19p13 with strong prognostic power.
The work described in this dissertation also leaves considerable room for future research. Here we list some of the drawbacks and future directions for improvement:

- Improvements to the visual analytics tools. Through use case studies, both of our tools show promising results on the effectiveness of exploring multi-omics data. However, the current prototype implementation suffers from a lack-of-scalability, the size of the data set that can be meaningfully explored is limited. We believed this problem could be tackled by using more advanced backend data management systems and applying different visualization techniques to better represent the omics data.

- Release the publicly available tools. Since we published iGPSe and GRAPHIE in BioVis 2014 and 2015 [164, 165], we have received positive feedbacks from the cancer research community. Many fellow researchers contacted us and requested for our tools for their research. To fill this need, we team up with Kitware, a company specializes in the research and development of open-source software in the fields of medical imaging, visualization, and technical software development, to develop a web-based online platform, OSU Muli-Omics analysis (OSUMO). The OSUMO platform aims for creating a one-stop website for researchers to perform integrative analysis on custom multi-omics data. Both GRRAPHIE and iGPSe will be integrated into OSUMO serving as key visual analytics components. As of May 2016, we have finished implementing a demo version of the iGPSe component and can be accessed by public at http://www.osum.org.
• Supervised integrative classification. Our subspace integrative clustering method is developed under the unsupervised setting. For biological systems, it is crucial to utilize the prior knowledge during the model building process. This requires the algorithms that can integrate supervised mechanisms. In our formulation, no supervised information or prior knowledge is involved. One of the future directions can be subspace integrative classification with the allowance of multi-omics data.

• More omics data types and disease types. In this dissertation, we mainly used the mRNA, microRNA expression and DNA methylation data collected from breast cancer patients. However, our work is not limited to breast cancer. It can also be used to analyze other types of diseases. We also aim to integrate with other types of omics data (e.g. genomics and epigenomics and proteomics data) to enable an integrative phenotypical study.

Our work is only a start of the emerging research of multi-omics data analysis. The listed future directions can further improves the tools and methods, which lead to more insights to cancer studies.
Appendix A: Supplementary Material for Chapter 2
<table>
<thead>
<tr>
<th>Name</th>
<th>Visualization type</th>
<th>Tool type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cBio Cancer Genomics Portal</td>
<td>Networks</td>
<td>Web tool</td>
</tr>
<tr>
<td></td>
<td>Matrix Heatmaps</td>
<td></td>
</tr>
<tr>
<td>CircleMap</td>
<td>Circular heatmaps</td>
<td>Command line web tool</td>
</tr>
<tr>
<td>Circos</td>
<td>Circular genomic coordinates</td>
<td>Command line</td>
</tr>
<tr>
<td>Caleydo StratomeX</td>
<td>Heatmap</td>
<td>Desktop application</td>
</tr>
<tr>
<td>Cytoscape</td>
<td>Networks</td>
<td>Desktop application</td>
</tr>
<tr>
<td>Genomica</td>
<td>Matrix heatmap</td>
<td>Desktop application</td>
</tr>
<tr>
<td></td>
<td>Genomic coordinates</td>
<td></td>
</tr>
<tr>
<td>Gitools</td>
<td>heatmap</td>
<td>Desktop application</td>
</tr>
<tr>
<td>Integrative Genomics Viewer</td>
<td>Genomic coordinates</td>
<td>Desktop application</td>
</tr>
<tr>
<td>IntOGen</td>
<td>heatmaps</td>
<td>Web tool</td>
</tr>
<tr>
<td>NAViGaTOR</td>
<td>Networks</td>
<td>Desktop application</td>
</tr>
<tr>
<td>Regulome Explorer</td>
<td>Genomic coordinates</td>
<td>Web tool</td>
</tr>
<tr>
<td></td>
<td>Networks</td>
<td></td>
</tr>
<tr>
<td>Savant Genome Browser</td>
<td>Genomic coordinates</td>
<td>Desktop application</td>
</tr>
<tr>
<td>The Cancer Genome Workbench</td>
<td>Genomic coordinates</td>
<td>Web tool</td>
</tr>
<tr>
<td></td>
<td>Heatmap</td>
<td></td>
</tr>
<tr>
<td>UCSC Cancer Genomics Browser</td>
<td>Genomic coordinates</td>
<td>Web tool</td>
</tr>
<tr>
<td></td>
<td>Heatmap</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Supplementary Material for Chapter 5

Figure B.1: Box plot of Gene expression level for MYBL2, CENPA, AURKB and KIF2C in the integrative clusters
Figure B.2: Box plot of copy number variations of chromosome 19p13 in the integrative clusters
Figure B.3: Plot of number of subtypes versus silhouette score values
Figure B.4: ROC curves on integrative subspace, mRNA subspace, miRNA subspace and DNA methylation subspace
Bibliography


122


[93] Sohaib Ghani, Bum Chul Kwon, Seungyoon Lee, Ji Soo Yi, and Niklas Elmqvist. Visual analytics for multimodal social network analysis: A design study with


