MATHEMATICAL MODELS FOR DATA MINING AND SYSTEM DYNAMICS TO STUDY HEAD AND NECK CANCER PROGRESSION AND CHEMOPREVENTION

A Dissertation
Presented to
The Academic Faculty

by

Chanchala D. Kaddi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Department of Biomedical Engineering

Georgia Institute of Technology
August 2015

MATHEMATICAL MODELS FOR DATA MINING AND SYSTEM DYNAMICS TO STUDY HEAD AND NECK CANCER PROGRESSION AND CHEMOPREVENTION

Approved by:

Dr. May D. Wang, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Brani Vidakovic
Department of Biomedical Engineering and School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Melissa Kemp
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Howard Weiss
School of Mathematics and Departments of Biology and Global Health in Rollins School of Public Health
Georgia Institute of Technology and Emory University

Dr. Dong M. Shin
Winship Cancer Institute and Department of Biomedical Engineering
Emory University and Georgia Institute of Technology

Date Approved: May 29, 2015

To My Mom and My Dad

ACKNOWLEDGEMENTS

This dissertation research would not have been possible without the advice and encouragement of many individuals. I sincerely thank all of them for their help and support.

First, I gratefully acknowledge my advisor, Dr. May D. Wang. I originally met Dr. Wang when I was an undergraduate student making initial, tentative inquiries into research. She took the time to talk with me, shared her vision on the importance of computing to the future of medicine and the biosciences, and gave me the opportunity to join her lab for my first research experience. This was the first of many, many opportunities she has provided to me over the years, both as an undergraduate and as a Ph.D. student in her lab. Her positive attitude is unwavering. When I have been hesitant or doubtful, she has consistently encouraged me to try new things and to take chances. Sometimes this led to obstacles and dead ends, but many other times it led to accomplishments that I am very happy about. I greatly appreciate her insight and how she has guided my professional development.

I am very thankful to the four professors who served on my Ph.D. thesis committee: Dr. Melissa Kemp, Dr. Dong M. Shin, Dr. Brani Vidakovic, and Dr. Howard Weiss. Over the last several years, they have been an invaluable source of advice and encouragement. All of them consistently took time out of their busy schedules to meet with me, listen to my updates, challenge my assumptions, and help me find paths forward. Their deep and thoughtful feedback into both the biological and computational aspects of my projects has greatly improved the quality of my work, and has helped me learn how to be a researcher.

I am also very grateful to the professors and researchers with whom I have had the opportunity to collaborate, on both thesis and non-thesis research projects. Dr. Facundo M. Fernández has provided guidance and encouragement on several mass spectrometry imaging projects, and has always been a kind and helpful source of advice. Dr. Rachel (Bennett) Stryffeler

was a Ph.D. student in Dr. Fernández's lab, and has been my friend and collaborator for many years. Dr. Georgia Chen and Dr. A.R.M. Ruhul Amin have both given me valuable insight into the biology of head and neck cancer, and also into how to design and analyze research projects from the perspective of a clinical researcher. I have greatly appreciated the technical discussions I have had with Dr. Selwyn Hurwitz about cancer modeling.

Next, I would like to thank several former members of the Biomedical Informatics and Bioimaging Laboratory (BioMIBLab). Dr. R. Mitchell Parry was my day-to-day mentor during my first few years as a Ph.D. student. I learned a tremendous amount from Mitch: how to define a research problem, how to think about and understand mathematical models, how to present ideas in a technical paper, and more. Chang F. Quo was my mentor when I was an undergraduate researcher; he introduced me to the field of systems biology, and has always been a great source of guidance and advice. Dr. Yachna Sharma has been a great friend and source of support for many years. Melissa Freedenberg Meyers and Erica Oden were my co-adventurers when I first joined the lab, and were also my first co-authors.

I thank the current members of BioMIBLab, both for their practical help – debating ideas, listening to my presentations, and critiquing my writing – and also for sharing the experience of graduate student life with me, with plenty of laughter and good cheer. In particular, I thank Janani Venugopalan, Dr. Chih-Wen (James) Cheng, Po-Yen (Leo) Wu, Ryan Hoffman, Li Tong, and Cheng Yang. I would also like to acknowledge the several undergraduate mentees I've had the opportunity to work with, including Sanaiya Sarkari and Sameer Mishra.

Others previously and currently at Georgia Tech deserve many thanks for their support: Dr. Mahera Philobos is one of the kindest and most encouraging people I've met. Dr. Esfandiar Behravesh was my supervisor for a required teaching practicum; he made it an awesome experience and gave me a great deal of valuable insight. Laura Paige, the Bioengineering academic advisor, is a fantastic source of help, advice, and morale-boosting. Sally

Gerrish, one of the Biomedical Engineering advisors, has been extremely helpful and kind. I would also like to thank the late Mr. Chris Ruffin and Dr. Andrés García.

I thank the NSF Graduate Research Fellowship program for my award, which funded the first several years of my Ph.D. studies. The P.E.O. organization also provided financial support for my studies, and I would like to particularly thank Mrs. Ruth Anne Paradice, as well as the other members of Chapter B and Lyra Halsten.

My friends outside of lab have helped me tremendously over the last several years, by, among other things, letting me vent my frustrations and then making me forget them and laugh instead. In particular, thanks to: Amanda Dannemiller, Erzsi Sleder, Sehar Mehmood, Aparna K. Swamy, Rashmi Prashanth, and thanks also to many other friends not listed by name. I also thank Poornima Kaddi and my other cousins for very similar reasons.

I close this section with my deepest thanks: to my Mom and to my Dad. They are the source of my inspiration. Without their constant support and encouragement, from as far as my memory stretches, I would not have been able to pursue advanced studies and follow my dream to conduct biomedical research. And most fundamentally, I give thanks to God: *Om Namo Bhagavate Vasudevaya Namaha.*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

HNSCC	Head and neck squamous cell carcinoma
MSI	Mass spectrometry imaging
TLC	Thin layer chromatography
RNAseq	RNA sequencing
RPPA	Reverse phase protein array
ODE	Ordinary differential equation
ABM	Agent-based model
CV	Cross-validation
DESI	Desorption electrospray ionization
MCC	Matthews correlation coefficient
AUC	Area under the receiver operating characteristic (ROC) curve

SUMMARY

Head and neck squamous cell carcinoma (HNSCC) is the 6th most prevalent cancer worldwide, and more than 12,000 deaths from this disease are anticipated in 2015 in the U.S. alone. The advent of the "Big Data" era for biomedicine, through the widespread use of genomic, transcriptomic, and other –omic data acquisition technologies, has enabled deeper exploration of the molecular-level mechanisms behind HNSCC development and progression. This knowledge in turn can lead to earlier diagnosis and better treatment strategies, resulting overall in better patient outcomes. However, the volume and complexity of –omic data present a major obstacle to fully realizing its potential to accelerate and enable basic and translational research for HNSCC.

The goal of this Ph.D. dissertation is to address several key technical challenges related to harnessing –omic data for clinical HNSCC research. These are (1) the lack of knowledge-driven modeling tools and systems for discovering biomarkers at the protein and metabolite levels; (2) the lack of effective strategies for integrating heterogeneous types of –omic data for prediction; and (3) the lack of systems-level representations of biomarker knowledge for effectively predicting responses to bioactive agents. This dissertation addresses these challenges through three specific aims:

1. Knowledge-driven Data Mining: To develop modeling tools to mine –omic datasets in HNSCC for biomarker discovery by harnessing existing knowledge
2. Integrated –Omic Modeling: To develop supervised learning models for predicting HNSCC progression through integration of –omic datasets

3. System Modeling: To develop dynamic system models for predicting response to combinations of multi-target agents against HNSCC

The research in this dissertation was completed in collaboration with the Winship Cancer Institute and Georgia Institute of Technology. The models and tools developed have been systematically evaluated and validated using a variety of –omic data types. These results and associated case studies demonstrate the contribution of this work to and its future potential in computational HNSCC research.

CHAPTER 1

INTRODUCTION

1.1 Head and Neck Squamous Cell Carcinoma (HNSCC)

Head and neck cancer arises in the upper aerodigestive tract, at regions including the oral cavity, oropharynx, larynx, hypopharynx and tongue, as shown in Figure 1.1. As the vast majority (more than 95%) of head and neck cancers are squamous cell carcinomas [1], hereafter in this dissertation the disease will be referred to as head and neck squamous cell carcinoma (HNSCC).

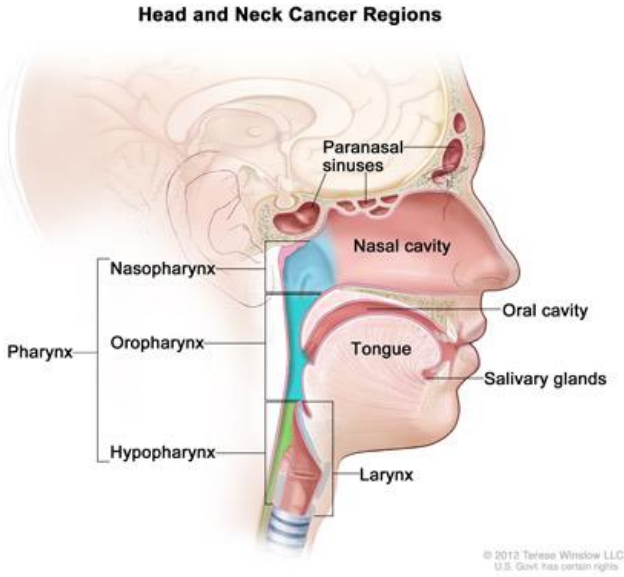


Figure 1.1: Disease subsites for HNSCC. Figure from National Cancer Institute: http://www.cancer.gov/cancertopics/factsheet/Sites-Types/head-and-neck

1.1.1. Statistics and Epidemiology

HNSCC is the 6th most prevalent cancer globally, with more than 600,000 new cases expected annually [1-4]. In the U.S. in particular, it represents 3% of all cancers, and in 2015 approximately 60,000 new cases and more than 12,000 deaths are expected.

Historically, HNSCC has been associated with alcohol and tobacco usage, and particularly by their use in combination [1, 5]. There is also a high prevalence of the disease in East and Southeast Asia, associated with the popularity of betel (areca) nut chewing [6, 7]. Overall, the disease is most prevalent among males over the age of 60. However, there is a growing subpopulation of HNSCC cases associated with human papillomavirus (HPV) infection. These patients tend to be younger, lack a history of alcohol and tobacco use, and predominantly experience cancer of the oropharynx.

1.1.2. Current Approaches for Treatment and Prevention

HNSCC treatment options include surgery, radiation, chemotherapy, or combinations of these treatments; specific treatment protocols vary based on disease sub-site and the stage at which the cancer presents [8]. Many patients with locally advanced (stage III/IV) disease respond favorably to treatments, and reach the so-called No Evidence of Disease (NED) status [9-13]. However, NED patients often later experience recurrence, secondary primary tumor (SPT) development, or metastatic disease. These factors give rise to the poor 5-year survival percentages (near 50%) observed for many HNSCC subsites [14].

1.2. The Role of Big Data: Opportunities and Challenges

The sequencing of the human genome in the previous decade heralded the era of "Big Data" in biomedicine. The concept of Big Data refers not only to the size of datasets being generated, but also to the complexity, quality, and utility of the measured features and to the speed of overall data acquisition. A common shorthand for these characteristics is the "five V's": volume, variety, velocity, veracity, and value [15, 16]. The recent

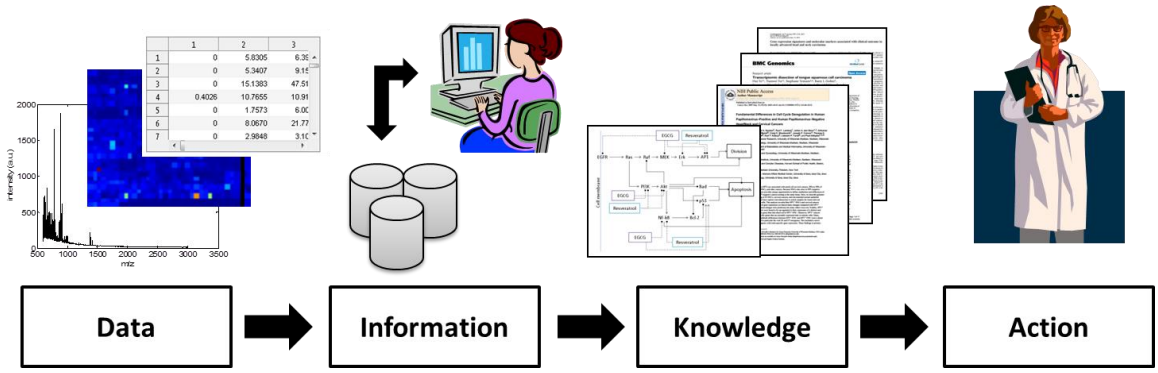


Figure 1.2: Analysis pipeline for Big Data in biomedicine

establishment of the Big Data to Knowledge (BD2K) initiative by the National Institutes of Health emphasizes the importance of this direction of research to multiple diseases, including cancers. The fundamental goal of Big Data analysis is to transform raw data, which may be too voluminous and complex for human interpretation, into organized, high-quality information and then into easily understandable knowledge. This is in turn used for decision-making and concrete actions. Figure 1.2 describes this pipeline.

In biomedicine, this pipeline is embodied by the identification of putative biomarkers that can be validated and then applied in clinical practice. Such biomarkers are individual genes, proteins, or other types of measurable characteristics which are associated with the disease state. There are two main reasons why biomarker research is important [17-19]. First, biomarkers may be harnessed for clinically relevant goals such as early detection, diagnosis, or patient stratification. Second, identified molecules may themselves be druggable targets, or they may interact with such targets, and therefore are of interest in pharmaceutical research.

One of the main sources of Big Data in biomedicine are the so-called '-omic' technologies, which are capable of measuring completely, or to a large coverage extent, the genome, epigenome (epigenetic modifications), transcriptome (mRNA transcripts),

proteome (expressed as well as functionally active proteins), metabolome (metabolites, including lipids), and other bio-molecular feature spaces. The availability of these large-scale, high-resolution data has helped to identify molecular expression patterns underlying many diseases, including HNSCC [20-22]. However, harnessing the potential of Big Data for HNSCC is far from complete. Here, I identify three major challenges related to applying –omic data for better understanding of HNSCC characteristics and for translating them into therapeutic strategies.

1.2.1. Genomics and Downstream –Omics: Knowledge-driven Mining for Transcriptomics, Proteomics, and Metabolomics

Cancer is a genetic disease in the sense that the transformation of normal cells to cancer cells is driven by alterations to the genome [23-27]. Great strides have been made in uncovering genomic changes associated with many types of cancers, including HNSCC. In particular, large-scale initiatives by The Cancer Genome Atlas have systematically explored the patterns in mutations, copy number variations, and epigenetic effects associated with HNSCC [28-30]. Validated oncogenes, such as PIK3CA and CCND1, and tumor suppressors, such as TP53 and PTEN, have focused attention on key biological pathways and processes that contribute to the development and progression of HNSCC. The most recent such analysis by TCGA also delineated four different sub-types of HNSCC [30].

However, identifying alterations at the genomic level is insufficient to fully characterize the disease state. Downstream effects of the altered genome propagate through and modify the expression of genes, proteins, and metabolites. The ongoing advancement of sequencing (RNAseq), mass spectrometry, and array-based technologies

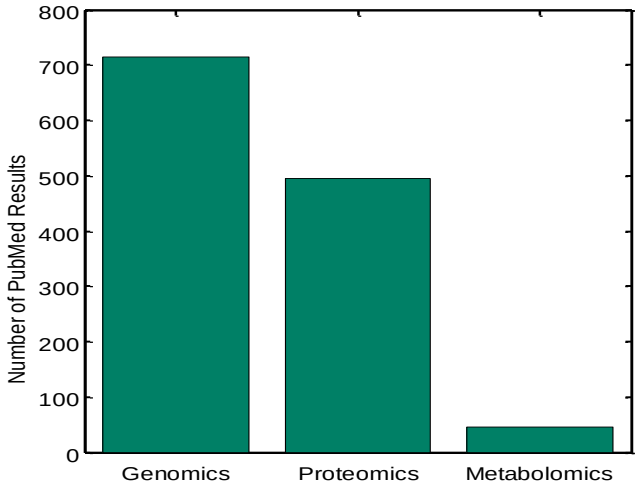


Figure 1.3: Comparison of the current number of PubMed hits for alternative –omics technologies in combination with head and neck cancer

provides powerful tools for data acquisition at these –omic levels. However, compared to progress in genomics, HNSCC research in proteomics – and especially in metabolomics – remains at an early stage. For example, Figure 1.3 compares publication counts in PubMed for these areas. One of the reasons is the inherent nature of the data: while genomic and transcriptomic data are completely described by nucleic acid sequences, amino acids describe only protein primary structure. Protein function and activity are determined by higher order structures and a complex network of regulatory interactions. Metabolites also exhibit great structural diversity, and are not encoded in the genome at all. In addition to this, high variation in abundance levels makes measurement and identification – and hence data interpretation – challenging [31]. Another critical reason is the lag in developing appropriate computational tools [32]. However, recent studies have underscored the importance of proteomics and metabolomics in understanding HNSCC development and progression [33-36]. Therefore, a key challenge is to develop

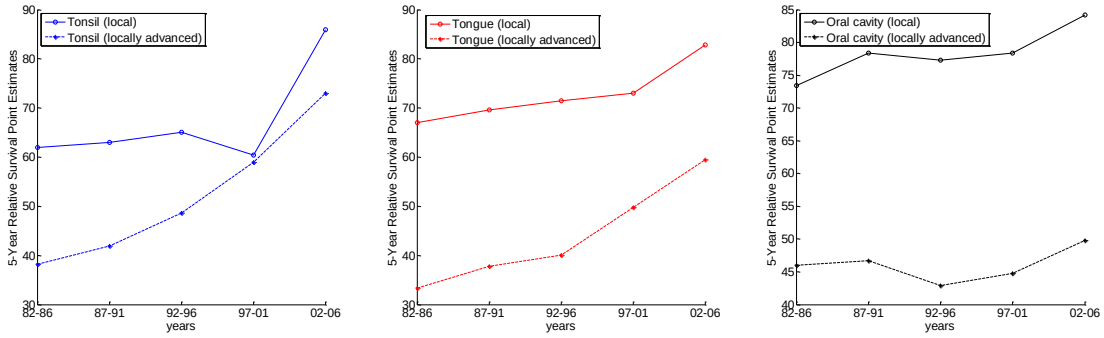


Figure 1.4: Comparison of the 5-year survival rates (point estimates) for early (solid lines) and locally advanced (dashed lines) HNSCC of the tonsil, tongue, and oral cavity, between 1982 and 2006. Data from [14].

mathematical models and tools for accelerating the identification of putative HNSCC biomarkers, particularly for proteomic and metabolomic data. And fundamentally, it is of interest to develop knowledge-driven models and tools, which can harness existing biological and biomarker knowledge to facilitate and accelerate data mining.

1.2.2. Integrated –Omics for Predicting Disease Progression

For most HNSCC subsites, there is a large difference in expected outcomes depending on the stage at which the disease is detected [14, 37]. Figure 1.4 shows trends for three subsites. Therefore, understanding the molecular-level changes that accompany disease development and progression could result in biomarkers for early diagnosis and in potential therapeutic targets. Recent modeling studies have investigated the differences between pre-malignant lesions and oral cancer using transcriptomic data [38, 39], and several transcriptomic, proteomic, and metabolomic studies have examined the differences between early and advanced stage HNSCC [34, 40-46]. However, these studies have yielded mixed results overall. Some identified discriminatory features

between early and advanced stage samples, while others did not. Moreover, some models for similar endpoints, using the same data type, identified non-overlapping feature sets.

Therefore, a key challenge in computational HNSCC research is effective data integration. Integration can occur both within and between –omic levels: by combining data of similar types across platforms (e.g., protein expression array and protein expression measured via mass spectrometry), or by combining different types of data (e.g., protein and gene expression data). Within –omic levels, some heterogeneity among datasets is expected due to experimental protocol- and platform-related factors; however, agreement is expected among fundamental putative biomarkers. Between –omic levels (e.g., genes and proteins), greater variation is expected because of the complex regulatory effects involved [47]. Better understanding of the key molecular features at each –omic level can provide insight into how these diverse molecular species collectively drive overall disease progression. Consequently, models which harness multiple types or levels of –omic data could provide better predictive performance and clinical utility.

1.2.3. Combination Strategies for Chemoprevention

Advances in –omic data acquisition and analysis have highlighted the importance of many signaling and metabolic pathways in HNSCC. As a result, molecularly targeted agents are emerging as a complement to conventional chemotherapeutics [2, 48]. However, due to factors such as pathway cross-talk, the response to individual targeted therapies has been limited, while those of combination therapies are promising. Some of these targeted agents, such as erlotinib and celecoxib, are also being applied for chemoprevention, in order to delay or prevent cancer progression [49, 50]. While these adjuvant chemoprevention therapies have shown promising effects in initial trials,

toxicity is a limiting factor. Therefore, the ongoing development of non-toxic agents for chemoprevention derived from natural dietary compounds, such as fruits and spices, is an important research direction [51, 52]. Aside from non-toxicity, one of the key strengths of these natural compounds, such as (-)epigallocathecin gallate (EGCG) from green tea, is that they are multi-target, interacting with key signaling pathways in complex manners. Moreover, combinations of natural compounds have demonstrated synergistic effects that can help compensate for limiting factors like low bioavailability [53-56].

Observations regarding individual versus combination strategies indicate that for predicting therapeutic and chemopreventive outcomes, it is insufficient to only identify molecular biomarkers. It is also necessary to gain a "systems-level" understanding of their roles in the context of signal transduction and metabolic pathways. This paradigm of data mining followed by modeling reflects the data → information → knowledge → action pipeline in Big Data research, since a system model is a higher-level representation of biomarker knowledge. Quantitative representations of cancer cell population and tumor growth have a long history in cancer research [57, 58]. In particular, the developing area of multi-scale cancer modeling explicitly links molecular-level observations, such as up-regulation of particular enzymes, to higher-level pathologically observed features, such as tumor aggressiveness. These representations are important for understanding system behavior and responses [59, 60]. However, previous modeling studies for HNSCC have focused on radiotherapy and chemotherapy, not targeted or multi-target therapeutic agents. Therefore, a key challenge is to develop a multi-scale modeling framework for HNSCC that can effectively predict the effects of combining multi-target agents.

1.2.4. Mathematical Modeling to Accelerate Translational Research

Mathematical modeling approaches are critical for addressing the challenges discussed in the preceding sections. The proposed solutions to these challenges are categorized into three main focus areas: (1) knowledge-driven mining, (2) data integration, and (3) system modeling. These approaches are all critical for handling the volume, variety, and velocity characteristics of Big Data. First, because of the volume and velocity of –omic data acquisition, modeling contributions that utilize existing knowledge to guide mining can help ameliorate the bottleneck imposed by analyzing large datasets. Second, data integration approaches are necessary for extracting knowledge from the volume (within –omic data types) and variety (between –omic data types) of large biological datasets. Lastly, system modeling provides a higher-level organization to the knowledge obtained through knowledge-driven data mining and data integration, and can generate specific, quantitative, and testable predictions.

1.3. Proposed Study and Organization of Dissertation

This dissertation focuses on addressing the three previously described key challenges related to HNSCC progression and chemoprevention. This is accomplished by developing mathematical models for data mining and for predicting system dynamics. The Specific Aims of this research are:

1. Knowledge-Driven Data Mining: To develop modeling tools to mine –omic datasets in HNSCC for biomarker discovery by harnessing existing knowledge

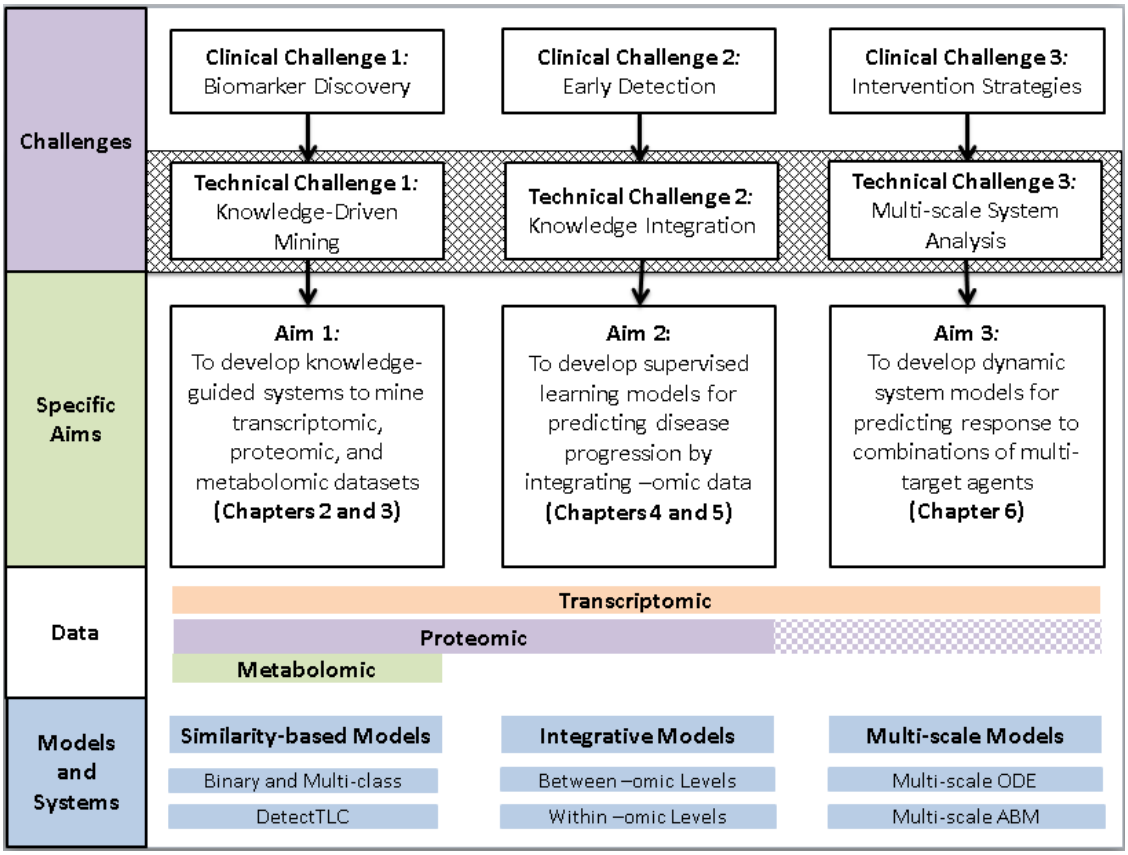


Figure 1.5: Workflow of Dissertation Research

2. Integrated –Omic Modeling: To develop supervised learning models for predicting HNSCC progression through integration of –omic datasets
3. System Modeling: To develop dynamic system models for predicting response to combinations of multi-target agents against HNSCC

In combination, this suite of modeling tools accelerates knowledge extraction from –omic Big Data in HNSCC. Chapter 2 of this dissertation focuses on Knowledge-Driven Data Mining, through the development of similarity measures applicable to multiple –omic data types. Chapter 3 also addresses Knowledge-Driven Data Mining, but focuses on constructing a system specifically for mining metabolomic data. Chapters 4

and 5 are focused on Integrated –Omic Modeling. Chapter 4 develops models for predicting HNSCC pathological stage by integrating transcriptomic and proteomic data. Chapter 5 proposes models for early detection of HNSCC through the integration of multiple transcriptomic datasets. Chapter 6 addresses System Modeling by developing multi-scale models for predicting the response to natural compound adjuvants for HNSCC chemoprevention.

Figure 1.5 describes the overall workflow of this dissertation, including the Specific Aims, the data types considered, and the developed models and tools. Chapter 7 concludes the dissertation by summarizing the key deliverables, including publications, and by discussing future directions for research.

CHAPTER 2

SIMILARITY MEASURES FOR EXPLORATORY DATA MINING

2.1. Applications of Similarity in Biomedical Research

Similarity measures are an important tool in the analysis of a wide range of biomedical data, with applications such as comparing peptide sequences [61] and gene expression data [62], as well as in text mining [63] and in image analysis [64, 65]. An important application of similarity measures is the detection of new and potentially functionally relevant patterns in large-scale biological datasets [62, 65, 66]. For example, if a particular gene is known to be associated with a disease, other genes potentially related to the disease may be detected by identifying highly similar patterns of expression. In this respect, similarity measures can be used to provide a shortlist of targets for further research.

Different similarity measures exhibit considerable variation in properties and performance [67, 68]. For example, many common measures do not have a probabilistic framework, although this is a useful property in terms of the interpretation of assigned similarity scores [69]. In this chapter, similarity measures with a probabilistic interpretation are proposed and developed. The first measure is restricted to two-class data, i.e., the comparison of binary images and data vectors. This model utilizes the hypergeometric distribution and Fisher's exact test. However, many types of biological data are not inherently binary in nature, and the thresholding process can discard useful information. Thus, the second measure utilizes the multivariate hypergeometric distribution and the Fisher-Freeman-Halton test to extend the first model to accommodate the comparison of non-binary, "multi-class" data.

2.2. Introduction to Mass Spectrometry Imaging

Mass spectrometry imaging (MSI) data is used several times in this chapter for testing the similarity measures, due to the unique combination of molecular and morphological information it provides. It is also used in Chapter 3. Therefore, this section provides a brief introduction to MSI in order to facilitate interpretation of later results.

MSI is an extension of conventional (non-imaging) mass spectrometry that can yield spatially-resolved information about the molecular composition of a biological sample. MSI datasets are generated by acquiring the complete mass spectrum at multiple points across the sample surface, yielding a three-dimensional (*x*,*y*: spatial dimensions, e.g. tissue, and *z*: spectral dimension) dataset as shown in Figure 2.1. The MSI dataset includes valuable information which is not obtainable through similar analyses using immunohistochemistry staining or conventional mass spectrometry. In traditional histological analysis, tissue is typically stained for a small number of molecular targets. In contrast, MSI is capable of simultaneously tracking thousands of *m/z* (mass-to-charge ratio) values. Depending on the MSI acquisition modality, each *m/z* value can be

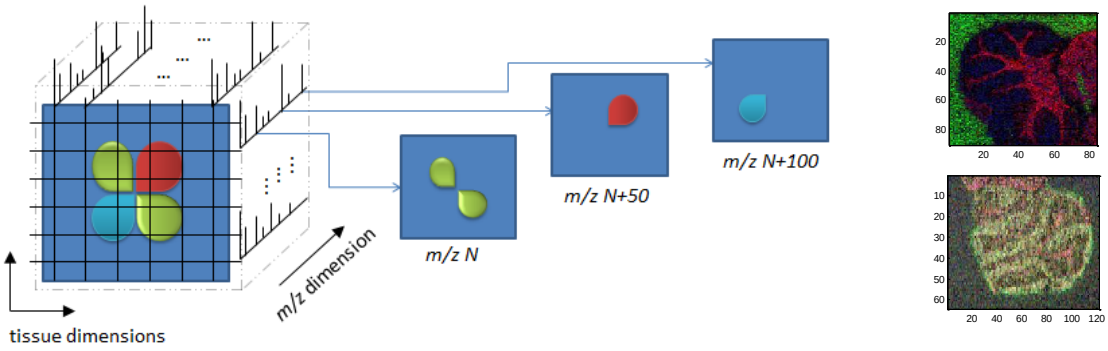


Figure 2.1: *(left)* Three-dimensional structure of MSI data. (*right):* False-color visualizations of multiple *m/z* values from MSI datasets of mouse models of Tay-Sachs/Sandhoff disease.

interpreted as a molecule or molecular fragment. Additionally, staining can only identify known molecular targets, while the large-scale data acquired by MSI enables discovery of sample components (and hence, potential biomarkers). Compared to mass spectrometry alone, MSI preserves the sample's spatial and morphological information. Thus, spectra corresponding to different regions of tissue samples like biopsies (e.g., tumor, marginal, or normal regions) can be differentiated, enabling more detailed and target-specific analysis. Due to these benefits, MSI is emerging as a popular experimental technique in proteomics [70], lipidomics [71], and metabolomics [72] research.

Because MSI is spatially-resolved, it is particularly relevant for research into diseases which have spatially localized characteristics – such as cancer. Recent MSI studies have investigated HNSCC [73], as well as cancers of the brain [74], breast [75], kidney [76], stomach [77], prostate [78], colon [79], pancreas [80], and bladder [81]. Other recent MSI studies have targeted diseases including Tay-Sachs/Sandhoff disease [82], Behçet disease [83], Parkinson's disease [84, 85], Alzheimer's disease [86], Duchenne muscular dystrophy [87, 88], Fabry disease [89], atherosclerosis [90] and stroke and ischemic injury [91-93]. In addition, MSI has been used to study bio-implant interfaces [94, 95] and drug distribution within tissues [96-101].

The spectral dimension of MSI data can be very large (e.g. tens of thousands of *m/z* values), making computational analysis essential to interpretation. Thus, it is critical to identify and to develop effective analytical methods for large-scale data mining and pattern recognition to effectively utilize MSI data. I have discussed the current state-of-the-art in MSI analysis techniques, including dimensionality reduction (e.g., principal

component analysis), clustering, and classification, in [102], but this content is outside the scope of this dissertation.

2.3. Binary Hypergeometric Similarity Measure

In this section, a binary similarity measure is proposed and developed, using the hypergeometric distribution and Fisher's exact test as a basis. The hypergeometric distribution has previously been used in bioinformatics to assess similarity in microarray functional analysis and tandem mass spectrometry [103-105]. The proposed hypergeometric similarity measure is compared with cosine similarity and Pearson correlation in terms of desirable properties related to formulation and behavior. Cosine similarity and Pearson have previously been used to assess similarity in mass spectrometry data for tasks ranging from protein identification to quality control [106-110]. The performance of the proposed similarity measure on synthetic data and experimental MSI data is studied, and examples are provided to demonstrate its advantageous performance in identifying and ranking similarities.

Desirable Properties of a Similarity Measure

The proposed similarity measure should sufficiently meet the following properties related to design and performance. The similarity measure should (1) be monotonically increasing between [-1, 1], in order to facilitate interpretation and comparison with other measures; (2) have good power of discrimination, i.e., should identify differences where they exist; (3) be consistently defined, i.e., there should not be sets of valid (observable) inputs for which the similarity measure output is undefined, and valid inputs should utilize the full dynamic range of the output.

Definition of Similarity Measure

Consider a dataset consisting of $i = 1, 2 \ldots m$ vectors $x_i \in \mathbb{R}^N$. The reference vector x_i contains n_1 dimensions with intensities greater than a selected threshold. When converted to binary form with respect to some threshold, this vector will have n_1 'on' dimensions and $N - n_1$ 'off' dimensions, which can be represented '1' and '0', respectively. A second, query vector x_j has n_2 'on' dimensions. The total number of dimensions at which both images are 'on' is *k*. The significance of overlap can be defined in terms of the probability of observing *k* given *N*, n_1, and n_2. If *k* of the n_1 dimensions from the first vector overlap *k* of the n_2 dimensions from the second vector, those *k* dimensions in the first vector may be arranged in $\binom{n_1}{k}$ ways. In the second vector, the ($n_2 - k$) dimensions which do not overlap may be arranged in $\binom{N-n_1}{n_2-k}$ ways. Thus, the total number of ways in which an overlap of *k* dimensions can occur, given n_1, n_2 and *N*, is $\binom{n_1}{k}\binom{N-n_1}{n_2-k}$. When divided by the number of ways in which the n_2 'on' dimensions in the second vector could be arranged if *k* of them were not constrained, this becomes the pmf of the hypergeometric distribution. I propose a similarity measure $h(k,n_1,n_2,N)$ which is defined, for any valid *k*, as the difference between the lower and upper "tails" of the hypergeometric distribution, as shown in equation (2.1).

$$h = \sum_{i=max(0,n_1+n_2-N)}^{k} \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} - \sum_{j=k}^{min(n_1,n_2)} \frac{\binom{n_1}{j}\binom{N-n_1}{n_2-j}}{\binom{N}{n_2}} \qquad (2.1)$$

The cumulative distribution function (cdf) of the hypergeometric distribution has previously been utilized as a similarity measure [111]. Since the population is discrete, additional information about k may be obtained by considering the probability of observing overlap at least as extreme. Both of these quantities can be considered p-values of hypothesis tests. In both cases, the null hypothesis H_0 is that the observed overlap occurred by chance. This is described by the urn model, in which an urn contains marbles of two colors, one representing a pair of overlapping 'on' pixels and the other representing non-overlap. When n_2 marbles are drawn from the urn without replacement and k of them are of the color representing overlap, the null hypothesis states that this has occurred by chance. The alternative hypotheses are that the observed overlaps are, respectively, larger or smaller than would be expected to occur at random for such an image pair. This implies that the images may be related, i.e., notably similar or dissimilar. Through the difference between these two probabilities, the proposed measure provides a scaled description of the unexpectedness of any observed overlap. The "tails" of the hypergeometric distribution also have upper bounds [112, 113]. For some parameter sets tested, the value of the hypergeometric pmf may be so small as to encounter machine resolution limits. Then, the proposed similarity measure may be implemented in terms of the upper bounds, as shown in equation (2.2).

$$h = \left(\left(\frac{p_1}{p_1+t_1}\right)^{p_1+t_1}\left(\frac{1-p_1}{1-p_1-t_1}\right)^{1-p_1-t_1}\right)^{n} - \left(\left(\frac{p_2}{p_2+t_2}\right)^{p_2+t_2}\left(\frac{1-p_2}{1-p_2-t_2}\right)^{1-p_2-t_2}\right)^{n} \tag{2.2}$$

Here, $n = n_1$, $p_1 = \frac{N-n_2}{N}$ and $t_1 = \left(\frac{n_1-k}{n_1} - p_1\right)$, such that $t_1 \geq 0$. Similarly, $p_2 = \frac{n_2}{N}$ and $t_2 = \left(\frac{k}{n_1} - p_2\right)$, such that $t_2 \geq 0$.

Similarity Measure Comparison and Assessment

First, the cosine similarity and Pearson correlation for binary vectors are expressed with the same variables as the hypergeometric pmf, allowing direct comparison of their formulae. Similarities and differences among the measures may be observed through their formulae; the binary expressions for cosine similarity and Pearson correlation are shown in equations (2.3) and (2.4). These are derived by noting that for binary vectors V_1 and V_2, the dot product is equivalent to k, and the norms to $\sqrt{n_1}$ and $\sqrt{n_2}$. Equation (2.4) is equivalent to Matthews correlation coefficient (MCC) [111, 114].

$$\frac{V_1 \cdot V_2}{\|V_1\|\|V_2\|} = \frac{k}{\sqrt{n_1}\sqrt{n_2}} \tag{2.3}$$

$$\frac{(V_1 - \overline{V_1}) \cdot (V_2 - \overline{V_2})}{\|V_1 - \overline{V_1}\|\|V_2 - \overline{V_2}\|} = \frac{k - \frac{n_1 n_2}{N}}{\sqrt{n_1(1 - \frac{n_1}{N})}\sqrt{n_2(1 - \frac{n_2}{N})}} \tag{2.4}$$

Performance on Synthetic and MSI Data

First, considering the mathematical expressions for each similarity measure, both cosine similarity and Pearson correlation are both linear with respect to k, and behave nonlinearly with respect to n_1 and n_2. Cosine similarity is independent of N, while mean-centering in Pearson correlation brings N into consideration. Pearson correlation asymptotically approaches cosine similarity for large N. The proposed measure, like Pearson correlation, considers N, but like cosine similarity, does not mean-center the data. Unlike both, it considers how unlikely it is to observe k by chance.

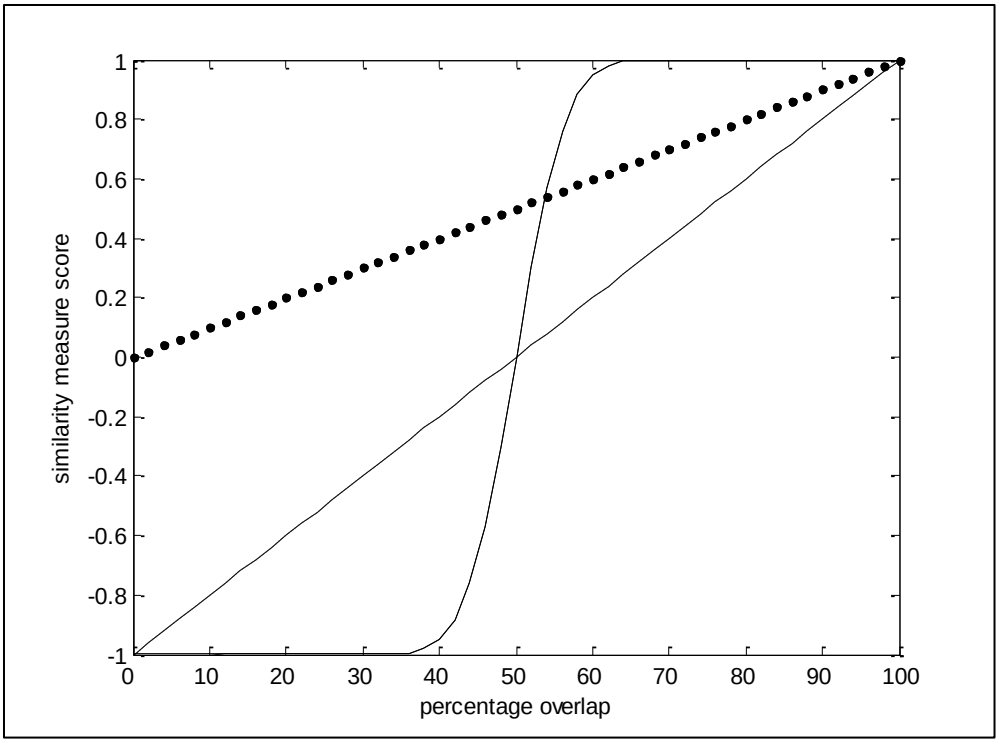
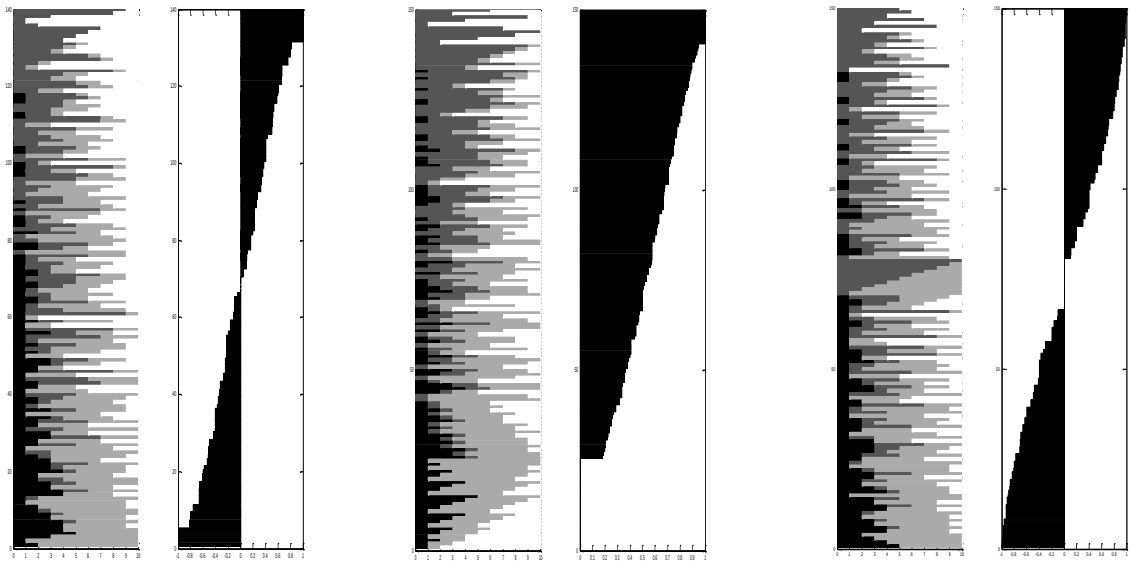


Figure 2.2: Hypergeometric similarity measure (solid), cosine similarity (dot) and Pearson correlation (dash) for $N = 100$, $n_1 = n_2 = 50$.

Second, the proposed similarity measure is compared with the cosine similarity and Pearson correlation by evaluating their output for binary image pairs having varying degrees of overlap. Figure 2.2 demonstrates that the proposed similarity measure satisfies criterion (1) regarding the desired properties of monotonicity and range. The hypergeometric similarity measure and Pearson correlation share a range of [-1, 1], while for positive data the range of cosine similarity is [0, 1]. The extremes of the hypergeometric similarity measure represent the limits of observable overlap k for a given parameter set N, n_1 and n_2.

Third, a synthetic dataset of binary vectors (images) with dimension $N = 10$ was created by considering all combinations of n_2, n_1 and k such that $N \geq n_2 \geq n_1 \geq k$, and such that k is greater than or equal to its minimum for any given n_1, n_2 and N (i.e., $k \geq n_1 + n - N$). This dataset consists of 150 vector pairs. The three similarity measures were



Pearson correlation Cosine similarity Hypergeometric similarity measure

Figure 2.3: Vector pair description (*l*) and corresponding similarity measure score (*r*). Color coding for the vector pair segments is as follows. Dark gray: k, number of overlapping 'on' pixels. Black: $n_1 - k$, number of non-overlapping 'on' pixels in image 1. Light gray: $n_2 - k$, number of non-overlapping 'on' pixels in image 2. White: $N - n_1 - n_2 + k$, total number of 'off' pixels. All vectors have length $N = 10$.

evaluated for each pair, and their outputs compared in terms of relative rankings. Figure 2.3 shows the performance of the similarity measures over the synthetic dataset.

The rankings show that the proposed similarity measure fulfills criterion (2), which addresses discrimination of differing cases. In particular, I examine the extreme cases of (a) no overlap, (b) complete overlap and (c) 'unsurprising' overlap. (a) Cosine similarity assigns 0 to all vector pairs with no overlap; a large segment of the dataset is so labeled with no additional sorting. Pearson correlation and the proposed similarity measure both sort this subset of vectors. However, only the proposed similarity measure recognizes that $k = 0$ is more surprising when n_1 approaches n_2, because there is more opportunity for at least some overlap. (b) The treatment of cases with complete overlap ($k = n_1 = n_2$) is also favorable with the proposed similarity measure because it orders them in a meaningful manner. It identifies $k = 5$ as the most 'surprising' case of complete

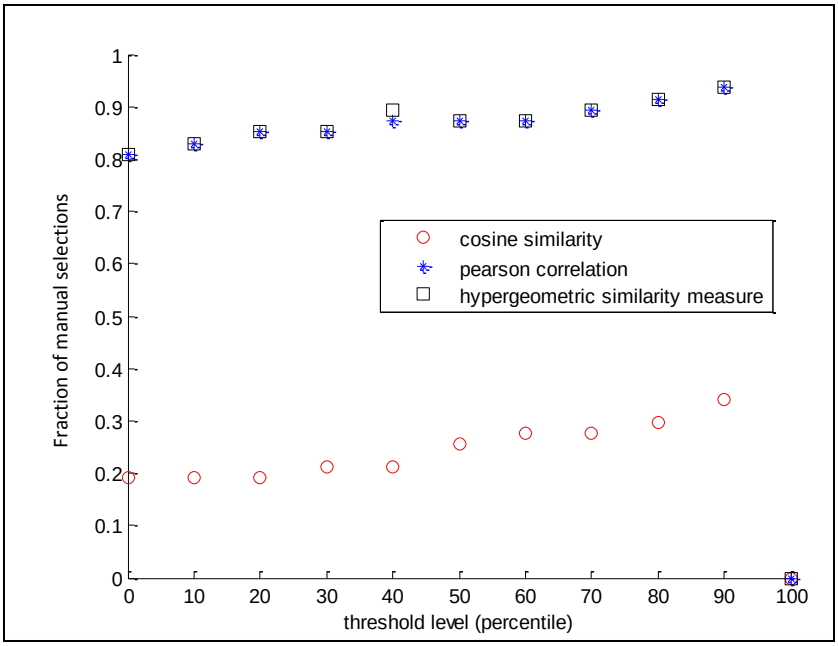


Figure 2.4: Fraction agreement between similarity measure and manual *m/z* selections across percentile-based binarization thresholds.

overlap, since there are the most opportunities for non-overlap to occur. The probability that $k = n_1 = n_2 = 5$ is equal to $\left(\frac{5}{10}\frac{4}{9}\frac{3}{8}\frac{2}{7}\frac{1}{6}\right)$. It also recognizes that $k = 6$ and $k = 4$ are equally 'surprising', since the probability of arranging $n_2 = 6$ pixels to completely overlap $n_1 = 6$ pixels is the same as arranging $(N - n_2) = 4$ pixels to completely overlap $(N - n_1) = 4$ pixels. The same pairings are observed for $k = 7$ and $k = 3$, and $k = 8$ and $k = 2$. In contrast, both cosine similarity and Pearson correlation assign 1 to this entire subset of vectors without further sorting. (c) The proposed similarity measure also meets criterion (3) regarding definition over the parameter space. Pearson correlation is not defined for vector pairs in which $n_1 = N$ or $n_2 = N$; this is evident from equation (4). The proposed similarity measure assigns 0 to these cases, because by definition, $k = n_1$. Thus, even though complete overlap occurs, it is not unexpected.

Fourth, the similarity measures were implemented on a biological MSI dataset. Although HNSCC tissue has previously been assessed using MSI [73], such data is currently not available to me for analysis. Instead, MSI data from a mouse model of Tay-Sachs/Sandhoff disease was used to test the similarity measure. Since general data characteristics for particular types of MSI data (e.g., MALDI, DESI, etc.) are expected to be similar regardless of the target tissue, it is reasonable to extrapolate conclusions to future performance on HNSCC MSI data. The experimental protocol for the MSI dataset investigated here is described in [82]. The image corresponding to *m/z* 890 was selected as a reference to due to its distinctive spatial pattern. The MSI dataset has a spectral dimension of 4,438 *m/z* values. It was manually inspected in non-binary mode to identify *m/z* values with similar spatial patterns; 47 *m/z* images were selected. The top 47 values selected by each similarity measure were compared to these values. The correspondence of the two lists was calculated for each similarity measure, and repeated for 11 alternative binarization thresholds. The upper bounds formulation shown in equation (2.2) used in this assessment due to the large variable values involved.

Figure 2.4 describes similarity measure performance, assessed as the fraction agreement between the top *m/z* values selected by each similarity measure and the manual selections. This comparison was carried out across multiple binarization thresholds based on the abundance percentiles of the mean spectrum. For this dataset, the 90^{th} percentile yields top selections from the similarity measures which correspond most closely to the manual selections. The selections of the proposed measure and Pearson correlation correspond highly with the manual selections, and also with each other. The

selections of cosine similarity consistently differed from the other two, and from the manual selections.

Discussion and Limitations of Binary Approach

In summary, a hypergeometric similarity measure is proposed as a tool for the exploration and analysis of biomedical data. Due to its definition as the difference between the upper and lower "tails" of the hypergeometric distribution, the proposed similarity measure explicitly defines the unexpectedness of any observed overlap. Using synthetic data, the proposed similarity measure was compared with cosine similarity and Pearson correlation in terms of three criteria related to design and performance, and it was shown to perform favorably. Tests on a biological, non-HNSCC MSI dataset showed that the proposed similarity measure is effective in identifying visually notable spatial similarities. Together, these results indicate that the proposed similarity measure can play a useful role in assessing similarity in biomedical data.

However, several caveats remain. First and foremost, abundance is a key feature of biological data, and analyzing binary data ignores this information. In the MSI dataset examined here, analysis of binarized data still revealed informative patterns. However, for some HNSCC datasets, retaining abundance information may be necessary for meaningful analyses. The flexibility to accommodate abundance to some extent is particularly important for the goal of developing a general similarity measure that provides useful output for multiple HNSCC –omic data types. Second, if binary data is to be used, the selection of appropriate thresholds to convert non-binary data to binary data is an open problem. Many alternative methods for thresholding images have been

proposed in image processing [115]. Results on this particular biological dataset indicated that increasing the threshold can increase agreement with the set of manually selected *m/z* values. However, the potential effects of inter-dataset variation on the performance of all three similarity measures have not yet been studied. That investigation would provide more insight into threshold effects and lead to more systematic recommendations for specific data types.

Instead, the next section in this chapter addresses this issue directly, by modifying the similarity measure to explicitly incorporate abundance.

2.4. Multivariate Hypergeometric Similarity Measure

In this section, the previous result is extended to present a general similarity measure that accommodates the comparison of non-binary, "multi-class" data. After defining the proposed multivariate hypergeometric similarity measure, I describe several tests using synthetic and biological data to investigate its performance. First, its patterns of sample ranking are again compared with those of cosine similarity and Pearson correlation, as well as with mutual information. These three similarity measures are used in the analysis of many types of biomedical data [62-65, 116]. Next, an algorithm called piecewise approximation, which facilitates the application of the proposed similarity measure to large samples, is developed and implemented.

Definition of Similarity Measure

Consider a dataset consisting of $z = 1, 2 \ldots m$ vectors $x_z \in \mathbb{R}^N$, with all intensities quantized into n bins, where N and n are positive integers. When comparing two such vectors, there are n^2 possible types of overlap between corresponding dimensions (i.e., in images, for spatially corresponding pixels). These overlaps can be represented as an $n \times n$ contingency table, as shown in Figure 2.5. Each class k_{ij}, for indices $i = 1...n$ and $j = 1...n$, represents the number of corresponding dimensions which are in bin i in the first ("reference") vector and in bin j in the second ("query") vector. The terminology is used in the sense that a given sample of interest would be selected as a "reference" and other samples in a dataset would be compared, or "queried" against it to find samples similar to the reference. The margins of the contingency table are fixed for a given pair of images: for each row i, $\sum_{j=1}^{n} k_{ij} = r_i$, the number of pixels in bin i in the reference image, and

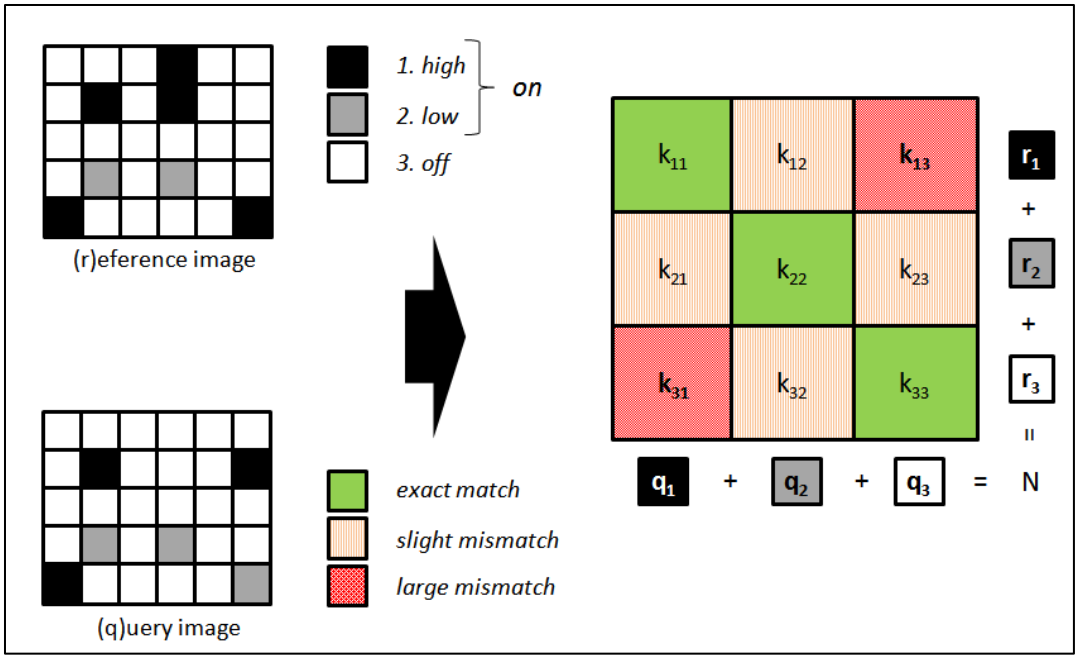


Figure 2.5: An image pair (reference and query images) with pixels intensities binned into three levels is represented as a 3×3 contingency table with fixed marginal totals.

similarly for each column $j, \sum_{i=1}^{n} k_{ij} = q_j$, the number of pixels in bin j in the query image. By definition, $\sum_{ij} k_{ij} = N$. The probability of observing a particular distribution of overlaps k_{ij}, i.e., the probability of observing a given contingency table, can be represented as the product of probability mass functions of the multivariate hypergeometric distribution. Considering only the first column of an $n \times n$ contingency table with row marginal totals r_i, the column sum is $q_1 = k_{11} + k_{21} + \ldots k_{n1}$. Each component k_{i1} is drawn from its row sum r_i. Since each draw is independent, the probability of observing a particular distribution of pixels is given as $\frac{\binom{r_1}{k_{11}}\binom{r_2}{k_{21}}\cdots\binom{r_n}{k_{n1}}}{\binom{N}{q_1}}$. This quantity is a probability mass function of the multivariate hypergeometric distribution. The probability of observing the second column is described similarly, but accounts for the pixels already assigned in the previous column: $\frac{\binom{r_1-k_{11}}{k_{12}}\binom{r_2-k_{21}}{k_{22}}\cdots\binom{r_n-k_{n1}}{k_{n2}}}{\binom{N-q_1}{q_2}}$. The same

pattern is followed through the $(n-1)^{th}$ column. Because the row and column sums are fixed, the configuration of the n^{th} column is determined by the preceding columns. Since the configuration of the available dimensions in each column (aside from the n^{th} column) is independent of the other columns, the probability p of the complete $n \times n$ contingency table is given by the product of the column probabilities. This quantity, shown in equation (2.5), is known as the probability for k-variate contingency tables [117, 118]. Here $q = [q_1, q_2 \ldots q_n]$, $r = [r_1, r_2 \ldots r_n]$ and $k = [k_{11}, k_{12} \ldots k_{nn}]$.

$$p(q,r,k) = \frac{\prod_{i=1}^{n} r_i! \times \prod_{j=1}^{n} q_j!}{N! \times \prod_{ij} k_{ij}!} \qquad (2.5)$$

In previous work focusing on binary data, the similarity measure was defined based on the hypergeometric distribution; the probability mass function of this distribution gives the probability of a 2×2 contingency table [119]. The similarity measure was defined as the difference between the lower and upper "tails" of the hypergeometric distribution defined by the marginal totals r and q. The values of r and q are a function of the particular reference image and query image being compared. The "tails" were defined with respect to the observed overlap, which was defined as the number of spatially corresponding pixels which are 'on' in both images, i.e., k_{11} in this terminology. To extend this approach from the two classes in binary data to n classes, I utilized the probability mass function of the $n \times n$ contingency table.

The statistical significance of a contingency table is evaluated by performing Fisher's exact test (in the binary case) or the Fisher-Freeman-Halton test (in the general case) [117, 120]. In both cases, the isomarginal family of tables (i.e. those tables having the same fixed margins r and q as the original table representing the reference and query image pair) is first generated, and the probability of each table within this family is

calculated. In the binary case, the hypergeometric distribution describes the isomarginal family. For each table in the isomarginal family, the value of a chosen statistic $S(k)$ is compared to that of the original table. With respect to $S(k)$, tables in the isomarginal family may be more extreme than the original table in two directions. The set of tables which are "more extremely large" have a larger than or equal value of the statistic, while the set of tables which are "more extremely small" have a smaller than or equal value of the statistic. The significance of a table in a particular direction is found by summing the probabilities of all tables within the respective set.

In the binary case, the choice of the statistic $S(k)$ is straightforward because due to the fixed margins, there is only one degree of freedom. $S(k) = k_{11}$ completely defines the table, and is reasonable because more similar images will have greater numbers of overlapping pixels. In the general case, however, there are $n^2 - 2n + 1$ degrees of freedom, and for $n > 2$, the choice of a statistic is not obvious. Here, I choose a vector of statistics – the set of diagonal elements of the $n \times n$ table – as $S(k)$, as shown in equation (2). These diagonal elements represent the exact matches – the spatially corresponding pixels in the reference and query images which are in the same class. While $S(k)$ may be defined in many alternative ways, I propose equation (2.6) as a reasonable choice for multi-class data because images which are more similar will have a greater number of each of the n types of exact matches.

$$S(k) = [\mathrm{k}_{11}, \mathrm{k}_{22}, \ldots, \mathrm{k}_{nn}] \qquad (2.6)$$

For each table in the isomarginal family, I performed an index-wise comparison of each diagonal element to the corresponding diagonal element in the original table. In other words, I compared each element in $S(k)$ with the corresponding element in S_0,

which is the instance of *S(k)* observed for the original table. If each diagonal element in the table is greater than or equal to its corresponding element in S_0, the table is assigned to set *G*, the set of "more extremely large" tables with respect to all elements of *S(k)*. If each diagonal elements is less than or equal to its corresponding element in S_0, the table is assigned to set *L*, the set of "more extremely small" tables. Equation (2.7) defines the proposed multivariate hypergeometric similarity measure *h* in terms of the probabilities of the tables in these two sets.

$$h = \sum_{L} p(q.r,k) - \sum_{G} p(q.r,k) \qquad (2.7)$$

Comparison of Similarity Measures

The sample rankings obtained from the proposed measure are compared with those from cosine similarity, Pearson correlation, and mutual information. Cosine similarity and Pearson correlation are defined for vectors V_1 and V_2 in equations (2.8) and (2.9), respectively. Mutual information is defined in (2.10), where x_i and y_j are the elements of V_1 and V_2, respectively.

$$\frac{V_1 \cdot V_2}{\|V_1\|\|V_2\|} \qquad (2.8)$$

$$\frac{(V_1 - \overline{V_1}) \cdot (V_2 - \overline{V_2})}{\|V_1 - \overline{V_1}\|\|V_2 - \overline{V_2}\|} \qquad (2.9)$$

$$-\sum_{i}^{n} \mathrm{p}(x_i)\log_2(\mathrm{p}(x_i)) - \sum_{j}^{n} \mathrm{p}(y_j)\log_2(\mathrm{p}(y_j)) + \sum_{i}^{n}\sum_{j}^{n} \mathrm{p}(x_i, y_j)\log_2(\mathrm{p}(x_i, y_j)) \qquad (2.10)$$

Design of Synthetic Dataset

First, the performance of the multivariate hypergeometric similarity measure is evaluated on synthetic data. While the proposed similarity measure is defined for any $n \geq 2$ classes, these synthetic experiments are performed using only three classes to clearly illustrate the method. Two synthetic datasets are used for this comparison. The first consists of the three-class isomarginal family defined by marginal totals (r_1, r_2, r_3, q_1, q_2, q_3) = (5, 5, 5, 5, 5, 5) and N = 15. The second consists of all three-class tables with N = 5.

2.4.1. Piecewise Approximation

Testing the significance of $n \times n$ contingency tables obtained from biomedical data, such as MSI data, poses a challenge due to data size. As the numbers of pixels in the images, and hence the marginal totals, increase, generating the isomarginal family of tables to perform the Fisher-Freeman-Halton test becomes demanding. The number of possible tables increases factorially as the numbers of rows, columns or total pixels increase [121, 122]. As an analytical example, the number of three-class contingency tables where all rows and columns sum to r is given by $\binom{r+2}{2} + 3\binom{r+3}{4}$ [123, 124]. When faced with a very large number of tables to enumerate in the isomarginal family, approximate solutions can be found through Monte Carlo testing [120]. However, in practice this may demand very large numbers of permutations to achieve satisfactory separation of similarity rankings.

Here I propose a piecewise method of approximation, in which the two images or data vectors to be compared are divided into a number of smaller subsections. The motivating idea is that similar samples will also have similar corresponding subsections.

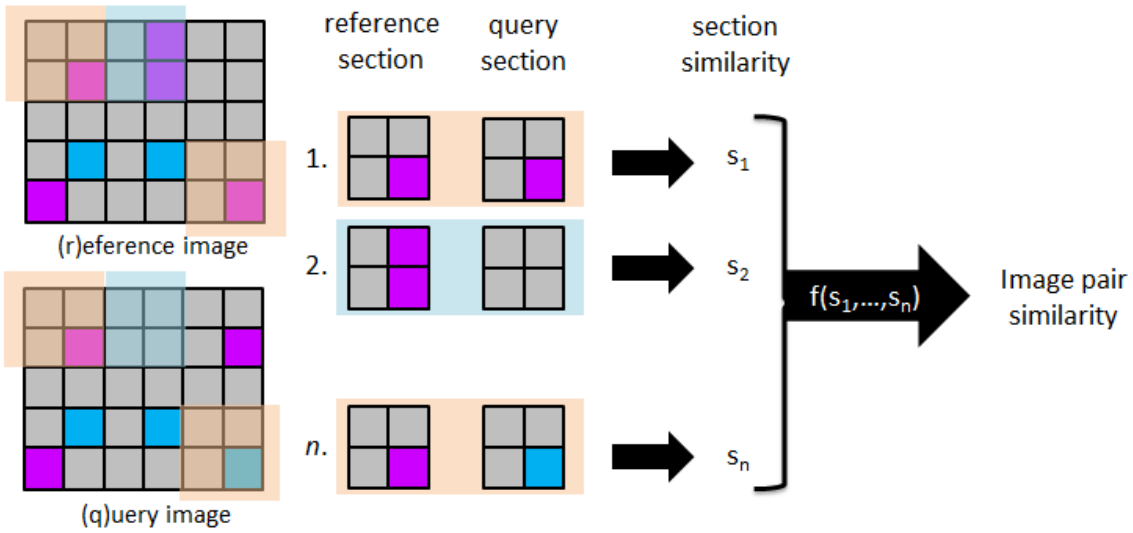


Figure 2.6: Overview of piecewise approximation process: subsection (1.) corresponds to the top-left 4x4 blocks of the reference and query images; (2.) to the 4x4 blocks to the immediate right of (1.); and (*n.*) to the bottom-right 4x4 blocks. The similarity is calculated for each spatially corresponding reference and query section, and the overall similarity of the reference and query is calculated as a function of the sub-section scores.

For each pair of reference and query subsections, an $n \times n$ contingency table is constructed and the multivariate hypergeometric similarity measure is calculated. The overall similarity of the image pair is computed as a function of the similarities of all subsections. Figure 2.6 illustrates this process.

Piecewise approximation requires choices in how images or data vectors are separated into subsections (e.g., different subsection sizes) and how the similarity scores for the subsections are combined to obtain an overall similarity score for the image pair (e.g., different functions). Alternative choices are examined here through experiments on synthetic and biomedical data. First, the previously described synthetic dataset (for N = 15) is used to examine whether there is a pattern between subsection size and the extent of difference observed between the piecewise approximation rankings and the exact rankings. In this test, the rankings for each sample obtained by using subsections of size

3, 4 and 5 pixels are compared with the ranking calculated using the whole sample. To avoid the comparison of single-pixel sections, if the sample is not evenly divisible at a particular increment size, the remainder pixels are added to the previous subsection to create one subsection larger than the others. In the same experiment, the effect of permuting the reference and query samples which correspond to a single $n \times n$ table is considered. While a given pair of samples yields a single $n \times n$ table, mapping a given table back to the sample space yields non-unique indexing of spatially corresponding pixels. This type of indexing difference would not affect the similarity score of a given reference and query pair if the whole sample is utilized. However, when piecewise approximation is employed, different subsections may contain different proportions of the pixels for each type of overlap k_{ij}. To examine how this may affect results, the 'randperm' function in MATLAB was used to generate a permutation of the sample indices, which was applied to both the reference and query samples before they were divided into subsections. This was repeated 10,000 times. The purpose of this step is to confirm that overall sample rankings in the synthetic dataset are not an artifact of arbitrary methods of generating synthetic samples from tables and subsectioning samples. For each subsection size, the sample ranking results shown are the mean across all permutations. Next, biomedical data was used to empirically compare alternative functions for aggregating the subsection similarity scores into an overall score for the image pair.

2.4.2. Performance on Synthetic Data

This section describes two sets of results. First, the performance of the proposed multivariate hypergeometric similarity measure is compared with the other similarity

measures using synthetic data. Second, the effects of subsection size and combination functions on the piecewise approximation method are investigated using synthetic and experimental MSI data.

In the first set of results, the rankings of samples in synthetic datasets by the three similarity measures are compared in Figures 2.7 and 2.8. Each sample (horizontal bar)

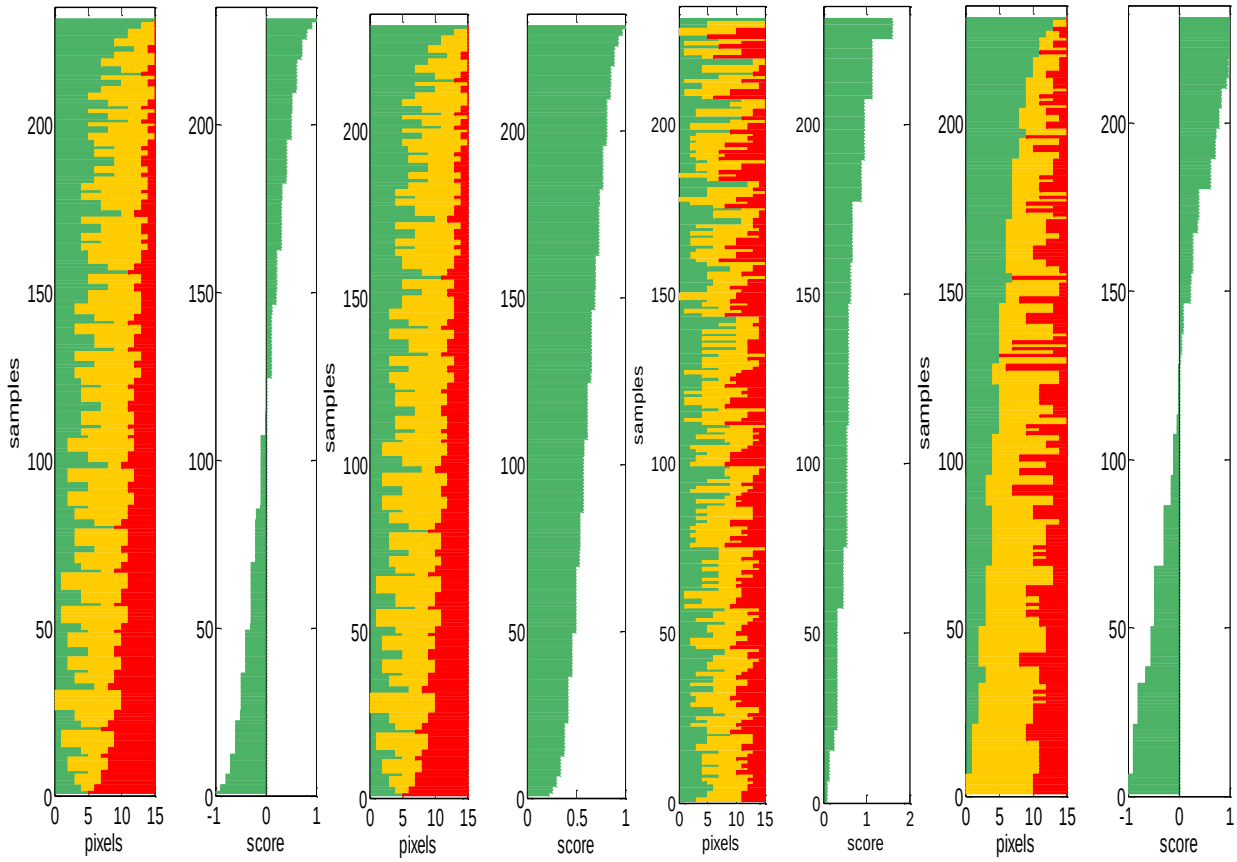


Figure 2.7: Comparison of sample rankings by the four similarity measures for the synthetic dataset comprising the isomarginal family given by (r_1, r_2, r_3, q_1, q_2, q_3) = (5, 5, 5, 5, 5, 5). Each sample (horizontal bar) represents a certain number of exact matches, slight mismatches and large mismatches (corresponding to [green, yellow, red], or [medium, light and dark] in grayscale). The length of each color segment corresponds to the number of that type of match in the sample. For each similarity measure, the similarity score corresponding to each sample is shown on the right panel.

represents a single table, with the green, yellow and red segments representing the number of exact matches ($k_{11} + k_{22} + k_{33}$), slight mismatches ($k_{12} + k_{21} + k_{23} + k_{32}$) and large mismatches ($k_{13} + k_{31}$), respectively. In Figure 2.7, there are 231 tables represented; these tables comprise the isomarginal family defined by marginal totals (r_1, r_2, r_3, q_1, q_2, q_3) = (5, 5, 5, 5, 5, 5).

All three similarity measures agree in that the highest score is assigned to the table with the largest number of exact matches. None of the similarity measures are monotonic with respect to the number of exact matches, but rankings from the proposed similarity measure are much closer to this trend than rankings from cosine similarity and Pearson correlation. Cosine similarity and Pearson correlation more closely sort by the number of large mismatches. For a single isomarginal family, the magnitudes and means of the two vectors are constant. The rankings of cosine similarity and Pearson correlation therefore depend on the value of the dot product, and the minimum dot product is observed when the number of large mismatches is maximized. The proposed similarity measure does not provide such distinction between slight and large mismatches, but it does provide a probabilistic interpretation which cosine similarity and Pearson correlation do not: the samples associated with extreme scores are the most "surprising" patterns of overlap observed. Mutual information assigns higher scores to cases where most pixels are concentrated in a few classes, but does not differentiate among the classes. For example, the tables with [k_{11}, k_{22}, k_{33}] = [5,5,5] (i.e., all exact matches) and [k_{31}, k_{22}, k_{13}] = [5,5,5] (i.e., many large mismatches) both receive equally high scores; as a result, the mutual information results do not show any trend with respect to exact matches, slight mismatches or large mismatches. In contrast, [k_{11}, k_{22}, k_{33}] = [5,5,5] is ranked highly by

the proposed similarity measure, while [k_{31}, k_{22}, k_{13}] = [5,5,5] receives a much lower score.

Figure 2.8 considers the rankings of the 1287 tables generated by considering every possible combination of marginal totals such that ($r_1 + r_2 + r_3 = 5$) and ($q_1 + q_2 + q_3 = 5$). Again, all of the measures agree in that the highest score is assigned to the table with the largest number of exact matches, but the proposed similarity measure more consistently assigns lower scores to tables with fewer exact matches. The rankings in this

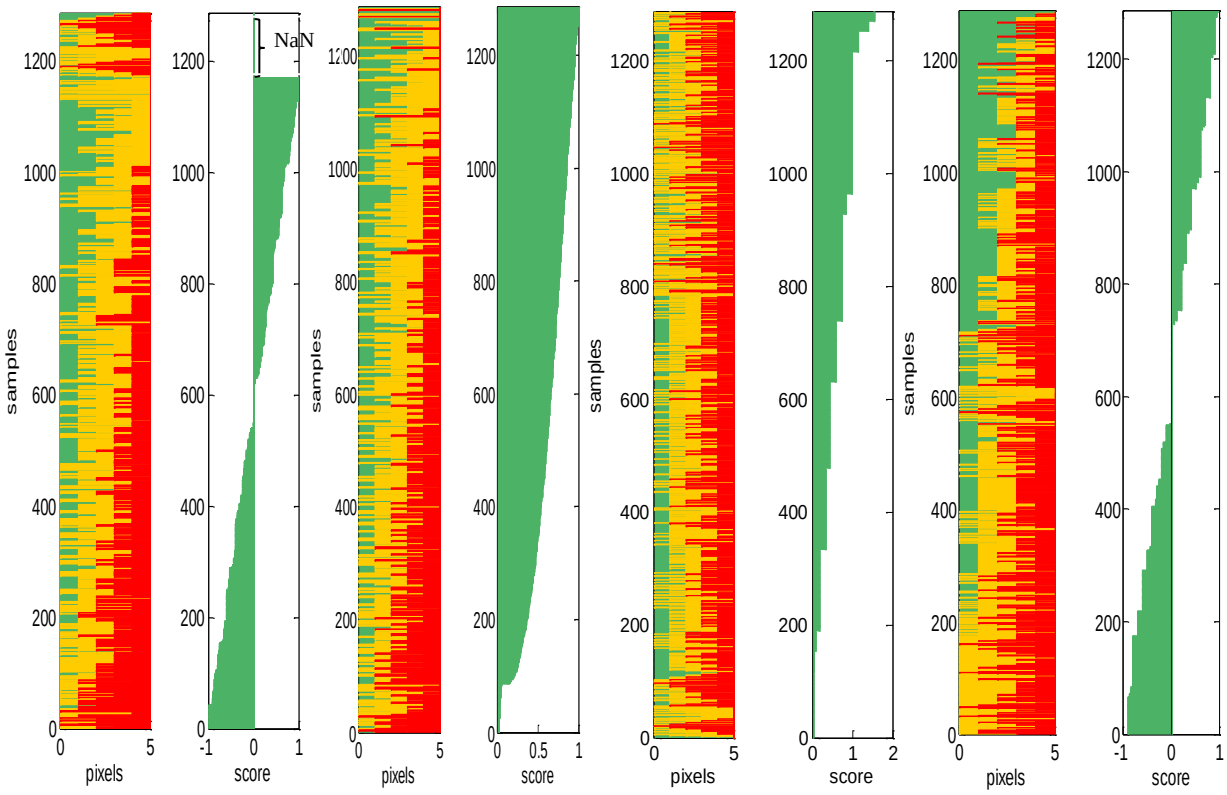


Figure 2.8: Comparison of sample rankings by the four similarity measures for the synthetic dataset containing all tables for $N = 5$. Each sample (horizontal bar) contains a certain number of exact matches, slight mismatches and large mismatches (corresponding to [green, yellow, red], or [medium, light and dark] in grayscale). The length of each color segment corresponds to the number of that type of match in the sample. For each similarity measure, the similarity score for each sample is shown on the right panel.

set of all tables for $N = 5$ illustrate additional probabilistic aspects of the proposed similarity measure. For example, the proposed measure can distinguish between instances of overlap with different distribution magnitudes. It assigns identical scores to the set of tables with $[k_{11}, k_{22}, k_{33}]$ as [3,1,1], [1,3,1] and [1,1,3], and a different identical score to the other possible set of tables describing only exact overlap, with $[k_{11}, k_{22}, k_{33}]$ as [2,2,1], [2,1,2] and [1,2,2]. Pearson correlation and cosine similarity do not distinguish between these two sets of tables. Mutual information distinguishes the two sets of tables, but again does not distinguish between case cases of exact matches and many large mismatches; for example, the cases where $[k_{11}, k_{22}, k_{33}] = [3,1,1]$ and $[k_{31}, k_{22}, k_{13}] = [3,1,1]$ are assigned the same score, and $[k_{11}, k_{22}, k_{33}] = [2,2,1]$ and $[k_{31}, k_{22}, k_{13}] = [2,1,2]$ are assigned the same score. For a second example, in the proposed measure, all tables which have marginal totals such that only one $n \times n$ table is possible are mapped to a score of zero. If only one set of overlaps k_{ij} can be observed for a particular pair of images or data vectors, then the overlap which is observed can be considered inherently "unsurprising." In contrast, this set of tables is undefined for Pearson correlation (i.e., these tables are assigned the value NaN, as shown at the top of the Pearson correlation plot in Figure 2.8). Cosine similarity does not group these tables together or otherwise distinguish them.

In the second set of results, the effects of subsection size on the piecewise approximation result are described in Figure 2.9. The 231 samples in the synthetic dataset shown in Figure 2.7 are plotted in order of increasing exact score. The piecewise approximation scores for each sample, across increments of size 3, 4 and 5, are compared. For all three subsection sizes, the mean score from 10,000 permutations of the

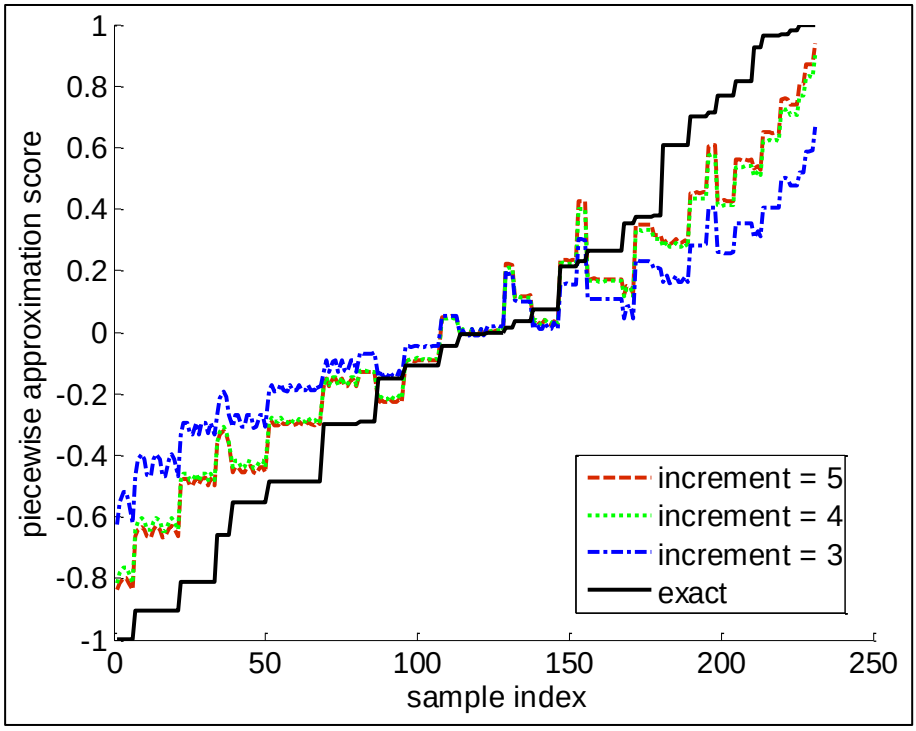


Figure 2.9: The mean rankings of synthetic samples using piecewise approximation at different subsection sizes (size 3: blue dash-dot line; size 4: green dotted line; size 5: red dashed line) compared to rankings from using the whole sample (black, solid line).

reference and query vectors is shown. Overall, the piecewise approximation scores follow the trend of the exact score, but there are notable deviations. In such cases, samples are ranked higher or lower as an artifact of the piecewise sectioning process. Interestingly, these cases tend to correspond across all of the subsection sizes; if a sample was scored much higher or lower than its adjacent samples by the piecewise method, the same jump or dip in score was observed across all three subsection sizes. However, the magnitudes of the piecewise scores indicate that, as expected, larger sections give scores closer to the exact result.

Next, different statistics for combining the similarity scores of subsections into a single overall score for the sample pair are compared empirically, using MSI data with a

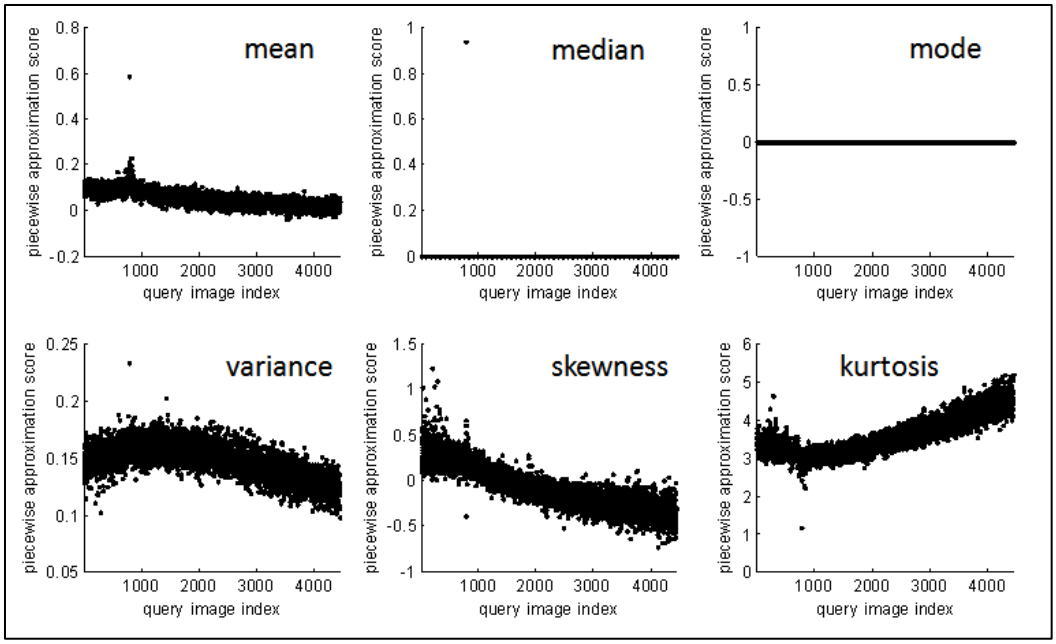


Figure 2.10: Empirical comparison of alternative functions for combining subsection similarity scores into an overall similarity score through piecewise approximation. Each dot on the scatter plot represents one query image (*m/z* value); 4,438 are in the dataset. The value (image pair score) of the dot represents the similarity score assigned to the query image based on the specified function of its subsection scores. For example, in the 'mean' plot, the image pair score of each query image is the average similarity score of its subsections.

subsection size of 4×4 pixels for piecewise approximation. Figure 2.10 shows the image pair similarity scores for each of the 4,438 *m/z* values, computed as the mean, median, mode, variance, skewness or kurtosis of all of their subsection scores. The *x*-axis of these plots, showing indices 1 through 4,438, represents the query *m/z* images; each is associated with a single score (dot) on the *y*-axis. This score is obtained by evaluating the specified function (e.g., the mean) over the set of subsection scores obtained for that query image when it was compared to the reference image. To interpret these results, it is necessary to consider that the reference *m/z* image corresponds to index 783. Since the most similar image in the dataset to the reference image should be the reference image itself, a well-performing function should assign the most extreme score to this index. This

result is observed for the mean, median, variance and kurtosis functions. During previous study of this dataset for the binary measure, 47 of the 4,438 images were observed to be qualitatively very similar to the reference *m/z* image, and those images were observed to be associated with indices relatively close to the reference index [119]. In contrast, lower indices were associated with noisy images (an artifact of MALDI MSI data acquisition), and higher indices with sparse images. A well-performing function would therefore exhibit a peak centered at the reference index of 783. The mean and kurtosis both show this feature by assigning extreme (higher and lower than most others, respectively) scores to indices close to the reference index.

2.5. Case Studies

Two HNSCC datasets were examined in this study. The first was a gene expression microarray dataset consisting of 25 cancer patient samples. This dataset was obtained from the ArrayExpress repository (ID: E-GEOD-6791), and is described in [125]. This study used the Affymetrix Human Genome U133 Plus 2.0 array platform, which contains 54,675 probes. To obtain gene expression values, raw .CEL files were processed with the robust multi-array average (RMA) algorithm in the Affymetrix Expression Console software. The second was a protein expression dataset consisting of reverse phase protein array (RPPA) data available from The Cancer Proteome Atlas [126]. This dataset consisted of 212 cancer patient samples and described the expression of 187 proteins. For both of these datasets, EGFR (which is up-regulated in more than 80% of HNSCC) was used as the reference gene and protein, respectively.

In order to (1) further investigate the generality of the similarity measure performance, (2) test a metabolomics (lipidomics) dataset, and (3) provide an easy-to-

interpret visual representation of performance, an MSI dataset was also investigated. This was the same experimental MSI data used previously, from a mouse model of Tay-Sachs/Sandhoff disease, and used to profile different lipid species in the brain [82]. The image corresponding to *m/z* 889.6 (located at index 783 within the dataset) was again selected as the reference image due to its distinctive spatial pattern. The MSI data has a spectral dimension of 4,438 *m/z* values, and all *m/z* images were tested as query images against the reference image of *m/z* 889.6.

For all of these datasets, the three-class cases were used for the experiments. Feature (gene, protein, *m/z*)-specific, percentile-based threshold pairs (x, y) were used to bin each expression value into "high" ($> y$), "medium" ($x <$ and $\leq y$) or "low" ($\leq x$) classes. For the gene and protein expression datasets, results from several alternative threshold pairs were compared. For the MSI dataset, the upper threshold y was arbitrarily selected as the 50^{th} percentile of the mean spectrum of the dataset, and the lower threshold x was 0. The piecewise approximation approach was used in all cases. For the gene and protein expression datasets, the primary subsection size was fixed at 10 features. In the MSI dataset, a primary subsection size of 4×4 was chosen after testing several sizes in an effort to balance section size and computational time. The proposed similarity measure score was calculated for each subsection. For the gene and protein expression datasets, the average score across subsections was taken as the overall similarity for the feature pair. For the MSI data, two functions (the mean and kurtosis) for combining subsection scores into an aggregate image pair score were compared. Finally, for all datasets, the top features selected by the proposed similarity measure using three

classes were compared to the top three-class results for cosine similarity, Pearson correlation and mutual information.

2.5.1. Gene Expression

Tables 2.1-2.3 show the top 20 gene rankings for each of the similarity measures for a single EGFR reference probe across three alternative percentile-based thresholds. First, for all cases, the reference probe '211550_at' is selected as the most similar, as it should be. Second, this set of results demonstrates that the multivariate hypergeometric similarity measure is successful in identifying genes which are associated with head and neck cancer. Moreover, across the three alternative thresholds considered, the proposed similarity measure identified 41 probes that were not in the top 20 rankings of the other similarity measures. Among these, 15 genes have been associated with head and neck cancer in recent studies: STX6 [127], BCL2L2 [128], RGS20 [129], SSSCA1 [130, 131], EHD2 [132], SYNPO2L [133], CNR2 [134], HCRP1 [135], CSNK1G2 [136, 137], EFNB1 [138], SH3GL2 [139], KRT31 [140], FKBP1A [141], SLC7A8 [142], and BCL2L14 [143]. In addition, HIBADH [144] and DSG1 [145] have been associated with head and neck cancer on the protein level.

These results also emphasize the value of integrating multiple forms of analysis in order to leverage complementary findings. One option is combining the results from alternative similarity measures. In addition, the benefits of examining a single dataset across alternative thresholds can be clearly observed through the notably different gene lists for each measure in Tables 2.1-2.3. Parallel assessments with different probes for the same gene are also important. For example, the top 20 rankings by the multivariate hypergeometric similarity measure for another EGFR probe gave relevant results such as

ITGBL1 [146] and TMCC1 [147]. Overall, these observations indicate that applying the multivariate hypergeometric similarity measure can yield relevant and useful results.

Table 2.1: Top 20 rankings by similarity measures on head and neck cancer microarray data, using percentiles [25, 50] as thresholds

Multivariate hypergeometric similarity measure		Pearson correlation		Cosine similarity		Mutual information	
Probe	Gene	Probe	Gene	Probe	Gene	Probe	Gene
211550_at	EGFR	211550_at	EGFR	211550_at	EGFR	211550_at	EGFR
202264_s_at	TOMM40	211716_x_at	ARHGDIA	211716_x_at	ARHGDIA	211716_x_at	ARHGDIA
1555140_a_at	BCL2L2	203411_s_at	LMNA	203411_s_at	LMNA	203411_s_at	LMNA
216293_at	CLTA	212103_at	KPNA6 /// LOC100652828 /// LOC100653335	212103_at	KPNA6 /// LOC100652828 /// LOC100653335	219764_at	FZD10
203256_at	CDH3	210128_s_at	LTB4R	1554097_a_at	MIR31HG	1555020_a_at	ARHGAP20
243313_at	SYNPO2L	1554097_a_at	MIR31HG	1555183_at	TERF2	231960_at	BRWD1
212127_at	RANGAP1	206642_at	DSG1	206642_at	DSG1	1566178_x_at	---
1559946_s_at	RUVBL2	210527_x_at	TUBA3C /// TUBA3D	209873_s_at	PKP3	220624_s_at	ELF5
203114_at	SSSCA1	214564_s_at	PCDHGC3	210128_s_at	LTB4R	222834_s_at	GNG12
201415_at	GSS	221001_at	---	210527_x_at	TUBA3C /// TUBA3D	237394_at	---
224314_s_at	EGLN1	230704_s_at	ITGB4	214564_s_at	PCDHGC3	242832_at	PER1
1552618_at	STX6	234743_at	LIMD1	218727_at	SLC38A7	219560_at	C22orf29
200815_s_at	PAFAH1B1	242211_x_at	WDR90	221001_at	---	204889_s_at	NEURL
1569303_s_at	RGS20	209873_s_at	PKP3	221801_x_at	NEFL	216369_at	---
221870_at	EHD2	221801_x_at	NEFL	228587_at	FAM83G	220510_at	RHBG
211716_x_at	ARHGDIA	228587_at	FAM83G	230704_s_at	ITGB4	237361_at	---
203411_s_at	LMNA	234894_at	ITIH6	234743_at	LIMD1	244233_at	---
216060_s_at	DAAM1	241405_at	LOC400604	234894_at	ITIH6	203256_at	CDH3
206642_at	DSG1	1555183_at	TERF2	241405_at	LOC400604	216293_at	CLTA
231955_s_at	HIBADH	218727_at	SLC38A7	242211_x_at	WDR90	1553297_a_at	CSF3R

Table 2.2: Top 20 rankings by similarity measures on head and neck cancer microarray data, using percentiles [25, 75] as thresholds

Multivariate hypergeometric similarity measure		Pearson correlation		Cosine similarity		Mutual information	
Probe	Gene	Probe	Gene	Probe	Gene	Probe	Gene
211550_at	EGFR	211550_at	EGFR	211550_at	EGFR	211550_at	EGFR
234591_at	---	1558378_a_at	AHNAK2	1558378_a_at	AHNAK2	209169_at	GPM6B
222834_s_at	GNG12	204996_s_at	CDK5R1	204996_s_at	CDK5R1	209505_at	NR2F1
220624_s_at	ELF5	206447_at	CELA2A /// CELA2B	206447_at	CELA2A /// CELA2B	208803_s_at	SRP72
233918_at	DCDC2B	234591_at	---	234591_at	---	227772_at	LATS1
1558378_a_at	AHNAK2	201183_s_at	CHD4	201183_s_at	CHD4	201893_x_at	DCN
216176_at	HCRP1	222834_s_at	GNG12	220624_s_at	ELF5	203685_at	BCL2
220365_at	ALLC	233918_at	DCDC2B	222834_s_at	GNG12	203822_s_at	ELF2
234749_s_at	POC1A	220624_s_at	ELF5	233918_at	DCDC2B	205880_at	PRKD1
1553137_s_at	KLF11	221115_s_at	LENEP	1552575_a_at	C6orf141	208249_s_at	TGDS
1559205_s_at	---	1552575_a_at	C6orf141	1553137_s_at	KLF11	208990_s_at	HNRNPH3
221115_s_at	LENEP	1559205_s_at	---	1554912_at	ESYT3	211896_s_at	DCN
221183_at	LOC100507388	217175_at	UGT2B15	1559075_s_at	BAHCC1	212462_at	KAT6B
204996_s_at	CDK5R1	217308_at	OR1F2P	1559205_s_at	---	213058_at	TTC28
1567068_at	OR4D1	220944_at	PGLYRP4	210553_x_at	LOC100507472 /// PCSK6	217954_s_at	PHF3
206586_at	CNR2	221629_x_at	FAM203A	213445_at	ZC3H3	220624_s_at	ELF5
206642_at	DSG1	227672_at	C8orf73	213681_at	CYHR1	222834_s_at	GNG12
225500_x_at	SCAF1	232409_x_at	FBXL16	217175_at	UGT2B15	225230_at	DRAM2
236614_at	LOC729683	1554912_at	ESYT3	217308_at	OR1F2P	226402_at	CYP2U1
237111_at	LOC388942	1559075_s_at	BAHCC1	220944_at	PGLYRP4	233918_at	DCDC2B

Table 2.3: Top 20 rankings by similarity measures on head and neck cancer microarray data, using percentiles [50, 75] as thresholds

Multivariate hypergeometric similarity measure		Pearson correlation		Cosine similarity		Mutual information	
Probe	Gene	Probe	Gene	Probe	Gene	Probe	Gene
211550_at	EGFR	211550_at	EGFR	211550_at	EGFR	211550_at	EGFR
214119_s_at	FKBP1A	201183_s_at	CHD4	201183_s_at	CHD4	201183_s_at	CHD4
206677_at	KRT31	1559205_s_at	---	1559205_s_at	---	1561319_at	OTX2-AS1
1561509_at	---	224239_at	DEFB103A /// DEFB103B	1570531_at	---	219764_at	FZD10
205938_at	PPM1E	1570531_at	---	224239_at	DEFB103A /// DEFB103B	231960_at	BRWD1
234191_at	BCL2L14	231737_at	CACNG4	1552575_a_at	C6orf141	221021_s_at	CTNNBL1
205341_at	EHD2	217175_at	UGT2B15	217175_at	UGT2B15	1561633_at	HMGA2
202945_at	FPGS	220021_at	TMC7	220021_at	TMC7	217237_at	---
229432_at	NAGS	1552575_a_at	C6orf141	220944_at	PGLYRP4	221069_s_at	TACO1
244258_at	---	221629_x_at	FAM203A	221259_s_at	TEX11	230810_at	JMJD4
1557514_a_at	---	231243_s_at	BHLHE41	221629_x_at	FAM203A	236119_s_at	SPRR2G
208337_s_at	NR5A2	231908_at	ZDHHC18	231243_s_at	BHLHE41	1570531_at	---
233069_at	PPP4R1L	232718_at	LINC00589	231737_at	CACNG4	212103_at	KPNA6 /// LOC100652828 /// LOC100653335
215835_at	LOC100653174	237430_at	---	231908_at	ZDHHC18	224239_at	DEFB103A /// DEFB103B
234933_at	CC2D2A	220944_at	PGLYRP4	232718_at	LINC00589	237900_at	KLHDC4 /// LOC100652950 /// LOC100653213
216603_at	SLC7A8	244202_at	---	237430_at	---	221812_at	FBXO42
202711_at	EFNB1	221259_s_at	TEX11	244202_at	---	235748_s_at	---
202574_s_at	CSNK1G2	1558871_at	---	1558871_at	---	244258_at	---
235748_s_at	---	231295_at	ME3	1561048_at	RARS2	59705_at	SCLY
205751_at	SH3GL2	217234_s_at	EZR	1564022_at	ZNF804B	207065_at	KRT75

2.5.2. Protein Expression

The similarity measure results for RPPA data are shown in Tables 2.4-2.6, for three alternative threshold selections. For all threshold selections – and for all similarity

Table 2.4: Top 20 rankings by similarity measures on head and neck cancer RPPA data, using percentiles [25, 50] as thresholds

Rank	Multivariate hypergeometric similarity measure	Pearson correlation	Cosine similarity	Mutual information
1	EGFR	EGFR	EGFR	EGFR
2	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068
3	E-Cadherin	eEF2K	eEF2K	eEF2K
4	eIF4G	VHL	VHL	VHL
5	VHL	Akt	Akt	Akt
6	HER2	p70S6K	p70S6K	E-Cadherin
7	Akt	beta-Catenin	beta-Catenin	Bap1-c-4
8	eEF2K	Tuberin	Tuberin	Tuberin
9	Tuberin	mTOR	mTOR	beta-Catenin
10	c-Jun_pS73	E-Cadherin	E-Cadherin	p70S6K
11	p70S6K	ERK2	ERK2	ERK2
12	Ku80	HER2	HER2	mTOR
13	mTOR	eIF4G	eIF4G	eIF4G
14	beta-Catenin	Bap1-c-4	Bap1-c-4	c-Met_pY1235
15	PDK1_pS241	PDK1_pS241	PDK1_pS241	SF2
16	INPP4B	Chk1_pS345	Chk1_pS345	CD31
17	c-Myc	Ku80	Ku80	Bax
18	ACC1	STAT5-alpha	STAT5-alpha	HER2
19	B-Raf	TSC1	TSC1	MEK1_pS217_S221
20	ERK2	B-Raf	B-Raf	c-Kit

Table 2.5: Top 20 rankings by similarity measures on head and neck cancer RPPA data, using percentiles [25, 75] as thresholds

Rank	Multivariate hypergeometric similarity measure	Pearson correlation	Cosine similarity	Mutual information
1	EGFR	EGFR	EGFR	EGFR
2	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068
3	E-Cadherin	eEF2K	eEF2K	eEF2K
4	eEF2K	VHL	VHL	CD31
5	eIF4G	Chk1_pS345	Chk1_pS345	VHL
6	mTOR	eEF2	eEF2	Bcl-2
7	Tuberin	beta-Catenin	beta-Catenin	PDK1
8	HER2	mTOR	mTOR	DJ-1
9	VHL	PDK1_pS241	PDK1_pS241	14-3-3_epsilon
10	p70S6K	Tuberin	Tuberin	eIF4G
11	ERK2	HER2_pY1248	HER2_pY1248	Akt
12	Bap1-c-4	INPP4B	INPP4B	PEA-15
13	TSC1	E-Cadherin	E-Cadherin	Bap1-c-4
14	Chk1_pS345	eIF4G	eIF4G	beta-Catenin
15	beta-Catenin	p70S6K	p70S6K	eEF2
16	Ku80	ERK2	ERK2	E-Cadherin
17	Akt	ACC1	ACC1	Tuberin
18	GSK3-alpha-beta	Paxillin	Paxillin	p70S6K
19	PDK1_pS241	MEK1	MEK1	Chk1_pS345
20	ACC1	Bap1-c-4	Bap1-c-4	p27

measures – the top-ranked protein was EGFR itself, as expected. The second-most similar protein was phosphorylated EGFR (Tyr1068). Overall, the selections among the different measures were highly congruent. However, among the three threshold-cases, there were several cases where relevant HNSCC-relevant proteins were selected by the multivariate hypergeometric similarity measure, but not by others. These included well-known cancer-related proteins like c-Myc [148], phosphorylated c-Jun [149], HER2 [150], and NF-kB [1], as well as proteins which have been implicated in HNSCC in recent studies, like INPP4B [151, 152] and ACC1 and AMPK [153]. Others highlighted only by the multivariate hypergeometric similarity measure in this case study were GSK3-alpha-beta [154], Ku80 [155], and TSC1 [156].

Table 2.6: Top 20 rankings by similarity measures on head and neck cancer RPPA data, using percentiles [50, 75] as thresholds

Rank	Multivariate hypergeometric similarity measure	Pearson correlation	Cosine similarity	Mutual information
1	EGFR	EGFR	EGFR	EGFR
2	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068	EGFR_pY1068
3	mTOR	mTOR	mTOR	CD31
4	p70S6K	VHL	VHL	mTOR
5	PDK1_pS241	PDK1_pS241	PDK1_pS241	VHL
6	beta-Catenin	HER2_pY1248	HER2_pY1248	p70S6K
7	VHL	Chk1_pS345	Chk1_pS345	PDK1_pS241
8	E-Cadherin	p70S6K	p70S6K	Akt
9	AMPK_pT172	eEF2K	eEF2K	14-3-3_epsilon
10	eEF2K	Tuberin	Tuberin	SF2
11	Ku80	E-Cadherin	E-Cadherin	PDK1
12	Tuberin	INPP4B	INPP4B	eEF2K
13	eIF4G	eIF4G	eIF4G	beta-Catenin
14	HER2	beta-Catenin	beta-Catenin	PRDX1
15	Chk1_pS345	Ku80	Ku80	Bax
16	ACC_pS79	Dvl3	Dvl3	Caspase-7_cleavedD198
17	p90RSK	ERK2	ERK2	HER2_pY1248
18	NF-kB-p65_pS536	p90RSK	p90RSK	E-Cadherin
19	c-Jun_pS73	ACC1	ACC1	c-Met_pY1235
20	INPP4B	MEK1	MEK1	Dvl3

2.5.3. Mass Spectrometry Imaging (Lipidomic) Data

In the final set of results, the similarity measure was applied to the experimental MSI data. The top 12 *m/z* images selected by each measure are shown in Figure 2.11. All measures agree that the reference itself is the most similar (selection 1: *m/z* 889.6). Notably, the proposed similarity measure gives results which are qualitatively very similar to the reference *m/z* image. The Pearson correlation and mutual information results for three classes also closely resemble the reference *m/z* image, while the cosine similarity results for three classes include several noisy images without a clearly discernible pattern. Interestingly, the top 12 results selected by the proposed similarity

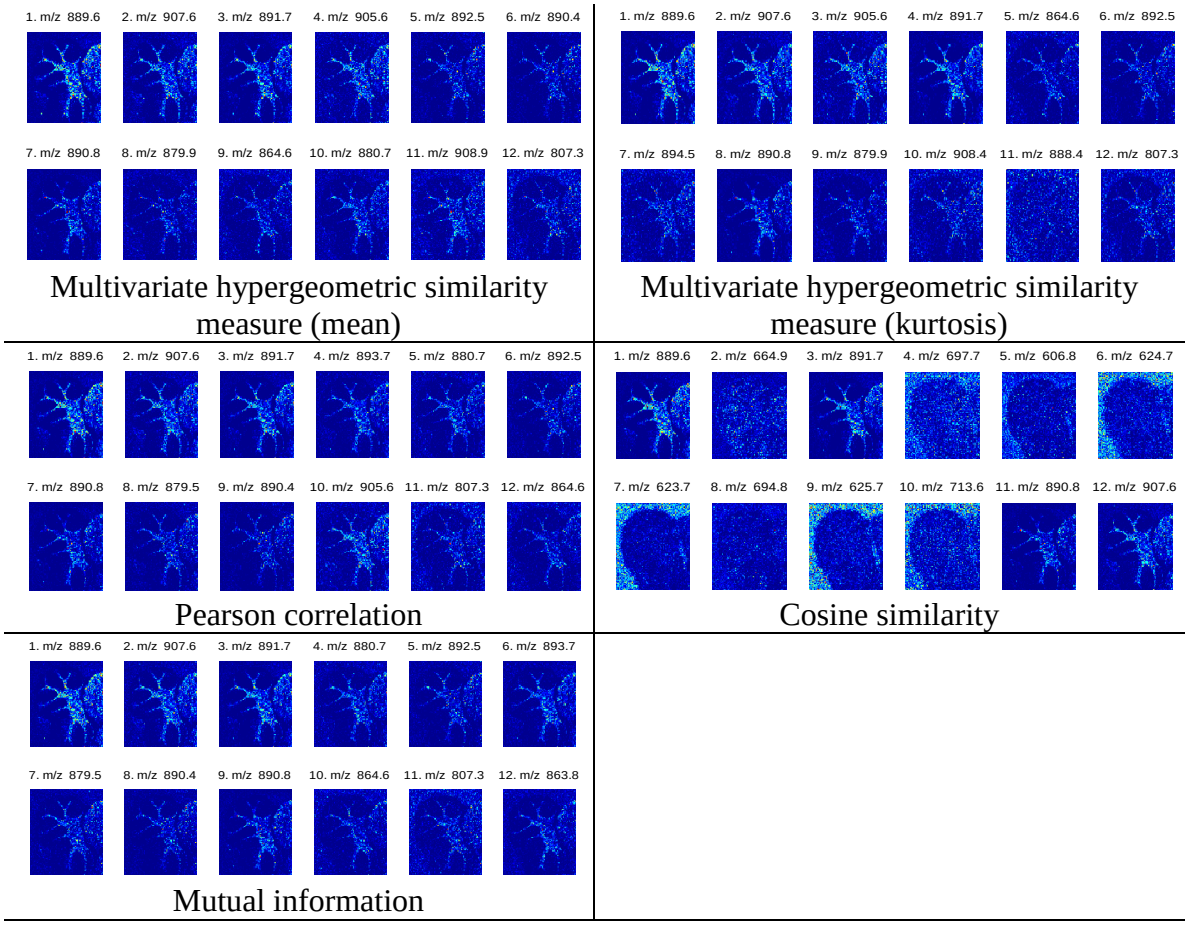


Figure 2.11: The 12 most similar *m/z* images, as ranked by the four similarity measures. The multivariate hypergeometric similarity measure results are shown with the mean and kurtosis as combination functions.

measure using the mean and kurtosis as combination functions are not identical. Moreover, neither set of results overlaps completely with the results from Pearson correlation, mutual information, and cosine similarity. For example, the proposed similarity measure, using the mean as the combination function, selects *m/z* 908.9, which none of the others select. Similarly, *m/z* 894.5, another unique selection, is picked by the proposed similarity measure when using the kurtosis as the combination function. Examining the top *n* results is common when applying a similarity measure to a dataset, and these observations indicate that applying the proposed multivariate hypergeometric similarity measure can yield relevant and useful results.

2.6. Discussion and Key Innovations

This chapter describes the design, development, and testing of two similarity measures. The second, the multivariate hypergeometric similarity measure, is the main result. It enables the pairwise comparison of images and data vectors featuring any positive integer number of intensity levels. This is an extension of initial work on the hypergeometric similarity measure, which was restricted to binary data. Using synthetic datasets, the proposed multivariate measure was compared to Pearson correlation, cosine similarity and mutual information in terms of sample rankings, and identified several favorable properties of the proposed measure. Next, a method of piecewise approximation was developed to facilitate the application of this approach to large datasets. Piecewise approximation was tested at several different subsection sizes on synthetic data, and was observed to follow the trend of the exact score. Functions for combining subsection similarity scores found through piecewise approximation were empirically assessed using biological data. The proposed similarity measure was tested on two HNSCC datasets: gene expression microarray data and reverse phase protein array

(RPPA) data. The proposed similarity measure was also demonstrated to be effective in identifying qualitatively similar images in a lipidomics MSI dataset. Critically, for all datasets, it made relevant selections which were not identified by other similarity measures in their top selections.

The results of this study highlight several avenues for further research on the multivariate hypergeometric similarity measure. For instance, this approach is defined for any positive integer number of classes, but the results in this study have considered only three classes. Three classes were chosen both for simplicity in examining similarity measure properties and to highlight the difference between the binary case and the multi-class case. Future research can assess the effect of increasing the number of classes. However, as previously noted, the generation of the isomarginal family becomes increasingly demanding as the number of classes increases [121-124]. Additionally, alternative definitions of the statistic $S(k)$ will be explored. Here, I chose $S(k)$ as the set of diagonal elements of the contingency table. In the future, it may be desirable to include sub- and super-diagonal terms when larger numbers of classes are considered. The selection of the appropriate number of classes – and of appropriate thresholds for separating classes – is another issue of interest. In this study, several percentile-based thresholds between classes were compared for the gene and protein expression datasets. From the perspective of practical biomedical applications, choices of thresholds for a particular dataset may be based on examination of descriptive data statistics, or by applying selected tests as a preliminary step [157]. The selection of functions for aggregating subsection scores obtained from piecewise approximation is another area for further study. Six functions were tested in this study, and many additional functions could

be tested. Interestingly, the set of top selections using the mean and kurtosis were not identical, indicating that it may also be useful to consider which combination functions may be complimentary.

The Key Innovations of this chapter are:

- Development of binary hypergeometric similarity measure using Fisher's exact test
- Development of multivariate hypergeometric similarity measure using the Fisher-Freeman-Halton test
- Development of a piecewise approximation algorithm to facilitate application of the multivariate hypergeometric similarity measure to high-dimensional data vectors
- Implementation on two HNSCC (transcriptomic and proteomic) and one non-HNSCC (MSI, metabolomic / lipidomic) datasets indicates that the proposed multivariate hypergeometric similarity measure makes relevant selections not identified by other similarity measures

CHAPTER 3

DETECT-TLC: EXPLORATORY DATA MINING FOR METABOLOMICS

3.1. Data Acquisition for Metabolomics

Metabolomics offers a perspective of the small molecules, including lipids, within an organism or patient [158, 159]. Compared to other –omics levels, the "chemical fingerprint" measured through metabolomics is highly dynamic, and has been shown to be a promising direction for the diagnosis and monitoring of disease [160, 161]. In HNSCC in particular, metabolomics approaches are demonstrating promising results for disease detection and early diagnosis [34-36]

The key data acquisition methodologies used in metabolomics are ^{1}H NMR and mass spectrometry. Mass spectrometry is used both alone and coupled to liquid (LC) and gas (GC) chromatography (LC-MS and GS-MS, respectively) [162, 163]. The reason for this coupling is because the chromatography step staggers the input sample flow to the mass spectrometer according to size, charge, or other properties, thereby generating sparser and easier-to-interpret mass spectra.

3.1.1. Coupling Thin Layer Chromatography with Mass Spectrometry Imaging

While LC and GC both return 1D data – i.e., spectra with intensities on the vertical axis and retention time on the horizontal axis – thin layer chromatography (TLC) is a 2D chromatographic separation process. Figure 3.1 shows how separated mixture components appear as spots on a TLC plate. TLC is a commonly used technique in synthetic and organic chemistry for the separation of complex mixtures due to its

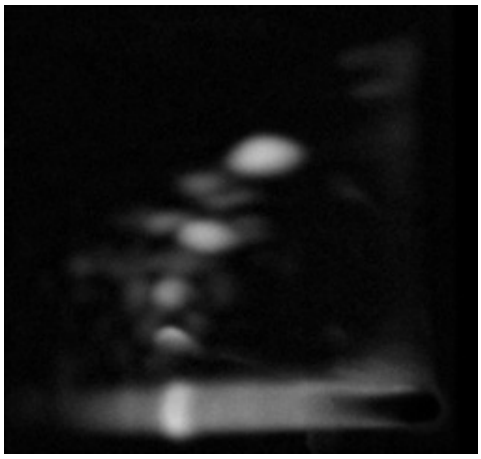


Figure 3.1: Optical image of a TLC plate. Image courtesy of Fernández Lab at Georgia Institute of Technology.

simplicity and speed [164]. In metabolomics, TLC alone has been applied to study bacteria [165, 166], but it is frequently combined with mass spectrometry analysis.

Due to the 2D nature of TLC, it can be coupled with mass spectrometry either by assessing an individual spot using conventional mass spectrometry, or by interrogating the entire TLC plate through MSI [167-170]. MSI analysis of TLC plates has been performed using different mass spectrometry ionization approaches, including including matrix assisted laser desorption ionization (MALDI) [171] and desorption electrospray ionization (DESI) [172, 173].

The advantage of TLC-MSI coupling is the molecular-level resolution: instead of being restricted to spots visible on the TLC plate, the MSI datacube can describe thousands of measurable spots. Examining a TLC-MSI dataset is straightforward if the analyte of interest is known, but for exploratory data mining purposes, the volume of data presents a challenge. This provides an opportunity for knowledge-driven mining in terms of implicit similarity: the goal is to identify all *m/z* images containing regions similar to a TLC spot, regardless of its spatial location or orientation. Currently, the state-of-the-art is

manual inspection of the thousands of images in the MSI dataset to detect such images of interest, which is a substantial data processing bottleneck. This chapter presents the development, testing, and validation of DetectTLC, a software tool for automatically detecting *m/z* images containing regions similar to TLC spots.

3.2. Development of Image Feature-Based Modeling Tool

The hypothesis behind DetectTLC is that *m/z* images containing spot-like regions are distinguishable from other images on the basis of quantitative image features. DetectTLC utilizes a five-step image processing pipeline, culminating in the extraction of such features. In the first step, smoothing filters are used to remove background noise from the *m/z* images, and very sparse and noisy images are excluded based on pixel counts. In the second step, the continuous-intensity *m/z* images comprising the MSI dataset are converted into binary images. In the third step, morphological image processing operations are used to fill in small holes in the binary *m/z* images. In the fourth step, quantitative image features are extracted for each *m/z* image in the dataset, with the goal of associating more extreme feature values with *m/z* images which contain TLC spot-like regions. In the fifth and final step, the *m/z* images are ranked in terms of the quantitative image features and are visualized in the graphical user interface. Alternative combinations of these steps were compared in order to identify well-performing pipelines. Each of these steps is discussed in detail in the following section.

3.2.1. Image Processing Pipeline

Step 1: Smoothing and Pixel-count Filters

Median filtering was used to remove background noise. For the MSI datasets examined in this study, 5×5 and 7×7 median filters were compared, but the difference in

performance was small compared to the effects of other factors, so only results from 7×7 filters are shown. Median filtered-results are also compared with un-filtered results.

The pixel-count-based filter was useful for removing sparse and streak-filled images from consideration. It was observed that many DESI-MSI images were sparse or streak-filled. In different datasets, the necessity of performing this filtering step may vary, as few such images may exist. Typically, binary *m/z* images with fewer than 5 and more than 1500 non-zero pixels were removed from consideration. Results with and without this filtering step were compared.

Step 2: Generating Binary Images

Two different methods for generating binary images were compared in this study. The first is a manually-selected threshold: if any signal S was present at a pixel (x, y) in the original *m/z* image above the threshold value T (i.e., $S(x, y) > T$), the value of the pixel in the binary image $B(x, y) = 1$. Otherwise, $B(x, y) = 0$. Users may select the desired threshold through the Advanced Options menu of the DetectTLC interface, which also provides a visualization of the selected threshold with respect to the average spectrum of the dataset. The second technique is Otsu's method, which selects the threshold at which the within-class variance of the pixels assigned to each label is minimized [174].

Step 3: Morphological Operations

As shown in Figure 3.2, an *m/z* image may feature a spot-like region that is not solid, i.e., single pixels or clusters of a few pixels where no or low signal was detected may occur between pixels where signal was detected. To the user, this area is interpreted

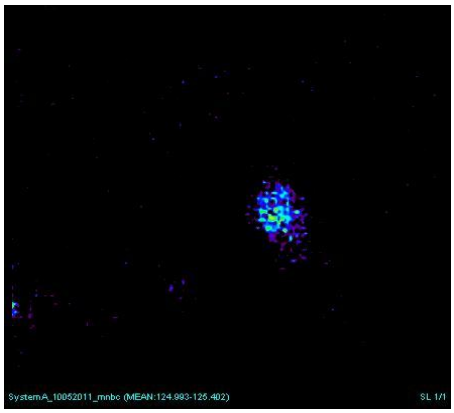


Figure 3.2: Example of a TLC spot-like region in an *m/z* image. Note that the spot consists of both high- and low-intensity pixels.

as a single spot-like region regardless. However, these discontinuities can influence the automated, image-feature based detection of spot-like regions. To address this issue, DetectTLC applies morphological image processing operations to the MSI dataset. Erosion and dilation are two basic morphological operators used in image processing. A structuring element of a particular shape – common shapes include disks, squares, and diamonds – is used to remove (in erosion) or add (in dilation) a layer of pixels from the image. The compound morphological operators of opening and closing are defined in terms of erosion and dilation: in opening, erosion is followed by dilation, and in closing, dilation is followed by erosion. In this study, we compared the performance of dilation and opening in generating homogenous spot-like regions: dilation fills in small holes in a single region and gaps between regions, while opening removes connections between separate regions.

Step 4: Scoring Based on Quantitative Image Feature Values

The performances of eight quantitative image features were investigated and compared in the development of DetectTLC. These included seven shape-based features:

area, compactness, convex area, eccentricity, extent, number of connected regions, and solidity; and one texture-based feature: entropy. Each of these features is described further in Table 3.1.

Table 3.1. Definition and description of image features investigated in the development of DetectTLC.

Image feature	**Definition**	**Description**
Area (*A*)	$A = \sum f(x)$, where x represents pixels with a value of one in the binary image, and f is a neighborhood operation function.	This feature is the weighted sum of pixels with a value of one in the binary image. Different spatial distributions of pixels are weighted. Images containing a spot-like region may have lower area values than images with other structures.
Compactness (*Co*)	$Co = P^2 / A$, where P is the perimeter of the non-zero region and A is its area.	Compactness is a regional descriptor defined as the ratio of an object's squared perimeter to its area. Compactness is minimal for disk-shaped regions [115], so images with a spot-like region may be characterized by lower compactness values.
Convex area (*Ca*)	$Ca = \sum x$ where x represents pixels which are in the convex hull of the image.	The convex area is the number of pixels inside the convex hull, which is the smallest convex polygon that contains the entire region of non-zero pixels. A smaller convex area implies a small, cohesive region of interest, so images containing a spot-like region may be characterized by lower convex area values.
Eccentricity (*Ec*)	$Ec = Dc / Dv$, where Dc is the distance from the center to the focus of the ellipse, and Dv is the distance from the center to a vertex.	Eccentricity is calculated by fitting an ellipse to the region of interest, such that the ellipse and the region share the same second moments. The image feature is then the eccentricity of the fitted ellipse. For a circular region, eccentricity would be 0; for a line it would be 1. Images containing a spot-like region may be characterized by lower eccentricity values.

Table 3.1 continued overleaf.

Table 3.1, continued

Entropy (*En*)	$En = -\sum p_i log_2(p_i)$, i = 0...1, where p_i represents the fraction of zero (i = 0) and non-zero (i = 1) pixels in the image.	Entropy is a measure of randomness used to describe image texture. For binary images, entropy is defined in terms of the fractions of zero and non-zero pixels. The quantity is maximized when the fraction of each pixel type is equal, so images with larger spot-like regions may be characterized by higher entropy values.
Extent (*Ex*)	$Ex = \sum x / \sum B$, where x represents pixels which are within the region, and B represents all pixels which are within the bounding box.	Extent is defined in terms of the bounding box, which is the smallest rectangular region that completely encloses the region. Extent measures the proportion to which the region of interest fills the bounding box. Images with a spot-like region may be characterized by higher values of *Ex.*
Number of connected regions (*Re*)	$Re = E + H$, where E is the Euler number and H is the number of holes.	The number of connected components is related to the Euler number (*E*), a commonly used shape-based image feature. The Euler number is defined as the difference between the number of connected components and the number of holes (*H*). In an image with solid spot-like feature, ideally there would be no holes (*Re = E)*, so the number of connected components was considered as a feature instead of Euler's number. Images with several distinct spot-like regions may be characterized by higher values of *Re.*
Solidity (*S*)	$S = A / Ca$, where A is the image feature Area and Ca is the image feature Convex area.	Solidity is a composite feature, defined in terms of Area and Convex Area. This image feature measures the fraction of pixels which are in both the convex hull and the region of interest. Images with a spot-like region may be characterized by higher values of *S.*

3.2.2. Features of Graphical User Interface

The user interface comprises four windows, with the main window shown in Figure 3.3. This window displays the spot-containing images identified through different

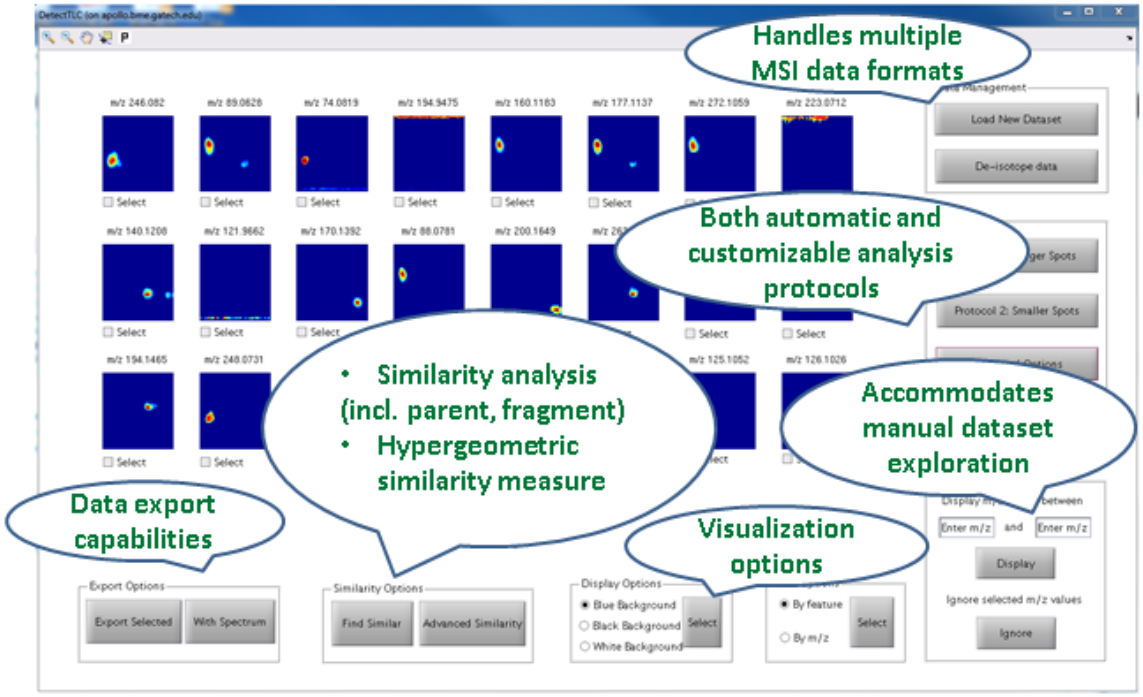


Figure 3.3: The main graphical user interface with the top 24 images containing spot-like regions are displayed. Data is first uploaded and (optionally) de-isotoped, following which the user may select from "Protocol 1" or "Protocol 2" for feature selection, or design their own processing pipeline through the "Advanced Options" tab. Spots with similar spatial distributions may be identified using the "Similarity Options". Selected images and/or spectra may be exported using the "Export Options".

algorithms, and contains a control panel for accessing the "Load New Dataset", "Advanced Options", and "Similarity Options" menus. Algorithm results are displayed on the main graphical user interface (GUI), 24 *m/z* images at a time. After running a processing protocol, images are initially sorted based on a quantitative image feature-based score. Within each window, they may be re-sorted by ascending *m/z* value for convenience. A scroll bar is used to scan though all images, which may also be visualized with alternate color schemes if desired. Additionally, the user can narrow the examined mass range for more targeted examination of images containing spot-like regions. The main user options in the DetectTLC GUI are described below.

Load New Dataset

DetectTLC accommodates MSI data in Analyze 7.5 and mzXML format (with time and position information). It also accepts MSI data which have been imported into MATLAB and saved as matrices in '.mat' files. Thus, other MSI data formats can also be used with DetectTLC if they are first imported into MATLAB.

De-isotope data

DetectTLC currently uses a basic de-isotoping algorithm in which the highest-intensity *m/z* value in each 3 Da-window is retained. Each spectrum is de-isotoped individually, and the spectra are then re-assembled into a de-isotoped datacube. After de-isotoping is performed, all further processing protocols will automatically be performed on the de-isotoped data. In order to return to the raw data (with no de-isotoping), it is necessary to re-load the data files.

Visualization

The main graphical user interface displays 24 *m/z* images at a time. The scroll bar at the bottom of the tool screen allows users to scroll through the dataset, showing *m/z* images 25-48, 49-72, etc. Whenever a processing protocol is implemented, the interface will show the top 24 *m/z* images according to that protocol, and the user can scroll through the rest of the ordered selections. Additionally, the user can select among three color schemes to customize the visualization in order to enhance detection of relevant *m/z* images.

Pre-set automatic protocols

Two automatic protocols are offered. The first, "Protocol 1: Larger spots", uses entropy as the quantitative image feature for scoring. The second, "Protocol 2: Smaller spots", uses compactness as the image feature. The default settings for filtering by intensity, filtering by non-zero pixels and median filtering are implemented in both of these protocols.

Refresh current dataset

By pressing the "Refresh dataset" button, the user can return the display to the original MSI data before any processing protocols (pre-set or via the advanced options) were applied. If the dataset had been de-isotoped, the de-isotoped data will be shown.

Find m/z values

This option displays all images corresponding to the user-input *m/z* range. Any processing steps that are called after the "Find *m/z* values" command will operate only on the images within that *m/z* range. To process the entire dataset, it is necessary to press the 'Refresh dataset' button first.

Ignore m/z values

This feature can be used to select *m/z* images which are not of interest to the user (e.g., noisy images, or images with homogeneous signal intensity) and remove them from the current dataset view. The original MSI data can be retrieved by using the "Refresh dataset" option.

Sort by image feature value or m/z value

As a default, the *m/z* images returned by any processing protocol are sorted by the quantitative image feature value. To facilitate review, the 24 images within an individual screen may also be sorted in order of ascending *m/z* value.

Export (with or without average spectrum)

Two different export utilities are available in DetectTLC: (1) Export Selected Images and (2) Export With Spectrum. In (1) Export Selected Images, the user can use the checkboxes below each *m/z* image to make selections, and then click the button labeled "Export Selected Images". All selected images will be saved in '.fig' format to the user-specified directory, and an ASCII file listing the selected *m/z* values will be created in the same directory. Multiple *m/z* images within one screen (24 images) can be exported simultaneously. In (2) Export With Spectrum, the user will instead click the button labeled "With Spectrum". A new figure will appear, showing the selected *m/z* image. The user can use the drawing cursor to select a region of interest by drawing a line through the spot-like region or around it. After the region of interest is selected, a composite figure containing both the *m/z* image and the average spectrum in the region of interest (Figure 3.4) will be saved as a MATLAB '.fig' file in the user-specified directory. Again, multiple *m/z* images (24 per screen) can be selected simultaneously. If multiple *m/z* images are selected, the process of manual region of interest selection will be repeated for each image, and an ASCII file listing the selected *m/z* values will be created in the same directory.

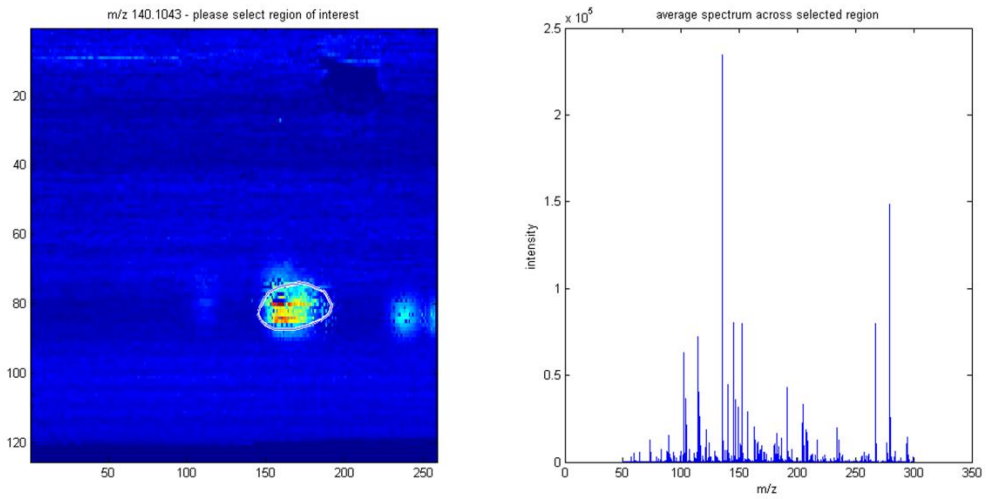


Figure 3.4: Example of selected *m/z* image (*m/z* 140.1043) to draw a region of interest (ROI, outlined in white) and resulting average spectrum for selected pixels.

Advanced Options GUI

The Advanced Options GUI (Figure 3.5) provides users with more control over how the MSI data is processed and analyzed. Four different pre-processing control panels are available: (1) Generation of binary images; (2) Pixel-count filtering; (3) Median filtering; and (4) Image feature selection.

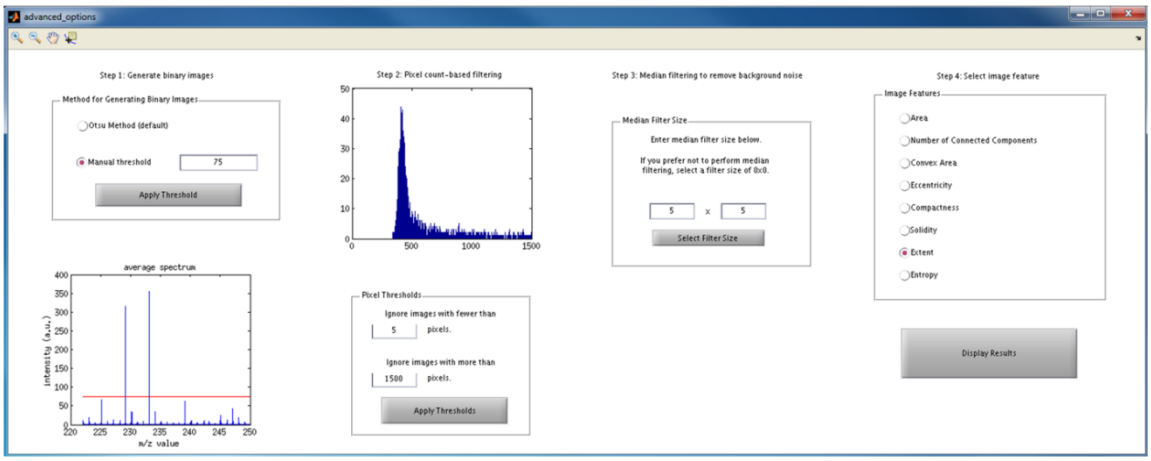


Figure 3.5: Advanced Options GUI in DetectTLC.

In DetectTLC, morphological operations and image feature scores are computed on binary *m/z* images. In the "Generation of binary images" control panel, the user has the option of using Otsu's method (default) or manually selecting a threshold for generating binary *m/z* images. The success of the thresholding process is highly dependent on spectral signal-to-noise; it is valuable to identify genuine spots that may have low abundance, but that are still sufficiently above baseline noise. Manual threshold selection allows for the user's knowledge of the spectral quality to be factored into the processing, but Otsu's method for threshold selection provides a satisfactory approach without the need for user input. In the case studies presented here, manual threshold selection and Otsu's method yielded comparable results across all other processing variables. The average spectrum across the MSI dataset is displayed below the thresholding panel in the GUI, and when the "Apply" button is clicked, the manually-selected threshold is overlaid on the spectrum as a red line.

The "Pixel-count filtering" control panel can be used to eliminate sparse images (i.e., *m/z* images with non-zero signal in very few pixels) and so-called streaky images (i.e., *m/z* images with high intensity signal in many pixels, but in a noisy, non-informative spatial pattern). In the MSI dataset analyzed in this paper, sparse images were generated as a result of the centroiding process. These images were eliminated by establishing a minimum of 5 pixels for a spot to be detected. Conversely, no more than 1500 pixels for a particular *m/z* could be present for a true spot, as the presence of that many pixels indicated streaks or widespread presence of a species across the entire TLC plate (e.g., an impurity in the DESI solvent). For general use, a histogram showing the distribution of *m/z* images with different numbers of non-zero pixels is displayed. The user can refer to

this histogram to select the upper and lower thresholds for pixel-count filtering. When the "Apply" button is clicked, all *m/z* images containing a number of non-zero pixels above the upper threshold or below the lower threshold are discarded from the dataset. The default setting is to discard images with fewer than 5 non-zero pixels and with more than 1500 non-zero pixels.

In the "Median filtering of *m/z* images" panel, the user can select the size of the two-dimensional median filter applied to remove "salt-and-pepper" background noise from the *m/z* images. The default setting is a 5×5 median filter. If no median filtering is desired, the filter size should be set to 0×0.

The fourth and final panel allows the user to select from among the eight image features investigated in this paper: area, compactness, convex area, eccentricity, entropy, extent, number of connected regions, and solidity. The *m/z* images remaining after the filtering steps, sorted according to the selected image feature, will be displayed in the main GUI. Selection of the image feature of interest is independent of the three pre-processing options. It is not necessary to perform any pre-processing before applying the image feature-based sorting – the default settings of Otsu's method, < 5, > 1500 pixel-count filtering, and 5×5 median filtering will be applied. The three pre-processing steps can also be applied individually or in any combination prior to selecting an image feature.

Similarity Assessment

Multiple protocols for performing similarity analysis are available in DetectTLC. The most basic method, which is implemented by selecting any *m/z* image via its

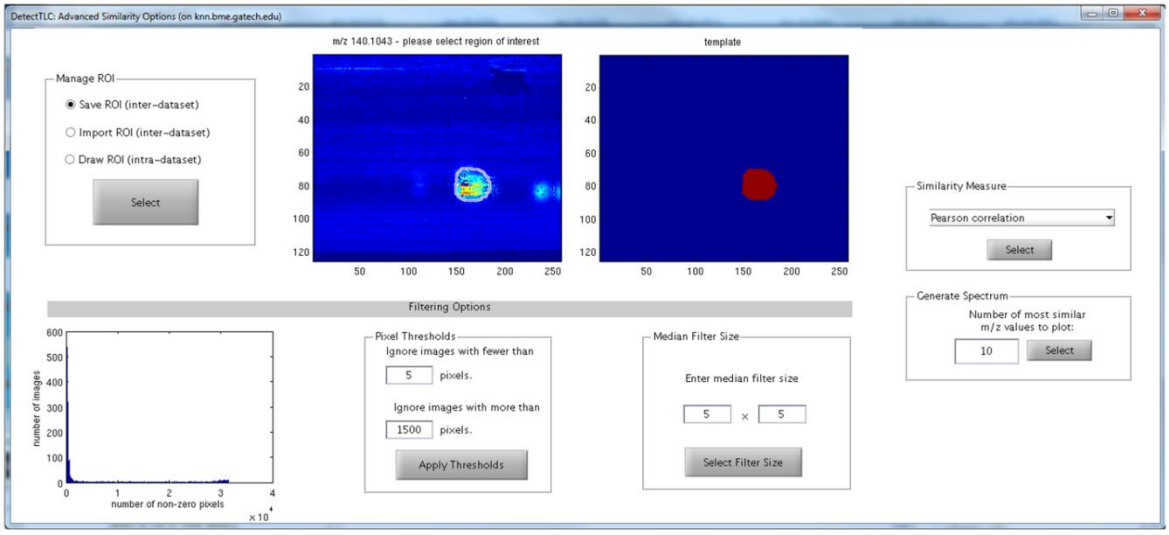


Figure 3.6: Advanced Similarity Options GUI in DetectTLC. The binary template corresponding to the selected *m/z* image for *m/z* 140 is shown as an example.

checkbox and clicking the "Find Similar" button, will return the most similar *m/z* images within the same dataset, as ranked by the binary hypergeometric similarity measure (described in Chapter 2). Alternatively, by clicking the "Advanced Similarity" button, the Advanced Similarity Options window will open, as shown in Figure 3.6.

The Advanced Similarity Options window provides three methods for managing the region of interest (ROI) for similarity assessment. These are: (1) Create and save a new ROI template (may then be used with the same or another dataset, i.e., "inter-dataset"); (2) Import an existing ROI template (may be from the same or another dataset, i.e., "inter-dataset"); (3) Create and implement an ROI template on the current dataset (i.e., "intra-dataset"). These options are described further as follows.

The first option, "Save ROI (inter-dataset)" enables the user to draw a binary ROI template and save it for later use, with the same or another MSI dataset having the same spatial dimensions. If the checkbox of any *m/z* image was selected in the main GUI, that *m/z* image will be provided as the guide for drawing the binary template. If not, the

average image across the loaded dataset will be provided. Once the template is drawn, the user may save it to a selected filename and directory. If a particular *m/z* image is used as the template basis, three variables are saved: the binary ROI, the *m/z* vector from current dataset, and the peak height of selected *m/z* in the current average spectrum. When the template is imported later, this data will be used to draw a spectrum of similar peaks. If the average *m/z* image is used as the template basis, only the binary ROI is saved.

In the second option, "Import ROI (inter-dataset)", a previously drawn ROI can be loaded. In order to use this ROI for similarity assessment, it must have the same spatial dimensions as the currently loaded dataset. Some variations in spatial dimensions can be handled by DetectTLC. These are:

1. Template image is rotated 90 degrees with respect to current dataset. DetectTLC will rotate the template so that the dimensions match.
2. Template image has one extra row and/or column. DetectTLC will delete the first row and/or column from the dataset.
3. Template image is rotated 90 degrees and has an extra row and/or column.

Before any of these actions are taken, DetectTLC will notify the user that the spatial dimensions of the current dataset and selected template do not match. If the mismatch falls into any of these three categories, DetectTLC will prompt as to whether the template should be automatically adjusted. If the mismatch does not fall into these three categories, DetectTLC will prompt the user to load a different template for use with the current dataset.

In the third option, "Draw template (intra-dataset)", the user can draw a template which will immediately be used for similarity assessment on the currently loaded dataset,

and which will not be saved. This differs from basic "Find Similar" option in that the user may select which similarity metric to use, and the spectrum of most similar peaks may be plotted.

After a template is available (either through the Import or Draw options), the median filter, pixel-count filter, "Choose similarity measure," and "Plot spectrum of most similar" panels become visible. The median filter and pixel-count filter options are as described previously. The user may select between two similarity measures: Pearson correlation or the hypergeometric similarity measure. As discussed in Chapter 2, analyses on MSI and other types of high-dimensional data have indicated that these two similarity measures tend to yield relevant but complementary (i.e., including non-overlapping, unique selections) top ranked results [175, 176]. Once the "Select" button is pressed, similarity assessment will proceed using the selected measure, and the *m/z* values in the current dataset will be sorted in the main GUI according to their similarity to the template. After similarity assessment has completed, the user may plot a spectrum of most similar *m/z* values by choosing the number to plot in the "Plot spectrum of most similar" panel. If the imported template was based on a particular *m/z* value, that *m/z* value (the precursor peak, in precursor-product analysis) will be indicated in red on the spectrum, and the similar peaks from the current dataset will be plotted in black. Otherwise, all peaks will be plotted in black.

3.3. Case Studies

The datasets used for the case studies are related to prebiotically-relevant abiotic synthesis of nucleic acids such as DNA and RNA. While these datasets are not linked to HNSCC, the advantage is that they are less complex mixtures than eukaryotic cell

lysates. Thus, they provide test cases for DetectTLC that have comparatively less chemical noise and are easier to interpret, thereby facilitating the testing and validation of the tool.

Three datasets were investigated. The first and second both involved reaction products that are part of the synthesis of pyrazin-2-one (PZO). The two datasets were generated by using two different solvent systems (A and B). PZO-A was used in the first case study, Pipeline Comparison, and PZO-A and PZO-B were investigated in the second case study, TLC Spot Detection. The third dataset, which was utilized in the Parent-Fragment Ion Detection case study, used the reaction synthesis mixture for 2-aminopyrazine (APZ). All datasets were DESI-MSI. The full details of the experimental data acquisition process are described in [177].

3.3.1. Pipeline Comparison

Considering the two thresholding methods, two morphological operators, two median filter options (none and 7×7), two pixel-count filter options (none and excluding < 5, > 1500), and eight image features, a total of 128 alternative processing pipelines are possible. These were compared for the PZO MSI dataset. Each pipeline was assessed by the number of images in the top 40 rankings which contained true spot-like regions, as determined by manual inspection and verification of selected spots by collaborators in the Fernández lab. The full results of this comparison are shown in Table 3.2.

For this MSI dataset, the best results were returned by the analysis pipeline consisting of (1) Otsu's threshold, (2) morphological opening, (3) application of a 7×7 median filter, and (4) removal of images with < 5 and > 1500 non-zero pixels. For all of the image features except area, all 40 of the top 40 ranked images contained TLC spot-

Table 3.2: Performance comparison of all 128 alternative processing pipelines in identifying *m/z* images containing TLC spot-like regions among the top 40 rankings.

Evaluation	Binarization	Morphological Operation	None		7x7		Median Filtering
			None	5<x<1500	None	5<x<1500	Pixel Filtering
Area	Manual	Dilation	0	0	26	28	
		Opening	20	35	24	28	
	Otsu	Dilation	0	0	27	38	
		Opening	24	27	25	36	
Number of connected regions	Manual	Dilation	2	1	24	37	
		Opening	11	39	24	**40**	
	Otsu	Dilation	1	1	21	37	
		Opening	11	39	25	**40**	
Convex area	Manual	Dilation	2	2	24	28	
		Opening	15	37	38	**40**	
	Otsu	Dilation	2	2	24	37	
		Opening	15	35	36	**40**	
Eccentricity	Manual	Dilation	0	0	26	31	
		Opening	17	36	34	**40**	
	Otsu	Dilation	0	0	24	36	
		Opening	17	34	34	**40**	
Compactness	Manual	Dilation	0	0	27	30	
		Opening	23	37	39	**40**	
	Otsu	Dilation	0	0	26	38	
		Opening	20	31	39	**40**	
Solidity	Manual	Dilation	9	15	23	36	
		Opening	33	**40**	27	**40**	
	Otsu	Dilation	8	10	22	**40**	
		Opening	33	**40**	27	**40**	
Extent	Manual	Dilation	6	10	23	31	
		Opening	36	39	34	**40**	
	Otsu	Dilation	5	7	21	36	
		Opening	36	39	33	**40**	
Entropy	Manual	Dilation	1	3	24	**40**	
		Opening	18	**40**	24	**40**	
	Otsu	Dilation	0	1	24	39	
		Opening	18	39	24	**40**	

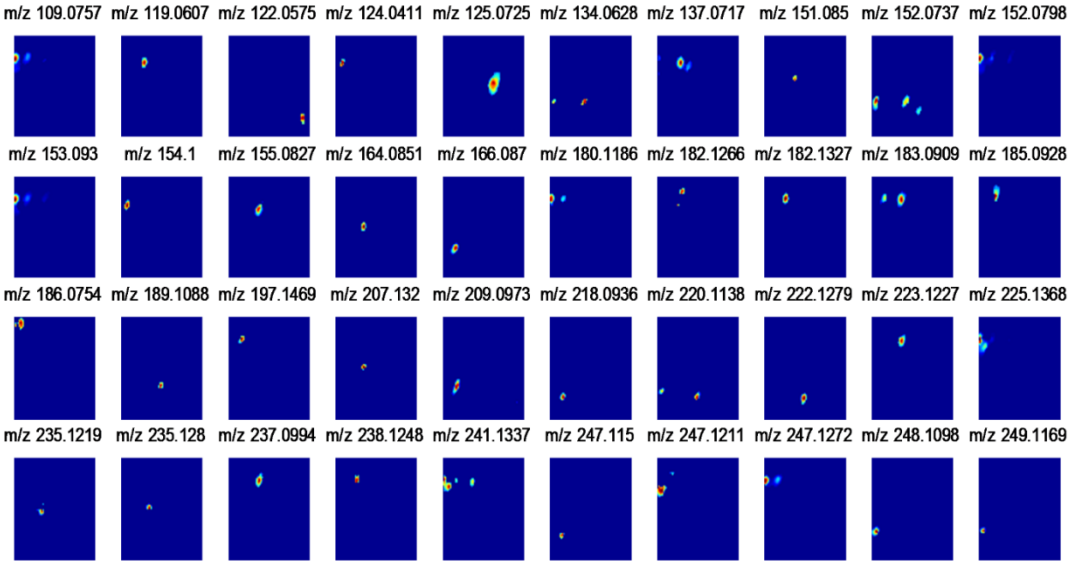


Figure 3.7: Example of *m/z* images featuring larger TLC spot-like regions. The top 40 selections are shown for the analysis pipeline consisting of (1) Otsu's method for generating binary images, (2) morphological opening, (3) 7×7 median filtering, (4) removal of images with < 5 and > 1500 non-zero pixels, and (5) entropy image feature.

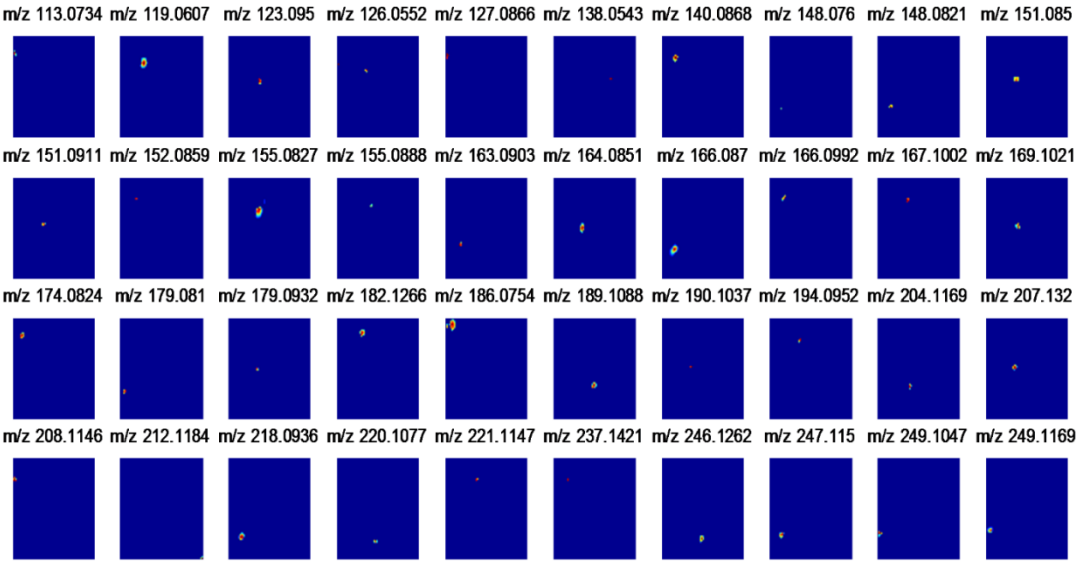


Figure 3.8: Example of *m/z* images featuring smaller TLC spot-like regions. The top 40 selections are shown for the analysis pipeline consisting of (1) Otsu's method for generating binary images, (2) morphological opening, (3) 7×7 median filtering, (4) removal of images with < 5 and > 1500 non-zero pixels, and (5) the compactness image feature.

like regions. The difference among the feature measures is most clearly demonstrated by the variety and size of the spots identified, as illustrated to an extent in Figures 3.7-3.8.

To further investigate these differences, the overlap among the top 40 rankings between each feature pair is shown in Table 3.3. This comparison confirms qualitative observations: image features which returned *m/z* images with smaller spot-like regions, such as compactness and convex area, had similar top 40 lists (e.g., 34/40 in common). Meanwhile, entropy, which returned *m/z* images with larger spot-like regions, and compactness had very different lists (e.g., 11/40 in common). Importantly, none of the image features were completely redundant in terms of their top 40 rankings. Image feature pairs which highlighted similar types (e.g. smaller or larger) of spot-like regions still returned unique *m/z* images. For example, entropy and extent both tended to

Table 3.3: Pairwise comparison of *m/z* images selected as the top 40 selections by different image features.

	Area	# Connected Regions	Convex Area	Eccentricity	Compactness	Solidity	Extent	Entropy
Area	40	16	26	17	26	7	18	0
# Connected Regions	-	40	24	23	23	21	26	21
Convex Area	-	-	40	28	34	14	27	11
Eccentricity	-	-	-	40	29	18	24	20
Compactness	-	-	-	-	40	15	25	12
Solidity	-	-	-	-	-	40	25	31
Extent	-	-	-	-	-	-	40	20
Entropy	-	-	-	-	-	-	-	40

highlight larger spot-like regions, but only 20/40 of their top-ranked images were in common. Thus, examining both of their top-ranked lists would be helpful during analysis, compared to considering only one image feature.

3.3.2. TLC Spot Detection

This case study demonstrates the application of DetectTLC in exploratory data mining through the PZO-A and PZO-B datasets.

DetectTLC Identifies Known Reaction Mixture Components

Three major products of the PZO synthesis process have previously been identified, and their chemical structures are known. These molecules are indicated in Figure 3.9(b). The TLC-MSI dataset was analyzed using DetectTLC. The top 20 results included images corresponding to the known products in terms of *m/z* and spatial location, as shown in Figure 3.9(c-d).

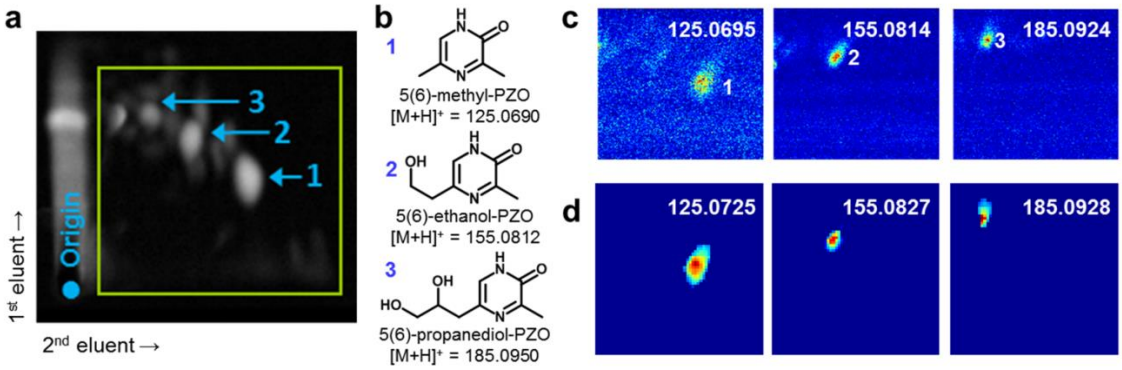


Figure 3.9: (a) Fluorescence image of a developed high-performance TLC (HPTLC) plate with the area imaged through DESI-MSI outlined in green (b) Known products of the PZO reaction, with numbers indicating the spatial location of each product in the fluorescence and MS images (c) Selected ion images acquired by DESI-MSI of reaction products previously identified, and (d) the corresponding images found by DetectTLC.

DetectTLC Can Help to Identify Relevant Components through Untargeted Analysis

When the PZO reaction mixture was analyzed using solvent system B, an intense fluorescent spot was observed that had not been seen using solvent system A, as shown in Figure 3.10(a). Processing the PZO-B dataset with DetectTLC yielded an ion image (*m/z* 167.0353) in which the spot was co-localized with that of the fluorescent image, as shown in Figure 3.10(c). This image was the 13th generated by DetectTLC, and appeared on the first page of results in the main GUI. Combining the DetectTLC-supplied *m/z* value with knowledge of reaction chemistry, my collaborators in the Fernández lab were able to tentatively identify this spot as 3,5(6)-dimethyl-4-oxoethyl-2-oxo-pyrazin-4-ium, a plausible reaction side-product of PZO synthesis.

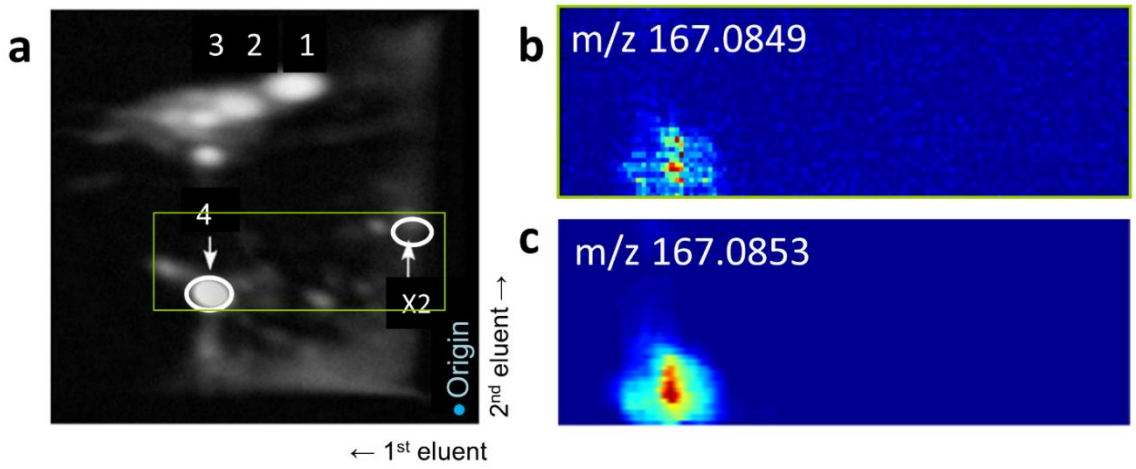


Figure 3.10: (a) Fluorescence image of a developed high-performance TLC (HPTLC) plate of the PZO reaction mixture, as separated by solvent system B. The green box indicates the area imaged using DESI-MSI, and Spot 4 indicates the unknown signal. (b) The manually plotted image of *m/z* 167.0353 and (c) the DetectTLC image identifying compound with *m/z* 167.0815 to be co-localized with Spot 4.

3.3.3. Parent-Fragment Ion Detection

The third and final case study applies the similarity assessment capabilities of DetectTLC to assist in structural identification of detected ions. In order to identify ions detected through mass spectrometry, it is necessary to obtain fragmentation data from high-energy ionization. In a typical tandem mass spectrometry (MS/MS) analysis, a precursor (parent) ion is selected for fragmentation. In MSI, selecting and fragmenting individual parent ions while maintaining high imaging throughput is challenging. A solution is to instead alternate between high- and low-energy scans while imaging, enabling both parent and fragment ions to be detected during a single experiment. This results in two datasets of equal or almost equal spatial dimensions, one consisting of parent ions, and another of fragment ions. DetectTLC utilizes spatial similarity measures to match potential fragments with parents.

This process is performed by allowing the user to select a parent ion of interest using the 'Advanced Similarity' GUI, and then identifying the most similar images in the fragment ion dataset. The intact parent ion of interest was 5(2-hydroxyethyl)-2-aminopyrazine (*m/z* 140.0817), a predicted side-product of the APZ reaction. The image corresponding to *m/z* 140.1 was first identified by DetectTLC from the low collision-energy dataset, as previously shown in Figure 3.6, and was used to create a template of the spot's location. Figures 3.11 and 3.12 show the results of the similarity assessment process, using Pearson correlation and the hypergeometric similarity measure, respectively. Both measures returned the same top 9 ranked ions, but sorted in different orders. Two of the top-ranked ions (*m/z* 122.0714 and 78.0345) were assigned to H_2O and CH_6N_2O losses from the parent ion, respectively. This observed fragmentation

pattern supports the structural assignment of 5(2-hydroxyethyl)-2-aminopyrazine to ions at *m/z* 140.0817.

In order to validate the fragment ions selected by DetectTLC, they were compared to liquid chromatography-tandem MS (LC-MS/MS) analysis performed on the parent ion.

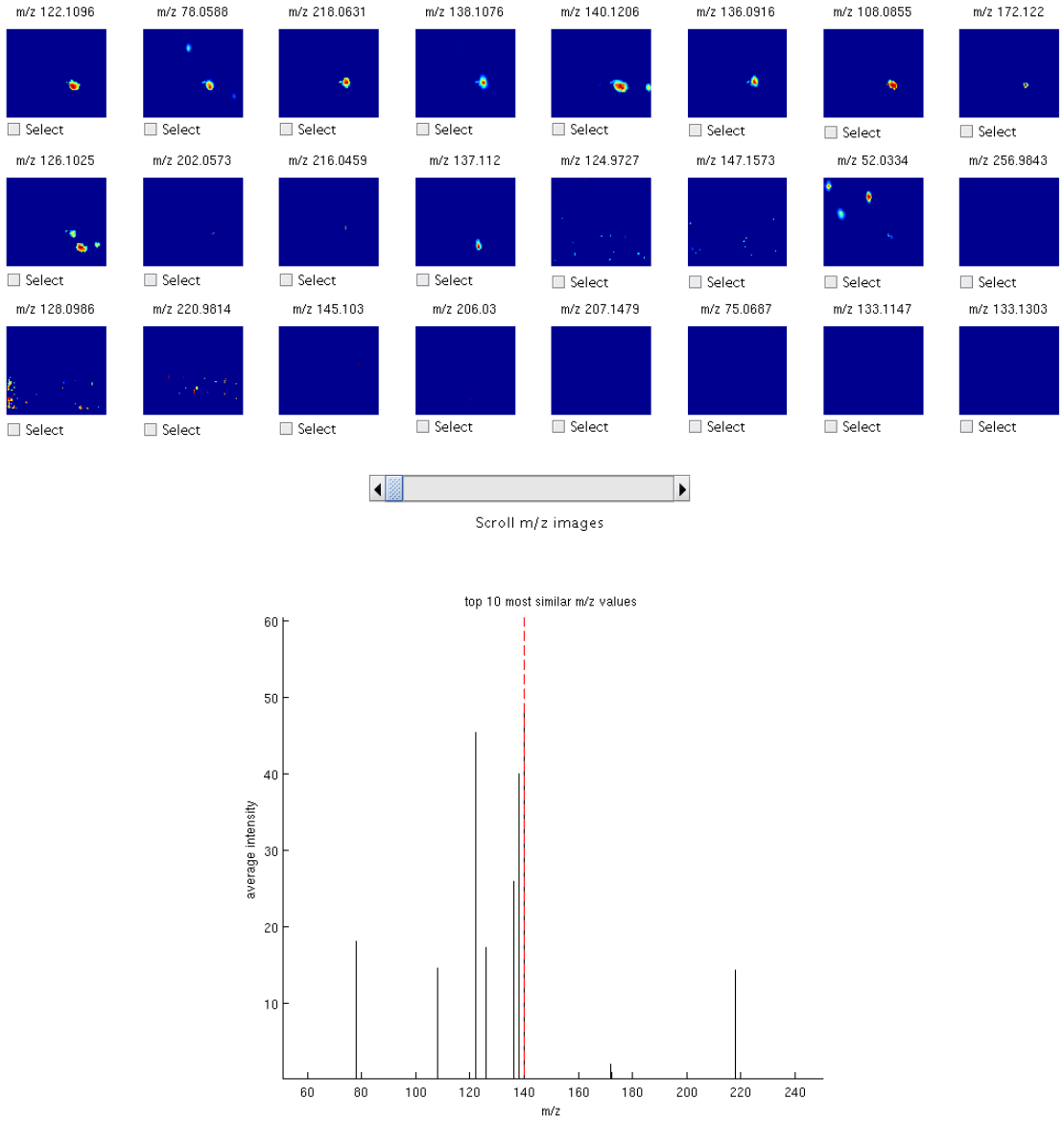


Figure 3.11: (*top)* The most similar fragment ion images observed when Pearson correlation was applied using the image of the precursor ion (*m/z* 140.1) as a reference. *(bottom*) The fragmentation mass spectrum showing the top 10 most similar *m/z* values (reference *m/z* indicated by dashed red line). DetectTLC automatically generates both of these figures during similarity analysis.

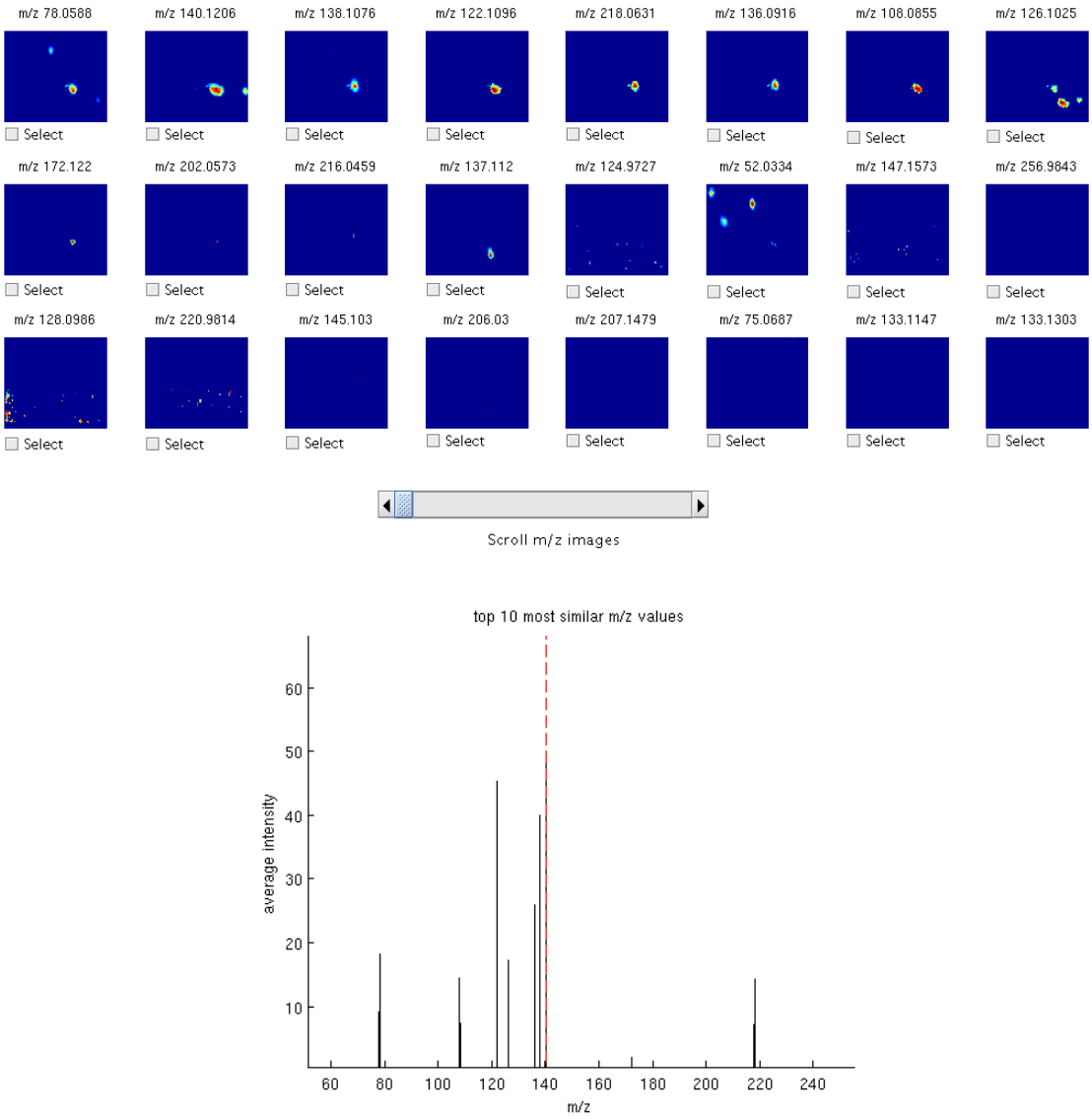


Figure 3.12: (*top)* The most similar fragment ion images observed when the hypergeometric similarity was applied using the image of the precursor ion (*m/z* 140.1) as a reference. *(bottom*) The fragmentation mass spectrum showing the top 10 most similar m/z values (reference *m/z* indicated by dashed red line). DetectTLC automatically generates both of these figures during similarity analysis.

These results are shown in Figure 3.13. Three of the ions, at *m/z* 78.0, 122.1, and 140.1, were identified in both datasets. The LC-MS/MS experiment also identified four other fragment ions which were not selected by DetectTLC: *m/z* 95.1, 105.0, 109.1, and 110.1. Investigation as to why these ions were missed by DetectTLC showed that three of

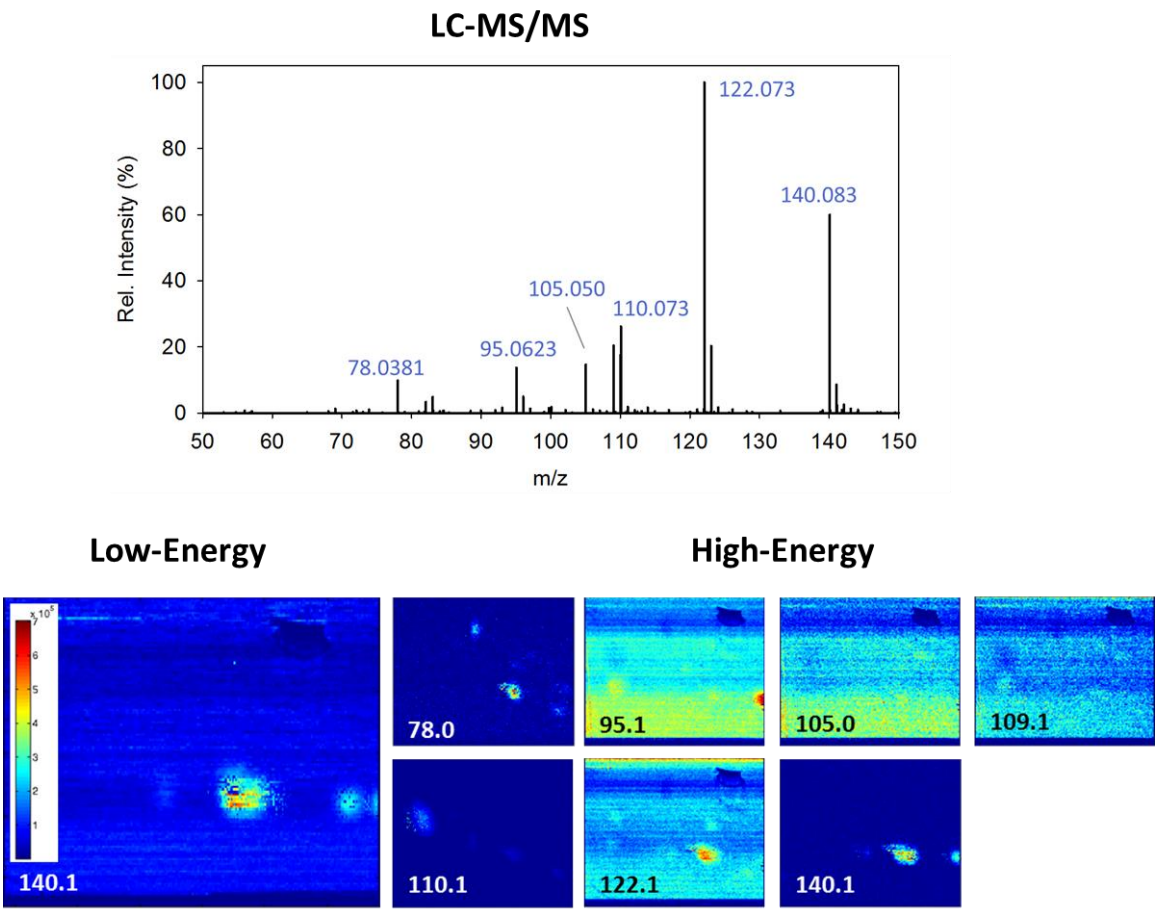


Figure 3.13: Comparison of LC-MS/MS analysis of selected APZ reaction products and the manually-selected ion images from low- and high-energy scans DESI-MSI. DetectTLC similarity matching was performed using *m/z* 140.1 as a template, and the high-energy ion images with red borders were in the top 8 outputs of DetectTLC. Figure courtesy of Fernández lab.

them (*m/z* 95.1, 105.0, and 109.1) had very low contrast between the TLC spot and the background. The deisotoping protocol was also found to be a factor in filtering out these relevant images; follow-up experiments (not shown) with smaller windows for deisotoping led to the selection of *m/z* 109.1 and 110.1 among the top 16 results.

3.4. Applications to HNSCC Research

DetectTLC is a general tool capable of processing TLC-MSI data from different biological contexts, including HNSCC. As previously noted, metabolomic research

through mass spectrometry in HNSCC is gaining momentum [35, 36]. In addition, TLC has been applied in HNSCC lipidomics research in several recent publications. For example, Gu and colleagues used TLC to separate and visualize lipids from five HNSCC cell lines in order to determine the mechanism by which RRR-α-tocopheryl succinate, a vitamin E analogue, induces apoptosis [178]. Yang and colleagues used TLC to track the effects of deguelin, which has been shown to have chemopreventive effects against other cancers, on HNSCC via the pro-apoptotic sphingolipid ceramide [179]. Thus, as research into metabolomics continues to grow, TLC-MSI analysis can help to uncover relevant metabolite- and lipid-centric patterns in HNSCC. DetectTLC can accelerate these experiments by removing the bottleneck of manual data processing.

3.5. Discussion and Key Innovations

In this chapter, I have described the design, development, and validation of DetectTLC, a software tool for accelerating metabolomics research through coupled TLC-MSI analysis. The previous state-of-the-art in assessing TLC-MSI datasets was manual inspection of the data to search for *m/z* images with TLC spot-like regions. DetectTLC automates this process, thereby removing a significant bottleneck to TLC-MSI experiments. While the TLC-MSI datasets tested during its development are related to prebiotic chemistry, DetectTLC is a general system that can be applied to metabolomics research in many different contexts, including HNSCC.

The utility of DetectTLC has been validated in the second and third case studies. First, it was demonstrated that DetectTLC is capable of automatically detecting spots corresponding to both expected and unexpected reaction mixture components. Second, it was demonstrated that DetectTLC can assist in structural identification of ions of interest

when multi-modal MSI is performed. During these validation experiments, some limitations of the current algorithms were also identified. Future versions of DetectTLC can incorporate image processing algorithms for the elimination of noisy background signals. They can also incorporate more sophisticated methods for deisotoping MSI data, in order to avoid inadvertent filtering of relevant *m/z* images in favor of neighboring high-intensity noisy images.

DetectTLC was developed and implemented in MATLAB. To enable widespread use of the tool, an executable (.exe) version of the tool has been generated. The tool will soon be freely deployed to the community via the website of Bio-MIBLAB at Georgia Institute of Technology.

The work described in this chapter was performed in collaboration with Dr. May D. Wang, Dr. Facundo M. Fernández and Dr. Rachel (Bennett) Stryffeler. Dr. Stryffeler performed MSI data acquisition and LC-MS/MS validation, while I implemented the DetectTLC tool and performed software-side experiments. The project idea and the design of DetectTLC tool capabilities and features were jointly developed.

The Key Innovations of this chapter are:

- Development of the first analytical pipelines using quantitative image features for identifying *m/z* images containing spot-like regions in MSI data
- Design, implementation, and validation of the first software tool, DetectTLC, for enabling and accelerating TLC-MSI studies in metabolomics by automatically finding mixture components of potential interest in TLC-MSI datasets

CHAPTER 4

SUPERVISED LEARNING MODELS FOR PATHOLOGICAL STAGE USING PROTEOMIC AND TRANSCRIPTOMIC DATA

4.1 HNSCC Disease Stage and Outcomes

The stage at which HNSCC is detected is important to therapeutic outcomes; patients with early stage (I and II) cancer have between 60-95% chance of successful local treatment, while those with advanced stage cancer are at high risk for recurrence or metastatic disease [37]. Greater knowledge of the molecular characteristics of different stages can provide insight into the mechanisms of HNSCC progression, and may help in identifying more effective therapeutic targets and strategies for treatment.

Previous research studies have analyzed gene expression, proteomic, and metabolomic data individually for studying differences between HNSCC stages, with mixed results. For example, three transcriptomic studies have related selected genes and gene signatures to different HNSCC stages [40, 45, 46], while two other transcriptomic studies did not find any discriminatory genes [41, 43]. A recent proteomic study using SELDI-TOF mass spectrometry data identified 11 *m/z* values differentially expressed between early- and late-stage oral SCC, but a satisfactory predictive model could not be developed [42]. Another recent study, using MALDI-TOF mass spectrometry data, identified several peaks that tended to correlate with clinical disease progression; however, no predictive model was developed [44]. A metabolomic study using ^{1}H NMR data identified several metabolite markers that discriminated between early and advanced stage HNSCC samples [34]. Additional bioinformatics studies – and in particular, the

development of predictive models that harness multiple data types – may help to gain additional insight into the differences between early and advanced HNSCC.

In this chapter, I investigated how quantitative functional proteomics, via reverse phase protein array (RPPA) data, can be used to develop predictive models for HNSCC stage. RPPA data is acquired by probing a sample with antibodies against specific proteins with regard to their activation states. With respect to HNSCC, RPPA data has been used to identify differentially expressed proteins between cancer and normal samples [180] and to identify proteins affected by the presence of an anti-invasion compound in nasopharyngeal carcinoma [181]. RPPA data has been applied to build predictive models for several other cancer types. Recent examples include for prognosis [182], drug response [183], and risk of recurrence [184] in breast cancer; for treatment response in ovarian cancer [185]; and for drug sensitivity in non-small-cell lung cancer [186].

In addition, I expanded upon previous efforts by developing predictive models for the same patient set using RNAseq data, and performing integrated analysis of RPPA and RNAseq data through functional assessment and ensemble model development. The goal of this investigation is to develop a set of improved predictive models, and thereby gather additional insight into HNSCC progression across multiple biological scales.

4.2 Protein and Gene Expression Datasets

RPPA data for HNSCC was downloaded from The Cancer Proteome Atlas (TCPA) [126] at http://bioinformatics.mdanderson.org/main/TCPA:Overview. This dataset consists of 212 patient samples and measures the response to 187 antibodies. TCPA provides a proteomic complement to The Cancer Genome Atlas (TCGA) [187] at

http://cancergenome.nih.gov/, where clinical, transcriptomic, and genomic data for the same patients are available. RNAseq data (Version 2) for HNSCC was downloaded from TCGA. Data was available for 210 of the same patients.

The downloaded RPPA data had been normalized and protein expression had been quantified using the "Supercurve Fitting" method. The details of these pre-processing steps are described in [126, 188]. In TCPA, antibodies are grouped into three classes: "validated", "under evaluation", and "use with caution." To perform a more conservative analysis, only those proteins with antibodies described as "validated" in both [126, 188] were utilized in this study. 113 proteins were considered for further analysis. In TCGA, RNAseq (Version 2) data has been aligned using MapSplice and quantified using RSEM [189, 190]. This dataset describes 20,531 genes. The un-normalized data was used for differential expression analysis and the normalized data was used for classification.

The clinical data for the 212 patients was downloaded from TCGA. Pathological stage information was used to divide the RPPA and RNAseq datasets into two groups: patients with early stage (stage I and II) cancer, and patients with advanced stage (stage III, IVA, IVB) cancer. Pathological state was unavailable for 12 patients, so clinical stage was substituted. One patient for whom the pathological stage was unavailable and the clinical stage was IVC was not considered, because unlike the other advanced cases, stage IVC involves metastatic disease. For RPPA, the early stage group contained 50 patients, and the advanced stage group contained 161 patients. The two patients for whom RNAseq data was unavailable were both of advanced pathological stage.

4.3. Model Development

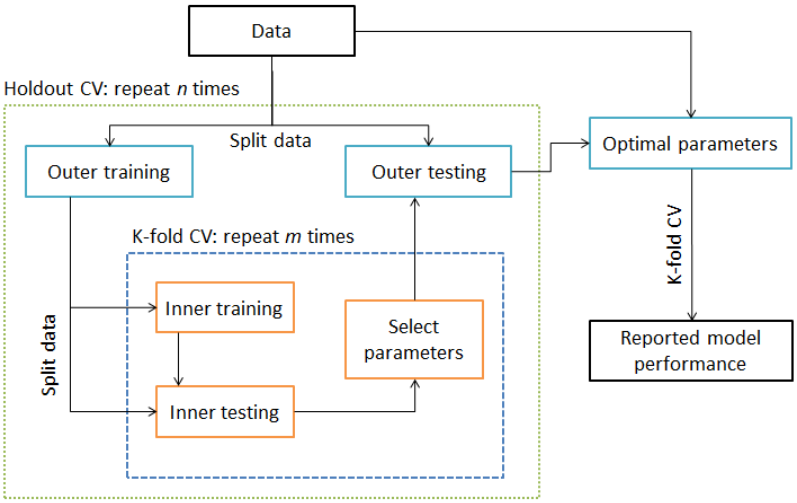


Figure 4.1: The nested cross-validation framework used in this study. The outer split was repeated $n = 3$ times, and the inner 10-fold cross-validation was repeated $m = 5$ times.

Classification Methods

Four individual binary classification methods and two ensemble classification methods were tested: k-nearest neighbors (KNN), support vector machine (SVM), naïve Bayes, decision tree, Adaboost, and bagging / Random Forests. Optimal parameters for each model were selected via nested cross-validation and grid search. Table I lists the range of parameters tested for each model, and Figure 4.1 describes the 3×5×10 nested cross-validation scheme. Optimization was performed with respect to the Matthews correlation coefficient (MCC). The area under the ROC curve (AUC) is also reported for the model having the maximum mean MCC. Analyses were performed using MATLAB (Mathworks, Natick MA).

Table 4.1: Classification model parameters examined via nested cross-validation

Classification Method	Parameters	Set of values
KNN	Number of neighbors (K)	$K \in [1,2,3,4,5,6,7,8,9,10]$
SVM	Kernel Soft margin cost (C) γ for GBRF	Kernels: linear, Gaussian radial basis function (GRBF) $C \in 2^m$, $m \in [-1,0,1]$ $\gamma \in 2^m$, $m \in [-1,0,1]$
Naïve Bayes	Prior distribution	Distributions: normal, kernel
Decision Tree	Splitting criterion	Criteria: Gini diversity index (GDI), Twoing rule, Maximum deviance reduction (MDR)
Adaboost	Number of trees (N)	$N \in [25,50,100]$
Bagging / Random Forests	Proportion (m) of all variables (p) to retain	$m \in [\sqrt{p}, \frac{p}{4}, \frac{p}{2}, p]$

Feature Selection

Three alternative feature selection methods were tested: two filter approaches and one wrapper approach.

The first filter method was based on differential expression. For RPPA data, the Wilcoxon rank-sum test was applied to identify proteins with significantly different expression between the early and advanced stage groups. Multiple testing corrections were applied by calculating the FDR for each protein, using the method of Benjamini and Hochberg through the R package p.adjust. To obtain a less conservative initial feature set, clinical stage was used to obtain a differentially expressed protein list. This yielded 11 proteins with FDR values ≤ 0.05, including the five proteins found when only pathological stage was used. A comprehensive examination of this feature space was performed by considering alternative classification models for every combination of the 11 features, i.e., $\sum_{i=1}^{11}\binom{11}{i} = 2047$ feature sets were considered. For RNAseq data, differential expression analysis was performed using two alternative tools, edgeR and EBSeq, of which the latter uses Bayesian methods [191, 192]. For a threshold of

FDR ≤ 0.05, edgeR identified 495 genes and EBSeq found 267 genes. These two lists had 108 genes in common. Due to the large number of differentially expressed genes identified by each method, comprehensive investigation of the feature space was not possible. Instead, model performances were compared across four feature sets: each differential expression result individually, the 108 common genes, and the 654 genes in the union of the selections of both methods.

The second filter method was mRMR (minimum redundancy maximum relevance), implemented using the FEAST toolbox [193-195]. The performance of each model was optimized for up to the top 50 features. The RNAseq data contained 1,414,819 unique count values, and the vast majority of values were observed only once. Due to this high dynamic range and memory limitations, the count values of the unscaled RNAseq data were binned prior to performing mRMR. The number of binned levels was chosen to balance performance and computational cost; 30,000 binned levels were the best alternative given the available computational resources.

In the wrapper approach, sequential forward feature selection (SFS) was performed. Model performance was optimized for up to the top 20 features. Due to the large number of genes in the RNAseq data, SFS was performed only after initial filtering based on differential expression. The input to SFS was the 654 genes found to be differentially expressed by edgeR and EBSeq in combination.

Data Scaling

Due to the high dynamic range of features in RNAseq data, two data scaling methods were tested. In the first – denoted scaled (1) – each feature was scaled by

dividing by the maximum value observed for that feature across any sample. In the second – denoted scaled (2) – each feature was scaled by subtracting its mean and dividing by its standard deviation, as suggested in [196]. The predictive modeling results for RNAseq with SFS are from unscaled data. For the differential expression and mRMR feature selection techniques, the best result among the two alternative scaling choices and unscaled data is shown.

Integrated Analysis

One of the fundamental goals of systems biology is to integrate information from multiple levels of biological complexity in order to increase actionable biological and clinical knowledge. However, this is a very challenging task. Several studies have demonstrated that mRNA and protein expression levels generally exhibit only moderate linear correlation; that is, mRNA expression may predict protein expression to a partial extent, but protein expression is also influenced by regulation at the post-transcriptional and post-translational levels [197-200]. Thus, models developed by integrating mRNA and protein features in some manner may potentially show improved performance over models using individual data types only. In a recent review, Haider and Pal discussed eight frameworks for performing integrated analysis of transcriptomic and proteomic data: union of data types, comparison of functional contexts, topological network analysis, merging datasets in individual domains, missing value estimation, multiple regression analysis, clustering, and dynamic modeling [201]. Due to the constraints of the available data, the techniques of merging datasets in individual domains, missing value estimation, multiple regression analysis, and dynamic modeling are not possible. In this

chapter, I examine model development based on the first two remaining methods: combination of the two data types and the results of functional assessment.

In the first case, RPPA and scaled RNAseq data were naively combined into a composite dataset. One dataset contained 221 features (113 RPPA and the 108 common RNAseq features) and the other contained 767 features (113 RPPA and the 654 union RNAseq features). SVM, KNN, and decision tree models with SFS were constructed using nested CV, with a maximum of 20 features. The better result among the two RNAseq scaling methods is reported.

In the second case, functional analysis of the genes corresponding to RNAseq and RPPA features in the best-performing models was performed using DAVID [202, 203] and the Reactome Analysis Tool [204]. I hypothesized that, if an ensemble of these models was created, individual models representing different functional categories would yield better-performing ensembles. This was tested by systematically evaluating all possible ensembles from nine SFS models: SVM, KNN, naïve Bayes, decision tree, and Adaboost using RPPA data, and SVM, KNN, decision tree, and Adaboost using RNAseq data. Ensemble decisions followed a majority voting scheme, and mean MCC values were compared across 100 repetitions of 10-fold CV.

4.4. Model Performance

4.4.1. Individual Data Types

Table 4.2 shows the predictive model performance of the six classifiers on the RPPA data. In general, performance is moderate, with several models achieving mean MCC values greater than 0.4 and AUC values greater than 0.7. The best performing

Table 4.2: Performance evaluation of alternative predictive models across feature selection methods for RPPA data

Classification Method	**Rank-Sum Test**		**mRMR**		**SFS**	
	MCC	**AUC**	**MCC**	**AUC**	**MCC**	**AUC**
SVM	**0.43±0.15**	**0.75±0.11**	**0.37±0.21**	**0.74±0.08**	**0.54±0.21**	**0.77±0.06**
Naïve Bayes	0.32±0.18	0.71±0.09	0.33±0.13	0.71±0.13	0.47±0.19	0.65±0.12
Decision Tree	0.16±0.17	0.64±0.13	0.12±0.27	0.62±0.11	0.40±0.20	0.68±0.11
KNN	0.35±0.28	0.74±0.13	0.28±0.13	0.77±0.09	0.46±0.22	0.73±0.11
Adaboost	0.11±0.24	0.71±0.11	0.25±0.34	0.71±0.14	0.45±0.13	0.71±0.11
Random Forests	**MCC**		**AUC**			
	0.04±0.16		0.65±0.10			

RPPA model was SVM with SFS feature selection. The SVM models outperformed the other classifiers for all feature selection methods on the RPPA dataset, and the SFS models outperformed the other feature selection methods for all classifiers. The naïve Bayes and KNN models were the next best in performance, while the decision tree models did not perform as well. The two ensemble classifiers showed markedly different performance. For mRMR and SFS, Adaboost outperformed the decision tree models, although it did not perform as well as the other individual classifiers. The Random Forests classifier gave surprisingly poor performance for the RPPA data, with an MCC value close to zero.

Table 4.3 shows the predictive model performance of five classifiers on the RNAseq data. The best RNAseq models, which achieve mean MCC values greater than 0.6, outperform the best RPPA models. Again, the SFS models outperformed the other feature selection methods for all classifiers. The highest performing RNAseq model was KNN with SFS; the Adaboost and SVM models with SFS performed almost as well in terms of MCC, though the SVM AUC value was non-informative. The Random Forests model for RNAseq data showed better mean performance than that for RPPA data, but it also had a large standard deviation. For differential expression and mRMR feature

Table 4.3: Performance evaluation of alternative predictive models across feature selection methods for RNAseq data. Legend: unscaled, *scaled (1),* scaled (2).

Classification Method	**Differential Expression**		**mRMR**		**DEG+SFS**	
	MCC	**AUC**	**MCC**	**AUC**	**MCC**	**AUC**
SVM	**0.52±0.27**	**0.91±0.11**	**0.42±0.27**	**0.50±0**	0.62±0.22	0.50±0
Decision Tree	0.23±0.22	0.62±0.11	*0.36±0.26*	*0.69±0.12*	0.52±0.14	0.74±0.11
KNN	0.35±0.22	0.67±0.08	0.26±0.27	0.68±0.09	**0.64±0.20**	**0.83±0.10**
Adaboost	*0.27±0.28*	*0.66±0.12*	0.32±0.23	0.73±0.10	0.62±0.13	0.73±0.14
Random Forests	**MCC**		**AUC**			
	0.30±0.30		0.79±0.10			

selection, the SVM models outperformed the other classifiers in terms of MCC. Under these two feature selection methods, the decision tree and Adaboost models showed better performance for RNAseq data than for RPPA data, but KNN was not notably different. In the majority of cases, scaled data showed better performance, and the second scaling method was more often better than the first.

4.4.2. Commonly Selected Features and Functional Analysis

The existence of well-performing models implies that the selected features are of functional importance. The five RPPA SFS models were compared, and 11 features were selected in at least two models. All of these have been associated with HNSCC in the literature: AR [205], C-Raf [206], CDK1 [207], Cyclin B1 [208], MAPK_pT202_Y204 [1], N-Cadherin [209], PDK1 [210], PI3K-p85 [211], VEGFR2 [212], c-Jun_pS73 [213], and p27_pT198 [214]. In particular, AR was selected by four models, CDK1 and Cyclin B1 by three, and the others by two. Table 4.4 shows the number of total common features between each model pair. The low counts show that some models achieved comparable performance using very different feature sets. Even greater feature diversity was observed for the RNAseq SFS models. Among the four models, 52 features were present in total, but only two features were selected in more than one model: FAM27B and KRTAP17-1.

Table 4.4: Comparison and functional analysis of the RPPA SFS models: The number of features, GO functional annotations, and pathways (KEGG and Reactome) in common between different models are indicated.

	SVM	**Naïve Bayes**	**Decision Tree**	**KNN**	**Adaboost**
SVM Features: 18 GO terms: 10 KEGG: 11 Reactome: 209	-	Features: 2 GO terms: 2 KEGG: 0 Reactome:93	Features: 2 GO terms: 1 KEGG: 9 Reactome:139	Features: 4 GO terms: 0 KEGG: 7 Reactome:86	Features: 2 GO terms: 0 KEGG: 0 Reactome:67
Naïve Bayes Features: 14 GO terms: 5 KEGG: 0 Reactome: 112	-	-	Features: 3 GO terms: 4 KEGG: - Reactome:97	Features: 1 GO terms: 0 KEGG: - Reactome:82	Features: 1 GO terms: 0 KEGG: - Reactome:59
Decision Tree Features: 11 GO terms: 15 KEGG: 25 Reactome: 208	-	-	-	Features: 2 GO terms: 1 KEGG: 12 Reactome:103	Features: 1 GO terms: 0 KEGG: - Reactome:71
KNN Features: 15 GO terms: 1 KEGG: 12 Reactome: 126	-	-	-	-	Features: 2 GO terms: 0 KEGG: - Reactome:56
Adaboost Features: 8 GO terms: 0 KEGG: 0 Reactome: 131	-	-	-	-	-

Functional analysis of the SFS feature sets was performed via DAVID and Reactome. DAVID was used to find significantly enriched Gene Ontology (GO) terms and KEGG pathways, while Reactome also returned significant pathways. In terms of specific features and GO terms, the five RPPA SFS models were diverse, with relatively few commonalities. However, many common pathways were found, both through KEGG and through Reactome. Seven KEGG pathways were common among the SVM, decision tree, and KNN RPPA models. These consisted of three signaling pathways: ErbB,

neurotrophin, and insulin signaling, and four cancer-related pathways: pathways in cancer, colorectal cancer, pancreatic cancer, and chronic myeloid leukemia. Reactome returned many more significant pathways than DAVID, and 46 pathways were in common among all five models. Most of these related to signal transduction and mitotic progression.

Notably, there were no results in DAVID for the four RNAseq feature lists from the SFS models. Reactome returned results for only the KNN RNAseq model. The nine pathways identified fell into four categories: regulation of gene expression and development in beta cells, visual transduction and phototransduction, retinoid metabolism and transport, and the synthesis of bile acids and bile salts. Retinoids are important therapeutics for many cancer types, including HNSCC [215], and recent studies have shown that bile acids may be associated with head and neck cancer [216, 217].

4.4.3. Integrated Analysis

The results for developing SFS models based on naïve combination of the RPPA and RNAseq datasets are shown in Table 4.5. All of the models outperformed the corresponding RPPA SFS models for the same classification method in terms of mean MCC values. However, only the SVM models showed improvement over the RNAseq SFS models as well. Moreover, only the models for the smaller composite dataset (221 features) utilized both RPPA and RNAseq features. The RPPA features selected by the SVM model for the smaller composite dataset were Cyclin B1 and p38_pT180_Y182. Cyclin B1 was one of the commonly selected features among the RPPA SFS models; p38 is a mitogen-activated protein kinase that has also been associated with HNSCC [218]. The models for the larger composite dataset (767 features) selected only

Table 4.5: Performance evaluation of alternative predictive models using two composite RPPA and RNAseq datasets. Legend: *scaled (1)*, scaled (2).

Classification Method		**SFS (221 features)**		**SFS (767 features)**	
		MCC	**AUC**	**MCC**	**AUC**
SVM		*0.68±0.15*	*0.82±0.09*	0.70±0.21	0.82±0.14
Decision Tree		0.50±0.20	0.75±0.13	0.46±0.16	0.69±0.10
KNN		0.53±0.21	0.77±0.11	0.64±0.26	0.83±0.13

RNAseq features. Thus, the improvement in MCC seen for the best-performing model (SVM with the larger composite dataset) cannot be attributed to integrating data types, but may be due in part to using scaled data.

Figure 4.2 compares the performance of single-data type models (RPPA and RNAseq) with ensembles comprised of only RPPA models, only RNAseq models, or both. The last category contains all possible ensembles with three to nine member models. Results represent the mean performance of 10-fold CV for the 209 common patients in the RPPA and RNAseq datasets, across 100 repetitions. The best single-data type ensembles had higher mean MCC values than individual models of that data type. Additionally, combination ensembles of multiple sizes were found which had better performance than any of the single-data type ensembles.

The performances of the RPPA-only ensembles were compared in terms of the previous functional analysis results. For example, the best performing RPPA-only ensemble (SVM, KNN, Adaboost) achieved a mean MCC value of 0.54, and 50 Reactome pathways were in common among the three feature sets. The worst-performing ensemble (SVM, Naïve Bayes, decision tree) had a mean MCC of 0.25 and 86 Reactome pathways in common. Among the RPPA-only ensembles overall, a correlation of -0.44 was observed between the mean MCC values and the number of Reactome pathways in

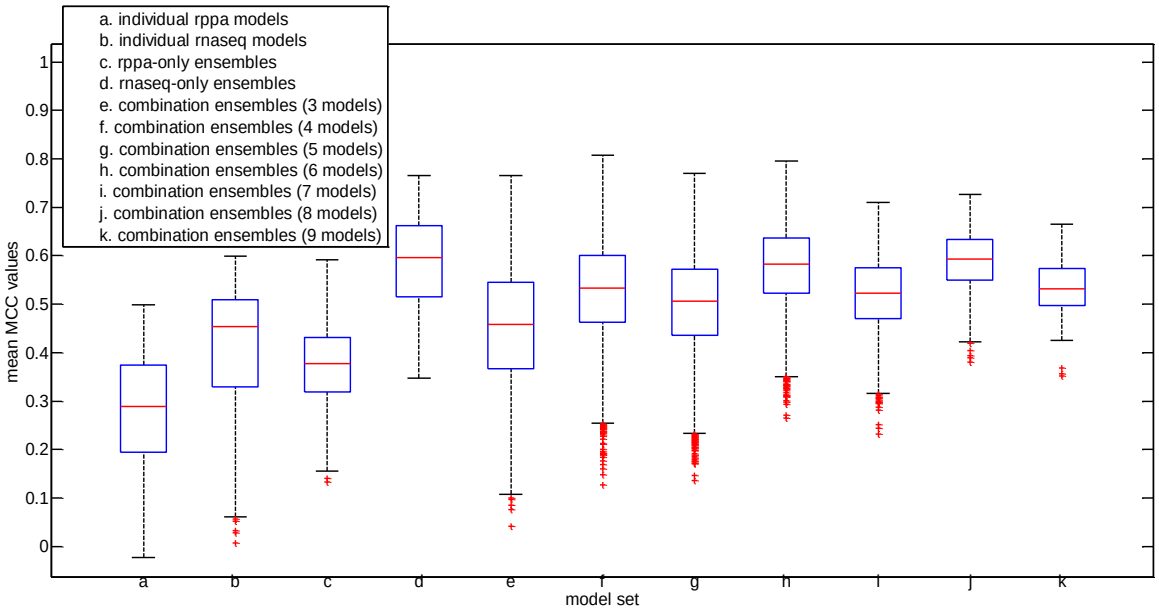


Figure 4.2: Comparison of individual and ensemble model performances over 100 repetitions of 10-fold CV. Combination ensembles, which allow for heterogeneity in both data type and component model type, outperform RPPA-only and RNAseq-only models.

common among the ensemble member models.

While the RNAseq ensembles had the highest median performance, several combined RPPA and RNAseq ensembles had higher overall performance. Among all of the ensembles tested – RPPA only, RNAseq only, and combination – 27 ensembles were identified which had better performance than the best-performing individual RNAseq model in more than 90 of the 100 CV repetitions. Of these, two were RNAseq-only ensembles. Another two were combination ensembles containing three and four models, respectively, in which only RNAseq models were chosen as members. The other 23 notable ensembles all contained both RPPA and RNAseq member models.

Among these 23 was the best performing ensemble overall, which achieved a mean MCC value of 0.80. This is higher than any of the model performances reported for previous tests. Steiger's Z test was used to compare the MCC performance of this

ensemble model with those of the highest performing RNAseq (KNN) and composite (SVM) models [219]. In both cases, the improvement was statistically significant ($p < 0.01$). This particular ensemble incorporated the Adaboost RPPA SFS model and the SVM, KNN, and Adaboost RNAseq SFS models.

4.5. Discussion and Key Innovations

In this study, I have performed an in-depth analysis of HNSCC RPPA data by implementing six different classification methods, using nested cross-validation to optimize parameters, and testing three alternative feature selection methods. This supervised approach contrasts with previous HNSCC studies using RPPA data, which have conducted unsupervised and differential expression analyses [180, 181]. It also differs from previous supervised studies on RPPA data [182, 183] in two ways. First, this study assesses the performances of several different combinations of feature selection methods and classification algorithms in order to identify the potentially relevant protein feature sets. Second, this study builds upon current research by developing integrated proteomic and transcriptomic models, and comparing them to RPPA-only and RNAseq-only models. In particular, I performed two types of integrated analysis: one by direct combination of RPPA and RNAseq data, and another by constructing ensemble models using both data types. To my knowledge, this is the first such comparative, integrated study for modeling progression in HNSCC.

From a modeling perspective, this study identified the integrated ensemble approach with both RPPA and RNAseq models as the best overall. The top-performing model for predicting HNSCC pathological stage was obtained using this approach, and had a statistically significant higher MCC value than the best performing individual

RNAseq and composite models. Notably, modeling results appear to support the initial conjecture that less functional agreement among the feature sets of member models will be associated with better performance. First, the RNAseq-only and the combination RPPA and RNAseq ensembles were observed to outperform the RPPA-only ensembles. Second, a moderate negative correlation was observed between the performances of RPPA-only ensembles and the numbers of common Reactome pathways among ensemble members. These observations indicate that higher-performing ensembles tended to be more functionally diverse in terms of member model feature sets. Investigation on larger datasets, as well as assessment using ensemble diversity measures and different ensemble construction techniques [220], are directions for further research. More rigorous examination of how and why different classifier and feature selection method combinations tend to vary in performance on RPPA and RNAseq data is also an important task.

A related question of interest is performing multi-class classification to study bio-molecular expression patterns among individual HNSCC stages, rather than grouping them into early and advanced disease. Another is investigating the differences between normal and early stage HNSCC samples. For investigating these questions, the availability of sufficiently large – in terms of both patients and features – public datasets is a constraint. While matched tumor and normal RNAseq data is available on TCGA for HNSCC, RPPA data for matched normal samples is yet unavailable. In addition, an inherent limitation of RPPA data is that only a selected set of proteins is measured. A larger set of proteins could enable discovery, in that proteins which were previously not implicated in HNSCC – or cancer in general – might be identified as informative features

through modeling. TCPA is currently in the process of extending their antibody set to cover 500 proteins [126], which will help to address this limitation to some extent. The availability of more extensive proteomic data for HNSCC through mass spectrometry is a related promising avenue. The Clinical Proteomic Tumor Analysis Consortium (CPTAC), like TCPA, is currently building a proteomic complement to TCGA. CPTAC hosts a library of LC-MS/MS data from tumor samples that are also in TCGA. At the time of writing, data from breast cancer, ovarian cancer, colon adenocarcinoma, and rectum adenocarcinoma have been released. Future availability of such data for HNSCC would be valuable to researchers.

From a systems biology perspective, investigating multiple types of –omic datasets to gain insight into disease processes is an important area of research. Numerous individual proteins and genes selected as features in well-performing models in this study have been previously associated with HNSCC in the literature, including in a recent large-scale study by The Cancer Genome Atlas Network [221]. Additionally, functional analysis of the features selected in the top-performing models revealed notable patterns. Many processes – e.g., signal transduction pathways including those through EGFR and ERBB2, and events related to mitotic progression – were commonly represented among the RPPA model features. The RNAseq feature sets were much more diverse, but some of the associated biological processes have still been linked with HNSCC in the literature.

While this integrative modeling study of RPPA and RNAseq data can provide guidance for further research, integration in general should be interpreted with caution. Because RPPA is a tool for functional proteomics, it is several biological steps removed from the mRNA counts measured by RNAseq, and mRNA is itself distinct from genome-

level factors. Thus, further investigation into additional data types – e.g., copy number variations, mutations, DNA methylation, protein subunits and alternative activation states, metabolites – is needed for drawing conclusions about the specific mechanisms underlying HNSCC progression. Appropriate comparison and combination of multiple data types will help to fill in the gaps and provide greater insight into the process of disease development. By harnessing the diverse data from initiatives like TCPA, TCGA, and CPTAC, bioinformatics studies can lead to better understanding of the molecular bases of HNSCC and also other cancers.

The Key Innovations of this chapter are:

- Performed the first supervised modeling study for modeling progression in HNSCC by integrating both proteomic and transcriptomic data
- Developed between-omic level integrated ensemble models with significant improvement in performance for predicting HNSCC pathological stage

CHAPTER 5

SUPERVISED LEARNING MODELS FOR EARLY DETECTION USING TRANSCRIPTOMIC DATA MODELS

5.1. Transcriptomic Modeling Research in HNSCC

Investigation of gene expression patterns in HNSCC is an active area of research, with numerous studies conducted using gene expression microarrays within the last 10 years [222]. More recently, transcriptomic research has shifted towards RNA sequencing (RNAseq) because of its high sensitivity and dynamic range [223]. However, due to variations in the sample population, small samples sizes, and differences in experimental design and analysis methods, different transcriptomic studies on the same disease may report notably different lists of significant or key genes [224]. For this reason, integrated analysis of multiple transcriptomic studies is necessary for identifying consistent, fundamental gene expression patterns that indicate HNSCC status.

Previous predictive modeling studies have applied gene expression data to various problems related to HNSCC, including predicting metastatic disease [225, 226], the development of cancer in patients with oral premalignant lesions [38, 39], and the risk of recurrence and relapse [227, 228]. A key aspect of HNSCC research is early detection: if the cancer is detected at an early stage, patient response to treatment is relatively high, and five year survival rates for multiple disease subsites exceed 80% [14, 37]. However, most cases are detected only at locally advanced stages, which are associated with much worse outcomes. For the same disease subsites, survival for locally advanced cases ranged from 49.8-73%. Current screening recommendations for oral cancer are based on

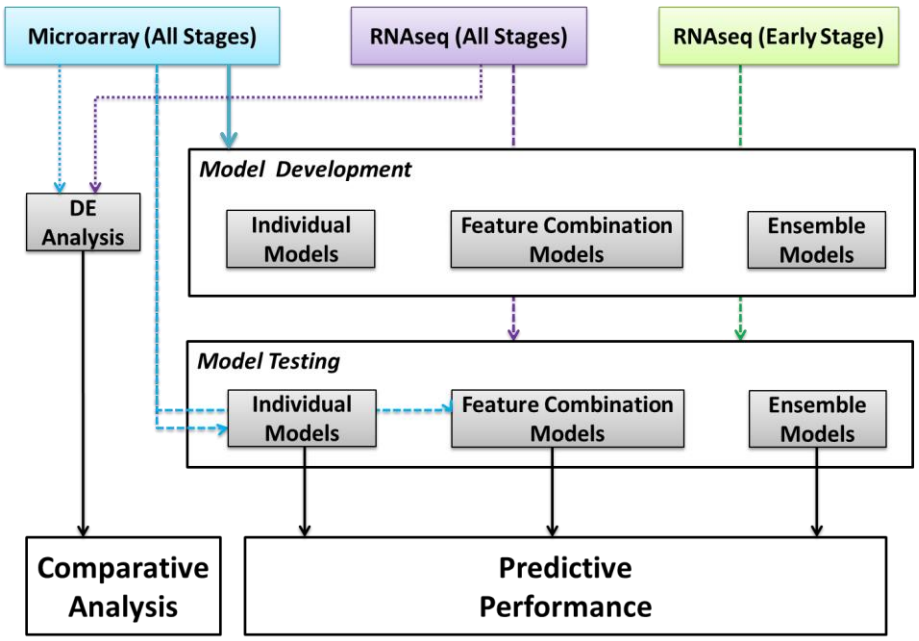


Figure 5.1: Modeling workflow describing roles for microarray, RNAseq (all stages) and early-stage RNAseq datasets

conventional visual and tactile examinations [229]. Effective supervised models for predicting HNSCC status – and in particular, for differentiating early-stage HNSCC patients from healthy individuals – based on molecular data could be useful clinical tools.

In this chapter, I present an integrated transcriptomic analysis of HNSCC, with the goal of developing robust predictive models for determining disease status. Because the lack of early stage samples is an obstacle, models are initially developed for predicting HNSCC status in general, and are then applied to predict early stage HNSCC in particular. The workflow for this study is shown in Figure 5.1. First, differential expression (DE) analysis was performed on several microarray datasets to identify common DE genes and to investigate the extent of variation among datasets. Second, classification models optimized on one microarray dataset were implemented on the others, in order to evaluate within-platform model robustness. Third, individual and

ensemble classification models developed using the microarray datasets were applied to RNAseq data, to test (i) between-platform robustness, i.e., if informative gene feature sets and model structures are transferrable across data types and (ii) performance in detecting early stage HNSCC. Finally, well-performing models were integrated into a software tool with a graphical user interface in order to make predictive models more accessible to HNSCC researchers and clinicians.

5.2. Microarray and RNAseq Datasets

Gene expression microarray datasets were obtained from the Gene Expression Omnibus (GEO) and ArrayExpress public repositories. To increase consistency in the downstream analysis, datasets selected for study met the following criteria: (i) data was from patient samples, not cell lines; (ii) data from both diseased and normal samples were available; (iii) the raw, unprocessed data was available; (iv) an associated publication was available; and (v) Affymetrix array platforms were used. These filtering steps led to five candidate Affymetrix datasets. Of these, three were chosen for further analysis because they shared the Human Genome U133 Plus 2.0 (54,675 probes) or U133A (27,777 probes) arrays, enabling direct comparison of probes. Only the 22,277 common probes were used for analysis. These datasets are described in Table 5.1. To obtain gene expression values, the raw .CEL files were processed with RMA in the Affymetrix Expression Console software.

Table 5.1. Description of gene expression microarray datasets examined in study

Dataset	Affymetrix Array Platform	Samples	Reference
E-GEOD-9844	Human Genome U133 Plus 2.0	25 cancer, 12 normal	[125]
E-GEOD-6791	Human Genome U133 Plus 2.0	42 cancer, 11 normal	[230]
E-GEOD-23036	Human Genome U133A 2.0	63 cancer, 5 normal	[231]

RNAseq data (Version 2) for HNSCC was obtained from The Cancer Genome Atlas (TCGA), along with associated clinical data. The data has been aligned using MapSplice and quantified using RSEM [189, 190]. Count data for 20,531 genes are available in this dataset. Un-normalized data was used for DE analysis, and normalized data for classification. At the time of analysis, matched tumor and normal RNAseq data was available for 40 patients. Of these, 17 patients were categorized as early stage (pathological stages I and II).

5.3. Model Development

Differential Expression Analysis and Feature Selection

Both the two-sample t-test and Wilcoxon rank-sum test were used to identify DE genes between the HNSCC and normal samples in the microarray datasets. Multiple

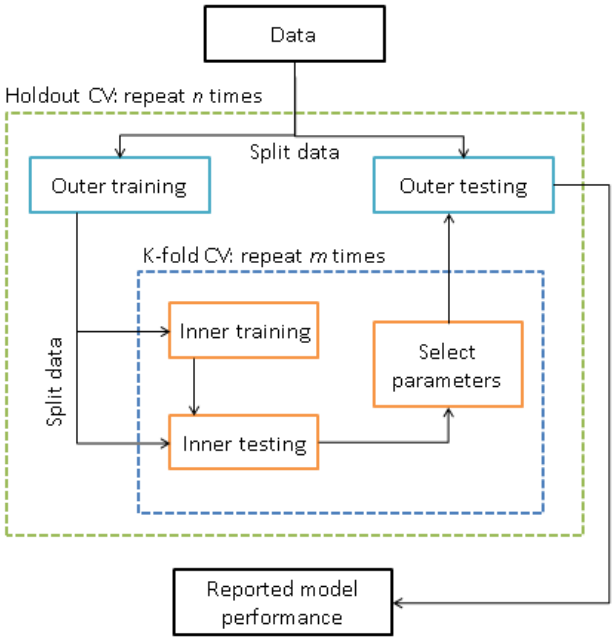


Figure 5.2: Schematic of nested cross-validation (nested CV) framework

testing correction was implemented by controlling the False Discovery Rate (FDR ≤ 0.05) or by implementing Bonferroni correction ($\alpha_{Bonferroni}$ = 0.05). DE analysis was performed on the RNAseq data using edgeR (FDR ≤ 0.05) [192]. To evaluate consistency in gene expression patterns across datasets, the DE gene lists were compared to each other. For classification, features were selected via mRMR (minimum redundancy maximum relevance) from the microarray datasets, implemented using the FEAST toolbox in MATLAB [193-195]. The performance of each classification model was optimized for up to the top 50 features.

Binary Classifiers

Three binary classification methods were tested: k-nearest neighbors (KNN), support vector machine (SVM), and decision tree (DT). Optimal parameters for each model were selected via grid search from the ranges shown in Table 5.2, using a nested cross-validation scheme as shown in Figure 5.2. Optimization was performed with respect to the Matthews correlation coefficient (MCC), and the area under the ROC curve (AUC) is also reported for optimized models. All analyses were performed using MATLAB (Mathworks, Natick MA).

Table 5.2. Classification model parameters examined via nested cross-validation

Classification Method	Parameters	Set of values
KNN	Number of neighbors (K)	$K \in$ [1,2,3,4,5,6,7,8,9,10]
SVM	Kernel Soft margin cost (C) γ for GBRF	Kernels: linear, Gaussian radial basis function (GRBF) C $\in 2^m$, m $\in$ [-2,-1,0,1,2] $\gamma \in 2^m$, m $\in$ [-2,-1,0,1,2]
Decision Tree (DT)	Splitting criterion	Criteria: Gini diversity index (GDI), Twoing rule, Maximum deviance reduction (MDR)

5.3.1. Evaluation of Model Robustness across Microarray Datasets

The robustness of each model was first evaluated by testing the model on the other microarray datasets, i.e., those which were not used in its development. In order to avoid the issue of batch effects entirely, models were not applied to other datasets directly. Instead, the model parameters and feature set associated with the top performing model of each classifier type for dataset i were used to train a model on dataset j ($\forall j \neq i$), with $i, j \in [1,2,\dots n]$, where n is the number of microarray datasets. In addition, this comparison was also carried out after combining the feature sets for alternative models. For example, such a model to be tested on E-GEOD-6791 would combine the optimized feature sets for both E-GEOD-9844 and E-GEOD-23036. The rationale behind this experiment is to test whether a composite feature library, defined as the union of optimal feature sets from multiple datasets, would help to improve average predictive performance on new incoming datasets.

Application of Microarray-Developed Models to RNAseq Data

Next, model robustness across data formats was investigated by applying microarray-developed models to RNAseq data from TCGA. As described for the microarray-only cross-performance tests, the model parameters and feature set associated with a given microarray model was used to train a model on the RNAseq data. In order to transfer the optimized microarray feature sets, microarray probes were mapped to RNAseq features on the level of gene symbols.

In addition to testing the performance of the nine individual and nine possible feature combination-based models on RNAseq data, two ensemble modeling frameworks

– majority voting and stacking [220] – were tested. Majority voting is the simplest ensemble framework; given a set of predicted labels from alternative models, the ensemble-predicted label is the most-commonly predicted label. All possible combinations of the nine individual microarray models (three classifier types for three datasets) with at least three members were considered for voting-based ensembles, resulting in 511 alternative models. Stacking involves a two-step classification process. In the first step, label predictions are obtained from alternative models on the training data. This set of predicted labels serves as the features for a second classification model, which is used to generate the final predictions. Three stacking models were developed, using all nine individual microarray models as the first-level predictors, and SVM, KNN, and DT were tested as the three second-level classifiers.

Tool Design

The developed predictive models were integrated into a software tool with a graphical user interface (GUI) to make them more accessible to HNSCC researchers, and for easy application to new datasets. Users can apply previously developed individual or ensemble models to process incoming datasets, and then visualize and export the results.

5.4. Model Performance

Comparison of DE Gene Lists and mRMR Selections

Notable differences were observed among the DE genes selected in each microarray dataset, as shown in Table 5.3. The more conservative Bonferroni method resulted in only 5 common DE genes among the three datasets: *MMP1, ABCA8, MYO1B, ARHGEF10L,* and *SASH1*; all of these have been associated with HNSCC in recent

Table 5.3. Comparison of DE Genes across Microarray Datasets

	E-GEOD-6791		E-GEOD-9844		E-GEOD-23036		Common DE Genes	
	Bonferroni	FDR	Bonferroni	FDR	Bonferroni	FDR	Bonferroni	FDR
T-test	213	5763	84	2451	273	3154	5	682
Rank-sum test	14	6836	5	2759	0	3401	0	785

literature [125, 231-234]. When applying FDR, more than 600 common DE genes were identified across the three datasets for both statistical tests. Functional analysis was performed on the common FDR gene lists using DAVID [202]. Although not statistically significant, the top 10 Gene Ontology (GO) terms selected for both sets included GO:0006915~apoptosis, GO:0008219~cell death, GO:0012501~programmed cell death, GO:0016265~death, GO:0043588~skin development. Overall, these results indicate that while there is substantial variation across the microarray datasets, the commonly-selected DE genes are relevant to HNSCC.

This variation was also observed for the RNAseq data: 10,239 DE genes were identified in the RNAseq data through edgeR; 526 and 610 of these genes overlapped with the microarray common DE gene lists (FDR) for the t-test and rank-sum test, respectively. Some of the commonly selected DE features were also represented in the mRMR feature lists. The common DE genes selected using either test with FDR were compared with the top 50 mRMR-selected features for the three microarray datasets. The number of features in the intersection of these lists ranged from 17 to 23.

5.4.1. Model Performance across Microarray Datasets

Table 5.4 shows the performance of KNN models developed using one microarray dataset on the others. The top three rows show the performance of the KNN model optimized through nested CV for each dataset. For example, the best-performing

Table 5.4. Multi-Dataset Performance of KNN Models in terms of MCC (AUC)

Data for Model Development			Prediction Dataset		
E-GEOD-6791	**E-GEOD-9844**	**E-GEOD-23036**	**E-GEOD-6791**	**E-GEOD-9844**	**E-GEOD-23036**
•			1±0 (1±0)	0.89±0.10 (0.94±0.06)	-0.08±0.03 (0.54±0.01)
	•		0.71±0.14 (0.98±0.02)	0.95±0.09 (1±0)	1±0 (1±0)
		•	0.33±0.27 (0.68±0.19)	0.75±0.11 (0.85±0.04)	1±0 (1±0)
			Average Cross-Dataset Performance		
			0.52	0.82	0.46
•	•		-	-	0.90±0.18 (1±0)
•		•	-	1±0 (1±0)	-
	•	•	0.84±0.17 (0.95±0.05)	-	-

KNN model on the dataset E-GEOD-6791 was developed using the same dataset. This resulted in perfect performance on testing data, with MCC and AUC values of 1±0. The same model (i.e., KNN with a given parameter set and feature set) also performed well when applied to another dataset, E-GEOD-9844, giving MCC and AUC values of 0.89±0.10 and 0.94±0.06. However, when applied to the third dataset, E-GEOD-23036, very poor performance was observed, with the mean MCC near zero and the mean AUC near 0.5. This example demonstrates the lack of model robustness across datasets. Similar patterns are observed for the other datasets for KNN, as well as for the DT and SVM models in Tables 5.5 and 5.6, respectively.

The lower three rows of Table 5.4 show the model performances resulting from combining the optimal feature sets of the other two models. For example, the bottom-most row shows that using the combined optimal feature sets of the E-GEOD-9844 and E-GEOD-23036 KNN models to develop a KNN model for E-GEOD-6791 resulted in MCC and AUC values of 0.84±0.17 and 0.95±0.05, respectively. While this is lower than

Table 5.5. Multi-Dataset Performance of DT Models through MCC (AUC)

Data for Model Development			Prediction Dataset		
E-GEOD-6791	**E-GEOD-9844**	**E-GEOD-23036**	**E-GEOD-6791**	**E-GEOD-9844**	**E-GEOD-23036**
●			1±0 (1±0)	0.88±0.10 (0.94±0.06)	0±0 (0.5±0)
	●		0.93±0.12 (0.94±0.10)	0.76±0.11 (0.86±0.10)	1±0 (1±0)
		●	-0.07±0.09 (0.62±0.11)	0.82±0.19 (0.90±0.10)	1±0 (1±0)
			Average Cross-Dataset Performance		
			0.43	0.85	0.50
●	●		-	-	1±0 (1±0)
●		●	-	0.88±0.10 (0.94±0.06)	-
	●	●	0.93±0.12 (0.94±0.10)	-	-

Table 5.6. Multi-Dataset Performance of SVM Models through MCC (AUC)

Data for Model Development			Prediction Dataset		
E-GEOD-6791	**E-GEOD-9844**	**E-GEOD-23036**	**E-GEOD-6791**	**E-GEOD-9844**	**E-GEOD-23036**
●			1±0 (0.5±0)	0.89±0.10 (0.54±0.07)	0.09±0.31 (0.5±0)
	●		0.85±0.14 (0.98±0.03)	0.89±0.10 (1±0)	0.90±0.18 (1±0)
		●	0.51±0.28 (0.82±0.12)	0.88±0.10 (0.99±0.02)	1±0 (1±0)
			Average Cross-Dataset Performance		
			0.68	0.89	0.50
●	●		-	-	0.90±0.18 (1±0)
●		●	-	0.89±0.10 (0.5±0)	-
	●	●	0.76±0.22 (0.98±0.03)	-	-

the performance of the optimal model developed for E-GEOD-6791 itself, it is higher than the average cross-dataset performance observed from applying either of the two other models to this dataset. For this particular group of datasets, E-GEOD-6791 and E-GEOD-23036 showed poor performance during cross-prediction tests with the other, while E-GEOD-9844 showed more stable performance. Similar trends for cross-prediction are shown for all classifiers tested. Overall, utilizing a composite feature set

aggregated from multiple models appears to yield more robust predictions for previously unseen data.

5.4.2. Performance of Microarray-Developed Models on RNAseq Data

Overall, individual (non-ensemble) models developed using microarray data performed reasonably well on RNAseq data, as shown in Figure 5.3. This experiment tracks the distribution of mean MCC values of each model category across 100 repetitions of 3-fold CV on the RNAseq dataset. The median performance of the KNN, DT, and SVM models when applied to the RNAseq data was 0.73, with the best results for any CV repetition approaching 0.86 (result set (a)). The feature combination approach (b) showed a slight increase in median performance, but also resulted in many low-performing outliers. Voting using single-classifier models (c-e) showed slight increases in median performance – the KNN-only and DT-only ensembles achieved median performances of 0.78, and the SVM-only ensemble reached 0.81 – and also increases in

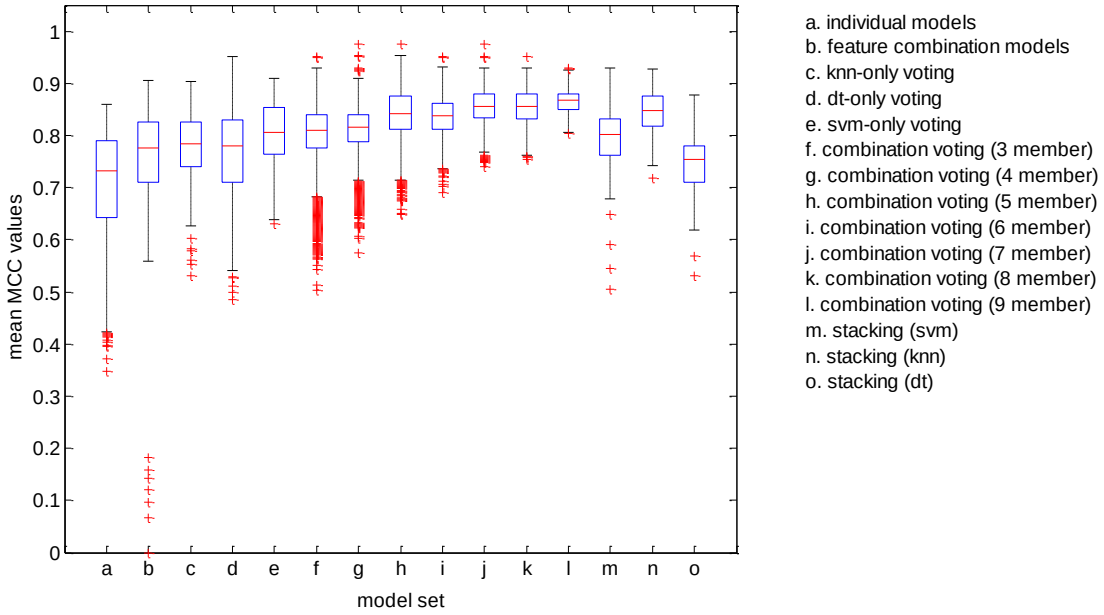


Figure 5.3: Comparison of alternative individual and ensemble models developed from microarray data when applied to predict HNSCC vs. normal samples from RNAseq data.

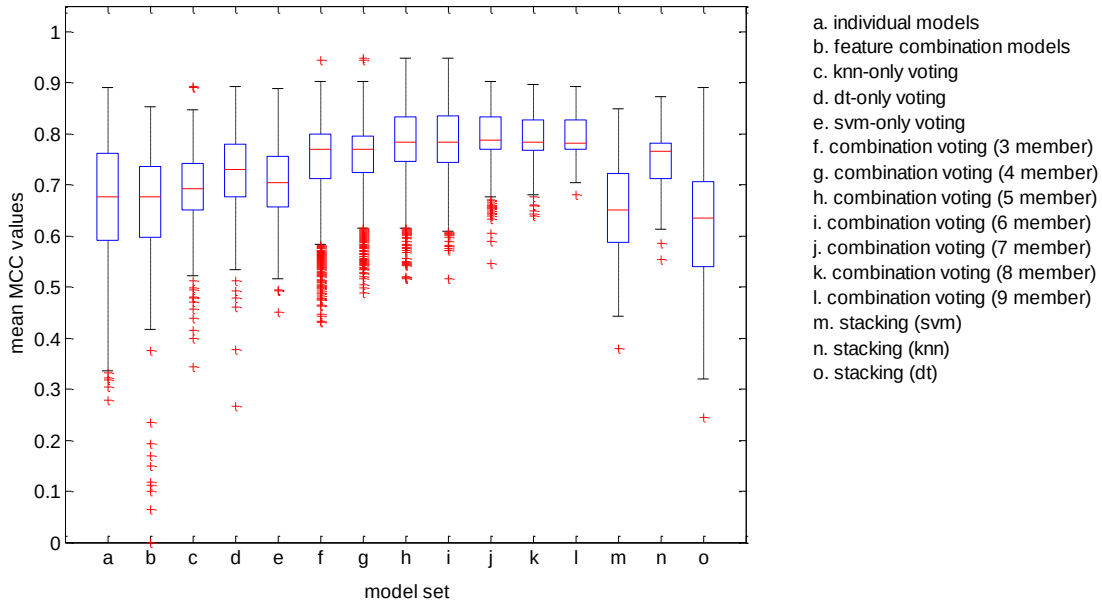


Figure 5.4: Comparison of alternative individual and ensemble models developed from microarray data when applied to predict early stage HNSCC vs. normal samples from RNAseq data.

the minimum performance. The combination voting approaches (f-l) had a trend of slightly increasing median performance and of lower variation as the number of models in the combination increased. The seven-, eight-, and nine-member ensembles had best overall median performances, ranging from 0.85-0.87. Among the stacking ensembles (m-o), the best median performance was observed with KNN as the secondary classifier.

In total, 42 models were developed which had better performance than the best-performing individual model (mean MCC = 0.8599) in at least 50 of the 100 CV repetitions. These models were part of the five-, six-, seven- and nine-member voting ensembles, which also had smaller amounts of variation than many of the other model categories. Most of these models had instances of statistically significant improvement ($p \leq 0.05$) over the best-performing individual model, as assessed by Steiger's Z-test [219].

Figure 5.4 shows the performance of the same models when applied to the early stage vs. normal RNAseq data, across 100 repetitions of 3-fold CV. Overall, performances are slightly lower and also more variable, reflecting the more challenging nature of the classification problem. Otherwise, similar trends were observed across the model categories. In terms of mean MCC, the median performance for the individual models (a) was 0.68. The median performances of the combination voting ensembles were in the range of 0.77-0.78. Unlike in the previous experiment, no models had better performance than the best-performing individual model across any CV repetition (MCC = 0.89) for more than 50 of the 100 CV repetitions. This is due in part to the overall lower model performances in most categories, as well as the slightly higher value of the maximum individual performance for this experiment. However, almost all of the combination voting models and one stacking model (f-o) exceeded the median performance of the individual models.

Tool Design

The suite of microarray models was integrated into a MATLAB GUI that allows users to (i) load a new dataset of interest, (ii) select a model developed using previously examined datasets, or compare all models, and (iii) implement the selected model(s) and visualize and export the results. Figure 5.5 shows a screenshot of the interface. The goal of developing this system is to enable HNSCC researchers, particularly those from more clinical-oriented, non-computational backgrounds, to take advantage of predictive modeling resources developed by the computational research community. In particular, by gathering multiple models from different datasets together in a single tool, it becomes

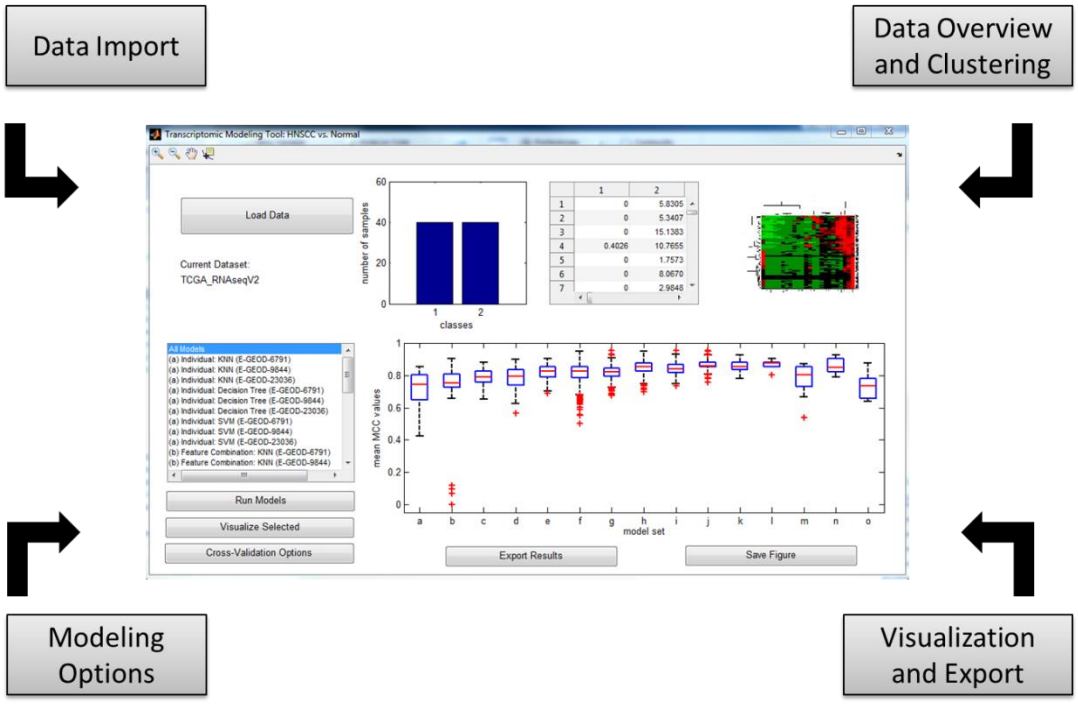


Figure 5.5: Screenshot of tool interface, displaying import, analysis, visualization, and export capabilities.

easier to implement ensemble approaches that improve overall performance.

5.5. Discussion and Key Innovations

Consistency among different studies lends support to research findings; in the same way, consistent performance among datasets increases confidence in a predictive model, and in the functional importance of the features that it utilizes. In this chapter, I developed predictive models for HNSCC using gene expression data that exhibit robust performance both within and between transcriptomic data types. Other recent transcriptomic HNSCC studies have also compared results across several datasets. De Cecco and colleagues used three microarray datasets to develop their model, and tested it on six other datasets, including TCGA RNAseq data [227]; however, the endpoint of interest in their study was risk of relapse, not diagnosis. Saintigny and colleagues tested

their model for risk of oral cancer development in leukoplakia patients on nine other microarray datasets [39]. In addition, neither of these studies considered multiple classification approaches and ensemble methods, as in this study. Ye and colleagues performed a meta-analysis across 63 HNSCC transcriptomic studies, considering premalignant lesions vs. normal samples, primary tumors vs. normal samples, and primary tumors vs. metastatic disease [125]; however, the study focused identifying key genes and pathways, and did not build predictive models. This study contributes to the existing literature on transcriptomic analysis for HNSCC by considering several alternative modeling frameworks for the endpoint of disease state, with an application of early diagnosis. Overall, multiple models with good performance (MCC $\geq$ 0.8, AUC $\geq$ 0.8) were identified. In addition, I compared and identified ensemble strategies that increased model performance for differentiating both general and early HNSCC from normal samples.

Another direction for further research is in the integration of protein and gene expression data for early HNSCC detection. Several recent studies have investigated the use of salivary RNA and/or proteins for detecting oral cancer [235-237]. Some of the salivary RNA markers validated in [235] – *IL-1B, IL-8*, and *H3F3A* – were also selected in the DE gene lists in this study, and *IL-8* was one of the mRMR-selected features for E-GEOD-9844. This observation is promising in both directions: applying other feature selection methods to the current group of datasets may reveal more previously-validated markers, and future validation studies may support the clinical relevance of features used in the current models. In addition, in Chapter 4, I have demonstrated that combining transcriptomic and proteomic models increases performance when predicting HNSCC

pathological stage [238]. Thus, models combining multiple –omic data types may also improve performance for early disease detection.

While current results are encouraging, more systematic testing, comparison, and refinement of models will be possible with additional and larger datasets. The three microarray datasets investigated here collectively include only 158 samples, and the number of matched tumor-normal RNAseq samples currently available in TCGA is also limited. However, data availability – and particularly for early stage disease – is always a limitation in cancer research. Therefore, one of the design goals of the modeling tool is to continually update its collection of individual and ensemble models as users upload additional labeled transcriptomic data. In this sense, it can serve to accelerate translational research. In the process, the tool can also be expanded to accommodate data and models for other prediction endpoints, such as length of survival, recurrence, and response to alternative therapies. It can also consider specific subsets of HNSCC, such as HPV+ vs. HPV- disease [239]. Thus, it can become a central component of a future clinical decision support system for assisting in HNSCC diagnosis and treatment planning.

The Key Innovations of this chapter are:

- Performed within–omic level integrative modeling study using microarray and RNAseq data for detection of HNSCC
- Translated ensemble models developed for discriminating between HNSCC and paired normal cases to the problem of early HNSCC detection
- Implemented tool to facilitate model translation and use of ensemble transcriptomic models in the HNSCC research community

CHAPTER 6

DYNAMIC SYSTEM MODELS FOR PREDICTION OF RESPONSE TO COMBINATION ADJUVANTS

6.1. Chemoprevention in HNSCC

Currently, HNSCC treatment options include surgery, radiation, chemotherapy, or combinations of these treatments [8]. Many patients with locally advanced (stage III/IV) disease respond favorably to initial treatment, but later experience locoregional recurrence, secondary primary tumor (SPT) development, or metastatic disease [9-13]. Chemoprevention is defined as the application of natural or synthetic agents to delay or prevent cancer progression. Adjuvant chemoprevention therapies have been shown to improve overall and disease-free survival in HNSCC; however, toxicity is a limiting factor [49, 240]. Therefore, the identification of safe, non-toxic adjuvant therapies for chemoprevention in HNSCC is of great clinical interest.

Because of these characteristics, natural compounds from dietary agents are promising as chemoprevention adjuvants for HNSCC. The primary catechin found in green tea, (-)epigallocatechin gallate (EGCG), has been shown to be an effective antioxidant and has a wide range of effects on signal transduction pathways implicated in cancer [240]. It affects multiple processes including cell proliferation and division, angiogenesis, and apoptosis. Recent phase II clinical trials have indicated that green tea extract is effective in preventing oral cancer development in patients with premalignant oral lesions [241, 242]. However, the effects of EGCG alone are limited by low oral bioavailability [54, 55, 243]. Thus, the identification of effective combinations of EGCG and other natural compounds is of interest, since natural compound combinations may

yield more-than-additive effects while maintaining low toxicity profiles. For example, green tea catechin in combination with curcumin, which is found in turmeric, has been shown to have synergistic apoptotic activity in larynx carcinoma cell lines [56]. EGCG in particular has been shown to synergistically increase apoptosis in HNSCC cell lines when combined with luteolin, an antioxidant found in many green vegetables [244], as well as with resveratrol, which is found in grape skins and red wine [53].

6.1.1. Prediction with Dynamic System Models

Predicting effective combinations of natural compounds is challenging due to their multi-target effects on complex biochemical signaling networks. Mathematical modeling for cancer is a diverse and growing area of research; although inherently much simpler than the complex biological systems represented, models provide tools for predicting outcomes and generating testable hypotheses [58-60]. Mathematical models can assist chemoprevention research by relating the activities of individual and combination agents to cellular-level responses, such as proliferation, survival, and apoptosis. For example, my prior work involved developing an agent-based model to predict the response of an HNSCC cell line to the combination of paclitaxel and the anti-angiogenic compound 2-methoxyestradiol (2ME2) [245]. These types of models may help to advance clinical research via the generation of specific, testable hypotheses, i.e., the response to alternative drug combinations, as well as the prediction of specific therapeutic targets that could increase favorable responses. In addition to the previously mentioned model, for HNSCC, models have been developed to predict the effects of radiotherapy [246, 247], and some models do so by incorporating clinical imaging data [248-250]. Other models focus on optimizing radiotherapy-chemotherapy combination

treatments [251, 252], and yet another focus is the prediction of nanoparticle drug uptake [253]. However, these models do not take into account the molecular pathway-level processes by which natural compounds exert their effects. In addition, previous mathematical models for chemoprevention for multiple cancer types mainly focused on cost-effectiveness, not biological effectiveness, and considered the effects of conventional chemotherapeutics [254-256].

To address this issue, this study develops a multi-scale dynamic model for predicting the response to natural compound chemoprevention agents in HNSCC, based on their targeted effects on signal transduction pathways. In computational cancer research, multi-scale dynamic models are those which describe behaviors at multiple spatial scales, and potentially also across different time scales. Possible spatial scales encompass the atomic, molecular, cellular, tissue, organ, and patient levels. The model developed here describes behaviors at the molecular and cellular levels. The model is applied to predict the combination effects of EGCG and resveratrol in several HNSCC cell lines. I also demonstrate how the multi-scale design enables use of the model for hypothesis generation, including the prediction of specific pathway targets and potential effective natural compound combinations. In addition, the initially developed multi-scale ordinary differential equation (ODE) model is then coupled to an agent-based model (ABM), which enables natural compound response prediction in complex, heterogeneous cellular environments. These models provide groundwork for advancing research into safer, non-toxic chemoprevention adjuvants for HNSCC from a computational perspective.

6.2. Model Development

6.2.1. Cell Lines and Dose Response Data

Dose response data from three HNSCC cell lines – Tu212, Tu686, and SQCCY1 – were used to develop and test the model. For Tu212, the percentage of apoptotic cells (early and advanced apoptosis) was measured for six dosage levels for resveratrol and 10 dosage levels for EGCG, as shown in Figure 6.1. The combination response was measured for four levels: 30µM EGCG with 10µM or 15µM resveratrol (abbreviated E30R10 and E30R15, respectively) and 40µM EGCG with 10µM or 15µM resveratrol (E40R10 and E40R15). For Tu686 and SQCCY1, the combination response was measured for 12 levels each: 15µM or 20 µM resveratrol with 30, 40, 50, 60, 70, or 80µM EGCG.

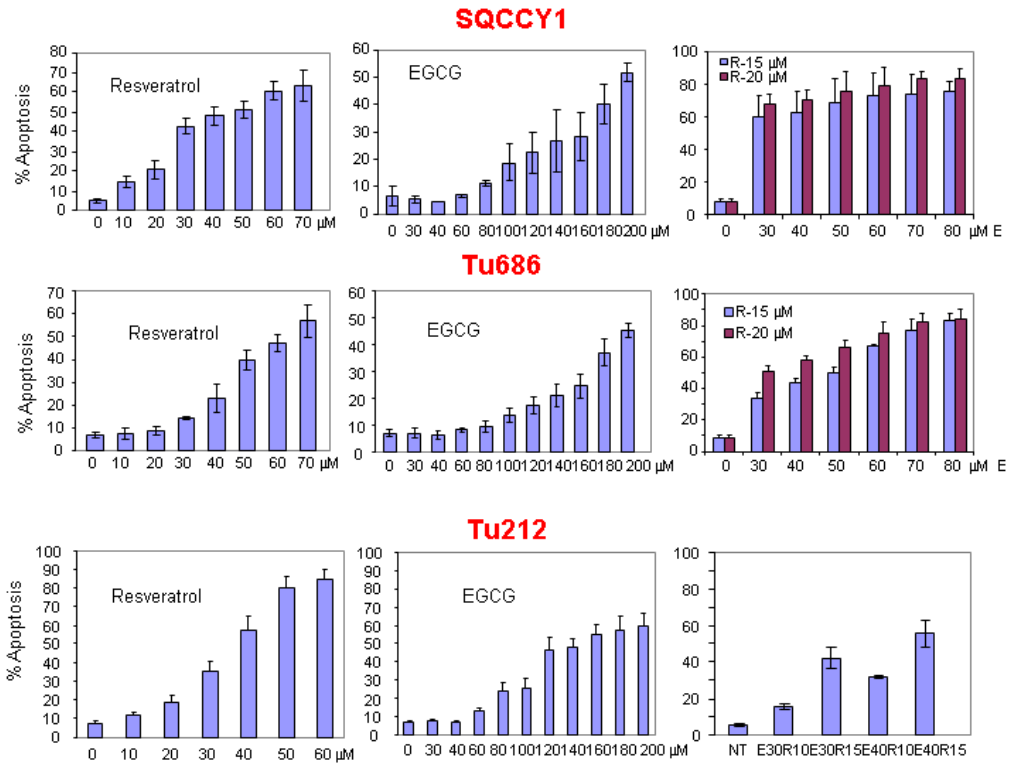


Figure 6.1: Dose response data for the Tu212, Tu686, and SQCCY1 cell lines. Image courtesy of Dr. A.R.M.R. Amin at Winship Cancer Institute.

Modeling Workflow

Available dose response data was separated into training and testing sets in order to estimate model parameters. Training data comprised of the EGCG-only and resveratrol-only dose response measurements, along with all but two of the combination responses. The remaining two responses were used for testing. For Tu212, $\binom{4}{2} = 6$ training-testing splits were evaluated, and for the other two cell lines, $\binom{12}{2} = 66$ splits were tested. Parameters were estimated by minimizing the root-mean-square error (RMSE) between the simulated and experimentally observed percentages of apoptosis. Two alternative optimization methods were tested, both with the constraint that all parameters be non-negative: sequential quadratic programming (SQP) and the genetic algorithm (GA). SQP is a deterministic, gradient-based optimization method in which a quadratic programming sub-problem is solved at each iteration. For SQP, a constant initial parameter estimate of $p \in \mathbb{R}^{33}$ was used, where $p_i = 0.1 \; \forall \; i \in [1, 2, 3, \ldots, 33]$. The GA is a direct-search optimization method which uses evolutionary mechanisms to explore the parameter space to approach a global minimum. The initial GA population had 1,250 entries, of which 250 were sampled from $U\sim[0,1]$; 250 were sampled from $U\sim[0, n]$, where n is twice the maximum value observed for the SQP-optimized parameters from any training-testing split for the Tu212 cell line; and 750 were obtained by adding normally distributed noise to the SQP-optimized parameters from all six trials for the Tu212 cell line: $p_{i,GA} = p_{i,SQP} + v_i \; s.t. \; v_i \sim N\left(0, \frac{2p_{i,SQP}}{10}\right), \forall \; i \in [1, 2, 3, \ldots, 33]$. Pairwise Pearson correlations between optimal parameter estimates obtained through each trial and each optimization method were used to assess the consistency of estimates.

6.2.2. Single-Scale Models

Two alternative models were considered to model the combination drug effects, using cellular-level processes only. In the first case, the living and apoptotic (both early and late apoptosis) cell dynamics are tracked by first-order ODEs, as shown in Table 6.1. EGCG and resveratrol effects are modeled through the apoptosis rate parameter k_{death}. The observed nonlinear effects are naively modeled through higher-order functions of natural compound concentrations.

In the second case, the Combination Index (CI), a measure used to assess drug combinations, was applied, as shown in Table 6.2. The CI indicates whether the combined effect of two drugs is additive (CI = 1), synergistic (CI < 1), or antagonistic (CI > 1). Chou and Talalay related the CI to administered drug ratios [257, 258]. The training

Table 6.1: Naïve ODE single-scale model

$[cells] = [living] + [apoptotic]$	The population is divided into living and apoptotic cells. Early and late apoptotic cells are pooled together.
$\frac{d[living]}{dt} = k_{division}[living] - k_{death}[living]$	
$\frac{d[apoptotic]}{dt} = k_{death}[living]$	
$k_{death} = k_{baseline} + r_1[res] + e_1[egcg] + \sum_{i=2}^{K=9} \lambda_i([res] + [egcg])^i$	Synergistic effects are naively modeled using higher-order concentration terms.

data was used to estimate the CI value and the Hill function parameters. The combination effect *E* for the testing data was then estimated using Nelder-Mead simplex direct search.

6.2.3. Multi-Scale Ordinary Differential Equation Model

The multi-scale ODE model modifies the naïve single-scale model by defining the division rate parameter $k_{division}$ and the apoptosis rate parameter k_{death} as functions of molecular species known to regulate these processes. In addition, the targeted effects of EGCG and resveratrol are modeled. Thus, both cellular-level and molecular-level factors are considered. The molecular-level model describes a system comprised of signal

Table 6.2: CI-based single-scale model

$\frac{C_A^*}{C_A} + \frac{C_B^*}{C_B} = CI$	C_A^*, C_B^* are the amounts of drugs A and B in the combination, while C_A, C_B are the amounts of drugs A and B that would yield the same effect as the combination if each was administered alone.
$\frac{C_A^*}{IC_{50,A}\left(\frac{E}{1-E}\right)^{\frac{1}{h_A}}} + \frac{C_B^*}{IC_{50,B}\left(\frac{E}{1-E}\right)^{\frac{1}{h_B}}} = CI$	The Hill function of order h_d relating the probability of response p_d to the drug concentration *x* is: $p_D(x) = \frac{x^{h_d}}{x^{h_d} + IC_{50,x}^{h_d}}$. Using this to model the response curves of drugs A and B individually, the CI equation can be expressed as shown. *E* is the combination effect.

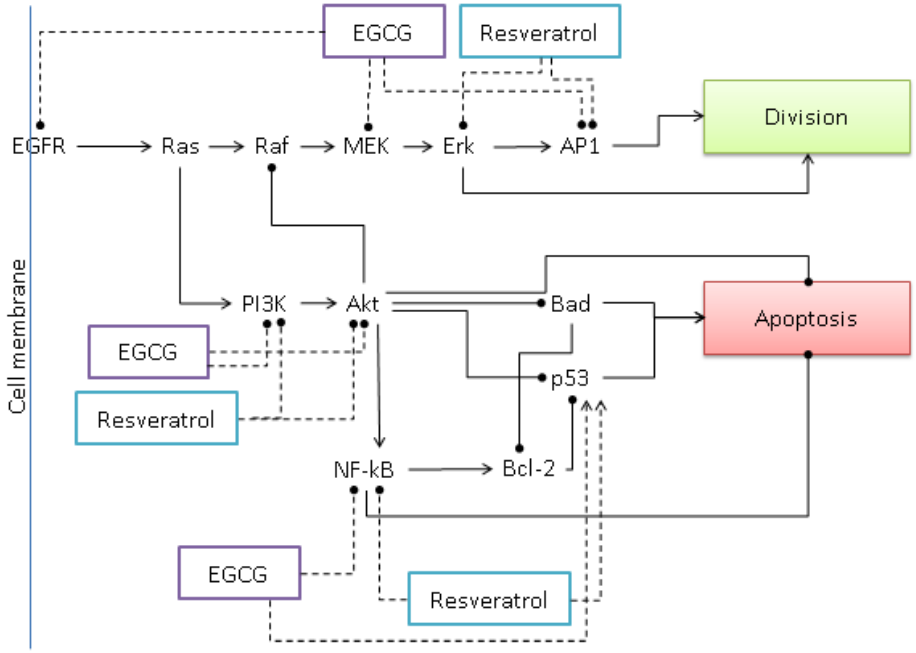


Figure 6.2: Signal transduction pathways represented in the molecular-level model

transduction pathways known to be highly relevant in many cancers, including HNSCC. These are the MAPK/ERK pathway, PI3K-Akt signaling, and their effects on modulators of apoptosis including p53, Bcl-2, and BAD. Figure 6.2 shows this signaling network, as well as the points at which the effects of EGCG and resveratrol are modeled. Some molecular targets are affected by both agents. EGCG has been shown to induce p53 expression in multiple cell types, and resveratrol is also associated with p53 activation [243]. In addition, both EGCG and resveratrol have been shown to inhibit NF-kB signaling [243], AP-1 activity [243, 259, 260], and PI3K-Akt [243, 261-263]. EGCG also inhibits phosphorylation of EGFR and association of Raf1 and MEK1 [243, 264]. Resveratrol has dose-dependent effects on phosphorylation of Erk1/2 [265, 266]. Table 6.3 shows the generalized mass action model for this system, and the relationship between the molecular-level model and the cellular-level model.

The general structure of the generalized mass action equations in Table 6.3 is as follows:

$$\frac{d[X]}{dt} = v_A[X][X_{activating\ factor}] - x_A[X]\left[X_{inhibiting\ factor}\right]$$

This is of course a basic approximation of the complex biochemical interactions occurring in the signal transduction network. Here, *X* is assumed to be the activated (e.g., phosphorylated) form of the molecule. The differential equation is second-order because the process of de-activation (e.g., dephosphorylation) is being described implicitly. That is, the kinetic parameter v_A is assumed to represent the rate of activation by $X_{activating\ factor}$ scaled by the proportion of *X* which has become de-activated, and which can thus be activated by association with $X_{activating\ factor}$. The same pattern is followed for the kinetic parameter x_A and $X_{inhibiting\ factor}$. A more complete representation of the activation or inhibition processes could include separate variables and equations for the active and inactive forms of *X*. For example:

$$[X_{total}] = [X] + [X_{inactive}]$$

$$\frac{d[X]}{dt} = v_A[X_{inactive}]\left[X_{activating\ factor}\right] - x_A[X]\left[X_{inhibiting\ factor}\right] - k_d[X]$$

$$\frac{d[X_{inactive}]}{dt} = -v_A[X_{inactive}]\left[X_{activating\ factor}\right] + x_A[X]\left[X_{inhibiting\ factor}\right] + k_d[X]$$

In this representation, active *X* would be generated by the association of $X_{activating\ factor}$ with the inactive form of *X*, and it would be removed either by association with $X_{inhibiting\ factor}$ or natural degradation back to its inactive state, governed by the kinetic rate constant k_d. This representation also makes the simplifying assumption that the total amount of *X* (active and inactive forms) remains constant under the time-scale of interest.

Table 6.3: Multi-scale ODE model

$\frac{d[EGFR]}{dt} = -e_1[EGCG][EGFR]$ $\frac{d[Ras]}{dt} = v_1[EGFR][Ras]$ $\frac{d[Raf]}{dt} = v_2[Ras][Raf] - x_1[Akt][Raf]$ $\frac{d[MEK]}{dt} = v_3[Raf][MEK] - e_2[EGCG][MEK]$ $\frac{d[Erk]}{dt} = v_4[MEK][Erk] - r_1[Res][Erk]$ $\frac{d[AP1]}{dt} = v_5[Erk][AP1] - r_2[Res][AP1] - e_3[EGCG][AP1]$ $\frac{d[PI3K]}{dt} = v_6[Ras][PI3K] - r_3[Res][PI3K] - e_4[EGCG][PI3K]$ $\frac{d[Akt]}{dt} = v_7[PI3K][Akt] - r_4[Res][Akt] - e_5[EGCG][Akt]$ $\frac{d[NFkB]}{dt} = v_8[NFkB][Akt] - r_5[Res][NFkB] - e_6[EGCG][NFkB]$ $\frac{d[p53]}{dt} = -x_3[Bcl2][p53] - x_4[Akt][p53] + r_6[Res][p53] + e_7[EGCG][p53]$ $\frac{d[Bcl2]}{dt} = v_9[Bcl2][NFkB] - x_2[BAD][Bcl2]$ $\frac{d[BAD]}{dt} = -x_5[BAD][Akt]$	The signal transduction pathway model includes 12 molecular species. The multi-target effects of EGCG and resveratrol are modeled directly on the various biochemical entities in the pathway. The model includes 33 rate parameters comprising activating inter-molecular interactions (v_i), inhibitory interactions (x_i), pro-proliferation (p_i) and pro-apoptosis (a_i) effects, and the effects of EGCG (e_i) and resveratrol (r_i).
$\frac{d[living]}{dt} = k_{division}[living] - k_{death}[living]$ $\frac{d[apoptotic]}{dt} = k_{death}[living]$	Again, the population is divided into living and apoptotic cells. Early and late apoptotic cells are pooled together.
$k_{division} = p_{21}[Erk] + p_{22}[AP1]$ $k_{death} = \max(0,\ a_{21}[p53] + a_{22}[BAD] - a_{51}[Akt] - a_{52}[NFkB])$	The division and apoptosis kinetic rate parameters are direct functions of pathway entity concentrations.

Another, even more complex and realistic representation could include protein transcription, translation, and degradation rates, as well as detailed enzymatic response patterns like Michaelis-Menten or multi-substrate kinetics. The reason for avoiding these representations is the lack of measured kinetic rate constants and the lack of time-series molecular expression data for estimating these parameters. This is problematic because the number of these parameters will increase as the model increases in complexity. In the future, as additional data become available, this foundational model can be expanded to accommodate more specific biochemical interactions of interest.

6.2.4. Multi-Scale Agent-Based Model

A key assumption behind ODE models is that the population is well-mixed, with no spatial gradients. This is appropriate for an *in vitro* study in which a homogeneous cell population is evenly distributed in its environment. However, research into natural compounds for chemoprevention also involves more complex *in vitro* environments (e.g. multicellular spheroids) and *in vivo* studies. ABMs are well-suited for representing heterogeneous cell populations, inter-cellular interactions, cell movement, and the complex spatial structure of tumors. The multi-scale ODE model was therefore used to develop a multi-scale agent-based model (ABM).

In this ABM framework each agent represents a single cell. Instead of kinetic rate parameters, ABMs utilize transition probabilities, i.e., the probability that each agent experiences division (p_d) or apoptosis (p_a) during a given time step. The multi-scale ODE is used to drive the ABM by enabling estimation of the apoptosis probability at each time step. For the purpose of estimating p_a, p_d is assumed to be constant, based on the cell cycle duration of the Tu212 cell line (~24 hours). The fraction of apoptosis predicted by

Table 6.4: Coupling the multi-scale ODE and agent-based models

Total living cells after *T* time steps: $N_0(1+p_d)^T(1-p_a)^T$ Total apoptotic cells after *T* time steps: $\sum_{i=1}^{T} N_0(1+p_d)^i(1-p_a)^{i-1}p_a$ $e = c \times \left(\sum_{i=1}^{T} N_0(1+p_d)^i(1-p_a)^{i-1}p_a + N_0(1+p_d)^T(1-p_a)^T \right) - \sum_{i=1}^{N} N_0(1+p_d)^i(1-p_a)^{i-1}p_a$	If N_0 cells are initially present, total numbers of living and apoptotic cells after a certain number of time steps (*T*) are represented in terms of probabilities of division (p_d) and apoptosis (p_a). The overall probability of apoptosis (*c*) can be used to estimate p_a.

the ODE (for a homogeneous population of the agent-cell under consideration) defines *c*, the cumulative probability of apoptosis over the prediction timeframe. As shown in Table 6.4, p_a can then be estimated by minimizing the error *e*.

In this manner, the ODE model for a homogeneous population can be used to govern the behavior of a single agent-cell, though it may exist in a heterogeneous population. In turn, population heterogeneity can also affect the behavior of single agent-cells. While the multi-scale ODE provides the forward drive to the ABM, the ABM can also provide feedback to the ODE model. Characteristics of the tumor microenvironment, such as hypoxia, can significantly modulate the activities of signal transduction pathways, including EGFR and PI3K-Akt signaling, thereby affecting the behavior of individual cells [267, 268]. The hypoxic effect experienced by a single cell can be estimated through the spatial environment of the ABM, and fed back into the ODE, in order to more realistically model microenvironmental effects on individual cells.

6.3. Model Performance

Multi-scale model outperforms single-scale models in predicting response to EGCG and resveratrol combinations

As shown in Figure 6.3, the naïve single-scale model was unable to replicate experimentally-observed trends in either the training or testing sets, demonstrating poor performance overall. The CI-based model was able to make reasonable predictions of combination effects, but lacked the structure to enable causal analysis for alternative

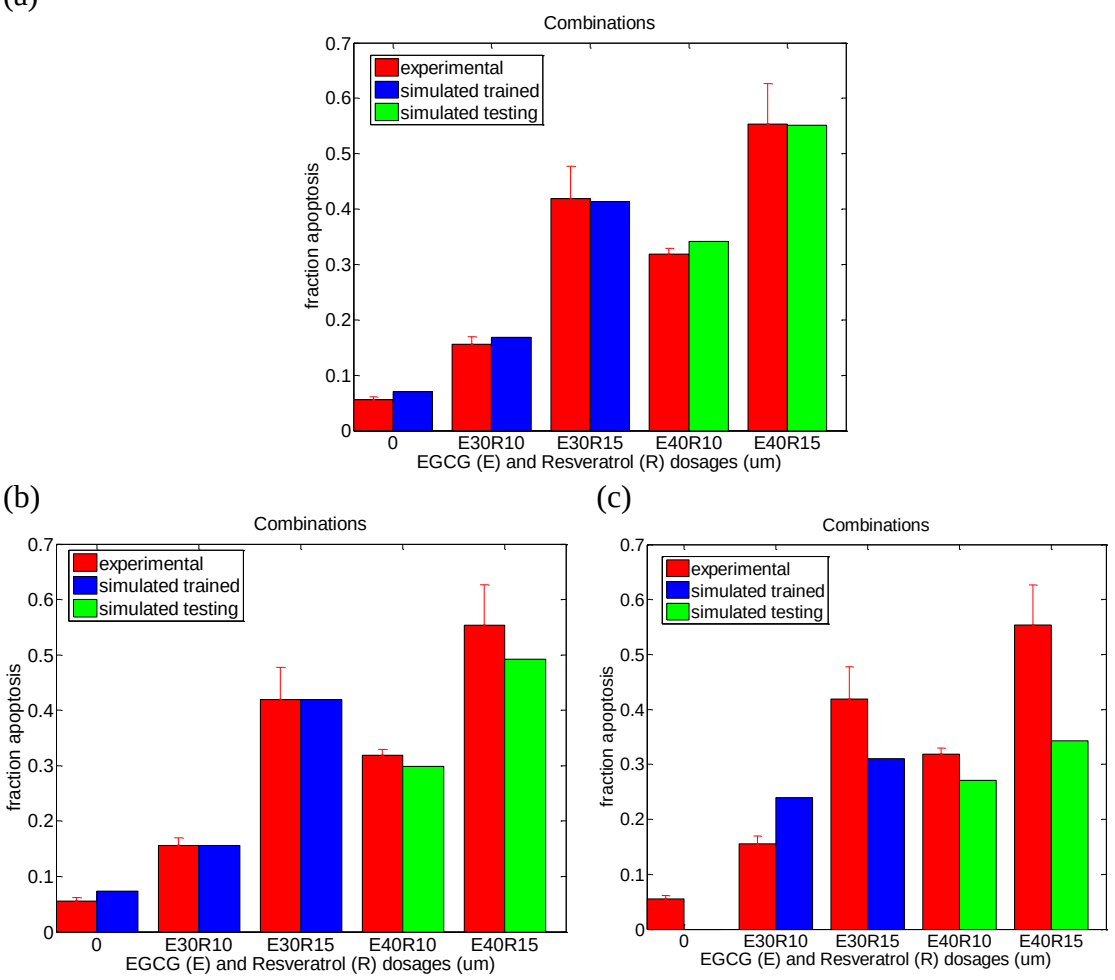


Figure 6.3: The performance of (a) the multi-scale model (RMSE = 0.0231), (b) the CI-based single-scale model (RMSE = 0.0578), and (c) the naïve single-scale model (RMSE = 0.1529) for a trial on Tu212 data, in which the E30R10 and E30R15 samples were used in training the models, and the E40R10 and E40E15 samples were used for testing. RMSE values are based on the testing samples only.

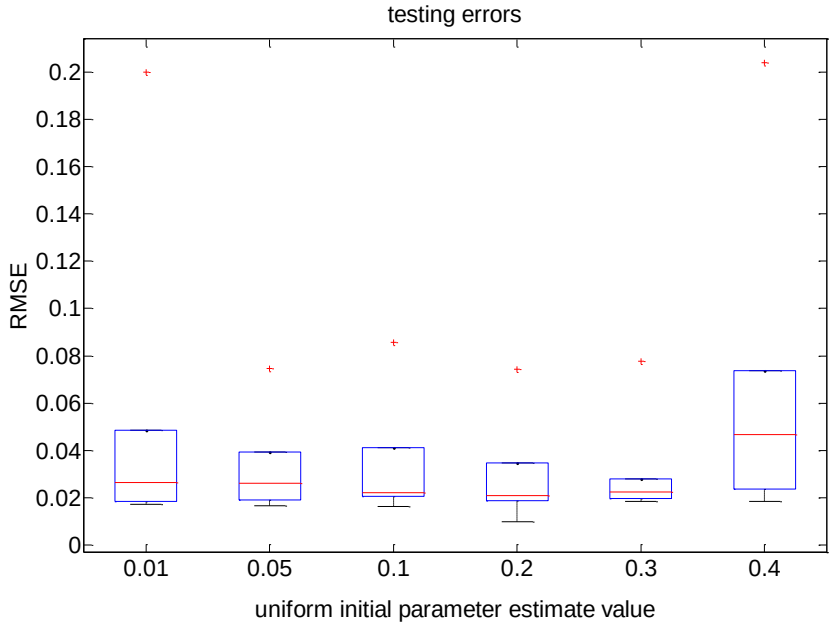
perturbations. The multi-scale ODE model predictions of EGCG and resveratrol combination effects were more accurate than those of either single-scale model. This improvement in performance was observed across all six trials for the Tu212 cell line. Overall, the improvement in performance by the multi-scale model (GA-optimized) was statistically significant ($p \leq 0.05$, Wilcoxon rank-sum test) compared to the CI-based and naïve single-scale models. The performance of the SQP-optimized multi-scale model was comparable; when discounting one outlier trial, it also showed statistically significant improvement over the single-scale models. Parameter estimates across trials and across optimization methods were generally consistent, exhibiting median Pearson correlation values exceeding 0.7.

In addition, at the final simulation time step, the normalized ratios of pathway elements (with respect to EGFR) were compared to the normalized gene expression ratios for the combination treatment case. While the ratios alone cannot be used to ascertain the correctness of the model, the high Pearson correlation value (r = 0.89) between the simulated and experimental ratios implies that the model predictions remain in a reasonable region of the state space.

SQP Optimization Shows Consistent Performance in Local Initialization Neighborhood

The results shown in the previous section were based off of a constant initial parameter estimate of $p \in \mathbb{R}^{33}$, where $p_i = 0.1 \; \forall \; i \in [1, 2, 3, \ldots, 33]$. Because the parameter optimization space may be complex and irregular, the effect of different initialization points on the SQP-based optimization was evaluated. The model performance for the Tu212 cell line across the six trials was evaluated in terms of RMSE

(a)



(b)

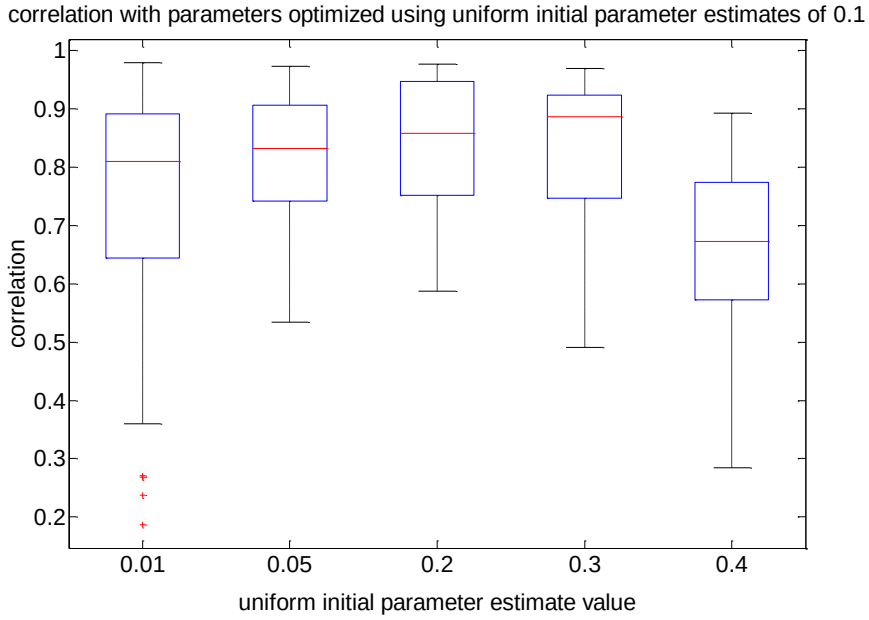


Figure 6.4. (a) Comparison of the SQP optimization performance at different initial parameter estimates for the Tu212 cell line, and (b) correlation of the optimized parameters obtained from different initial estimates with those obtained from the initial estimate of $p_i = 0.1 \; \forall \; i$.

for six alternative initialization points, as shown in Figure 6.4(a). Initializing at $p_i = 0.05 \; \forall \; i$, $p_i = 0.1 \; \forall \; i$, $p_i = 0.2 \; \forall \; i$, and $p_i = 0.3 \; \forall \; i$ gave similar distributions of testing

errors. Initializing at $p_i = 0.01 \; \forall \; i$ and $p_i = 0.4 \; \forall \; i$ gave an increased range of testing errors, and a much larger outlier value. Training errors (not shown) were consistently small for all initializations except for $p_i = 0.4 \; \forall \; i$.

Additionally, initializations with $p_i = 0.05 \; \forall \; i$, $p_i = 0.2 \; \forall \; i$, and $p_i = 0.3 \; \forall \; i$ yielded parameter estimates that had fairly high Pearson correlation values (median > 0.8) with those estimated using $p_i = 0.1 \; \forall \; i$, as shown in Figure 6.5(b). Notably, during the other two cases of $p_i = 0.01 \; \forall \; i$ and $p_i = 0.4 \; \forall \; i$, which were associated with higher testing error outliers and more variable correlations, stalling at low iteration counts was observed during the optimization process for some trials.

Model-based Comparison of Cell Line Responses

Alternative estimates of model parameters indicated differences in the responses of the three cell lines to the various combinations of resveratrol and EGCG. For example, Figure 6.5(a) shows the model predictions for the Tu686 cell line, using the parameters optimized for the Tu212 cell line. The combination response predictions track the experimental observations well, but the model over-predicted responses to treatment with resveratrol and EGCG individually, particularly for higher concentrations. This suggests that while Tu686 and Tu212 have similar responses to resveratrol and EGCG in combination, Tu686 is less sensitive to the individual treatments. Figure 6.5(b) shows the corresponding results for the SQCCY1 cell line; as for Tu686, the model over-predicted responses to treatment with resveratrol and EGCG individually. Moreover, compared to Tu686, the combination effect predictions for SQCCY1 when using the Tu212-optimized parameters were poorer. In comparison, model parameters optimized for the SQCCY1 cell line effectively predicted combination responses, as shown in Figure 6.6, as well as

individual responses to resveratrol and EGCG (not shown). However, the SQCCY1- and Tu212-optimized parameter sets exhibited low correlations, implying different patterns of activity.

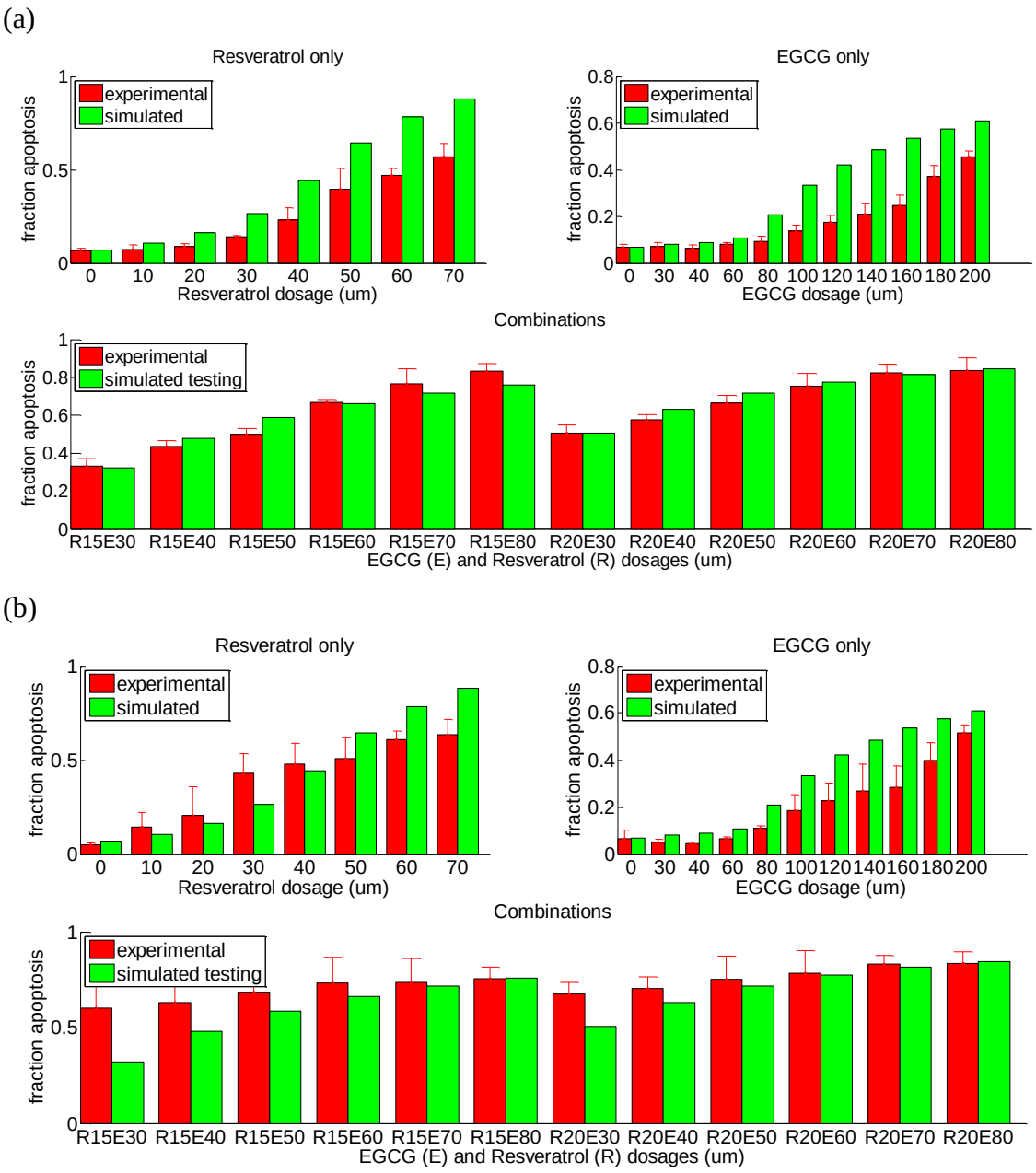


Figure 6.5. Comparison of the experimental and predicted responses to resveratrol alone (left), EGCG alone (right) and the 12 combination treatments (bottom) using Tu212-optimized parameters for (a) for the Tu686 cell line and (*b*) the SQCCY1 cell line.

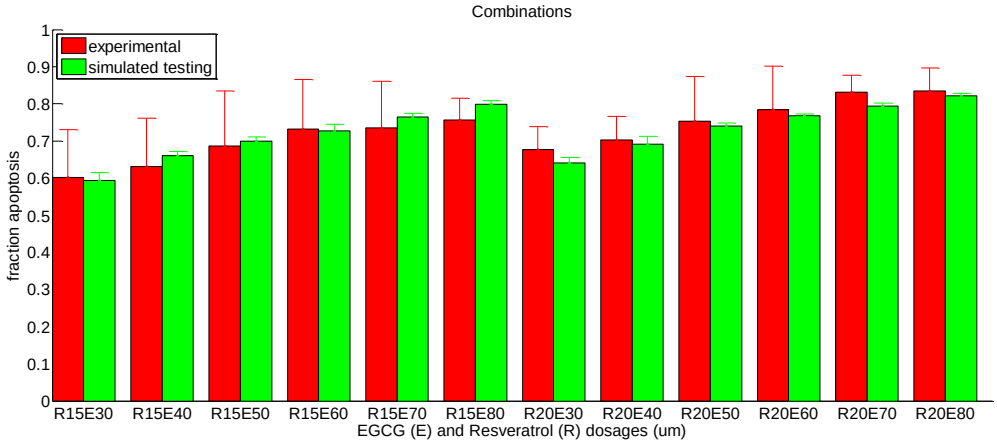


Figure 6.6. Comparison of experimental and predicted responses to the 12 combination treatments for the SQCCY1 cell line.

6.4. Case Studies

One of the main goals of developing dynamic models of biological systems is to predict how the system may respond to alternative perturbations. A perturbation may represent, for example, the effect of a drug on specific reaction or on a group of reactions initiated by a common molecule. The model can then be used to rank alternative perturbations in terms of their predicted effect. In this sense, the model can identify potential drug targets and generate testable hypotheses. In the first case study, the multi-scale ODE model is applied to predict which perturbations, in addition to the best-performing EGCG-resveratrol combination (E40R15), can further increase the fraction of apoptotic cells for the Tu212 cell line.

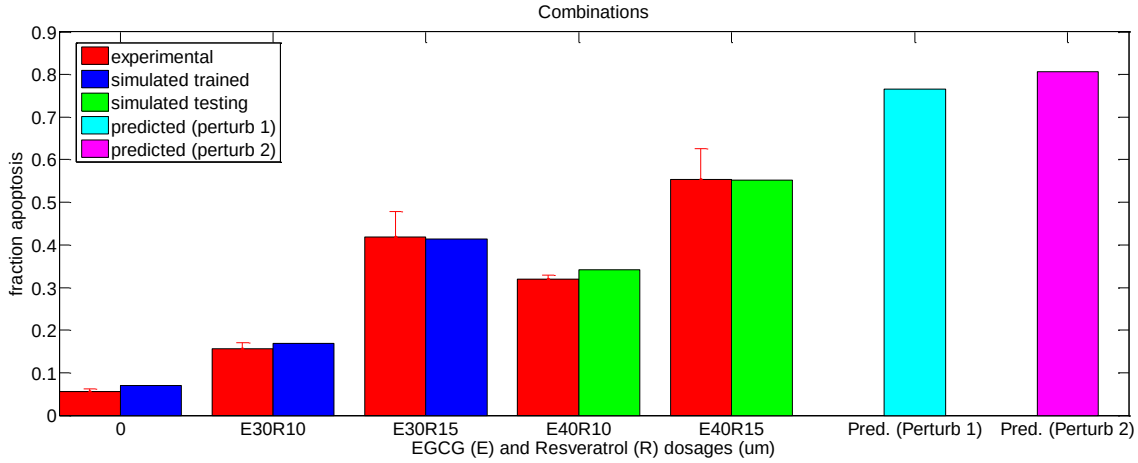
The experimental dose response data for EGCG and resveratrol in this study has been acquired from three HNSCC cell lines *in vitro*. In the second case study, the multi-scale ABM is applied to predict how the apoptosis patterns for EGCG and resveratrol combinations may change in more complex cellular environments, by considering the microenvironmental factor of hypoxia. Hypoxia is associated with many negative effects, such as suppression of apoptosis and increased cancer cell survival, increased

angiogenesis and invasiveness, and decreased sensitivity to both radiotherapy and chemotherapy [267, 269]. In HNSCC in particular, hypoxia measures are associated with poorer overall and disease-free survival [269]. One of the key signal transduction pathways affected by hypoxia is PI3K-Akt signaling. In HNSCC, hypoxia has been shown to increase activated Akt expression both *in vitro* and in xenograft models; moreover, this effect occurs independently of upstream EGFR status [267, 270]. The second case study will mimic these effects by varying the initial relative Akt activity input to the internal ODE model driving each agent cell, as a function of the degree of hypoxia experienced by that agent cell.

6.4.1. Target Prediction

Using the Tu212 cell line response data, 14 parameters modulating inter-molecular interactions were selected for perturbation analysis. The parameters related to EGCG and resveratrol effects and the proliferation and apoptosis rates were held constant. Two types of perturbations were considered: doubling and halving the original parameter values. Each type of perturbation was applied to each of the $\sum_{i=1}^{14}\binom{14}{i} = 16{,}383$ possible combinations of the 14 parameters. The combinations were ranked in terms of the fraction apoptosis observed after each perturbation. For both the SQP- and GA-derived parameter sets, the same top two perturbations were identified, as shown in Figure 6.7. If only one process was perturbed, halving Akt-mediated inhibition of p53 increased the predicted apoptotic fraction to above 0.75. If two processes were perturbed, halving Akt-mediated activation of NF-kB in addition increased the predicted apoptosis fraction to above 0.80. Further perturbations led to only slight increases in the predicted fraction of apoptotic cells.

(a)



(b)

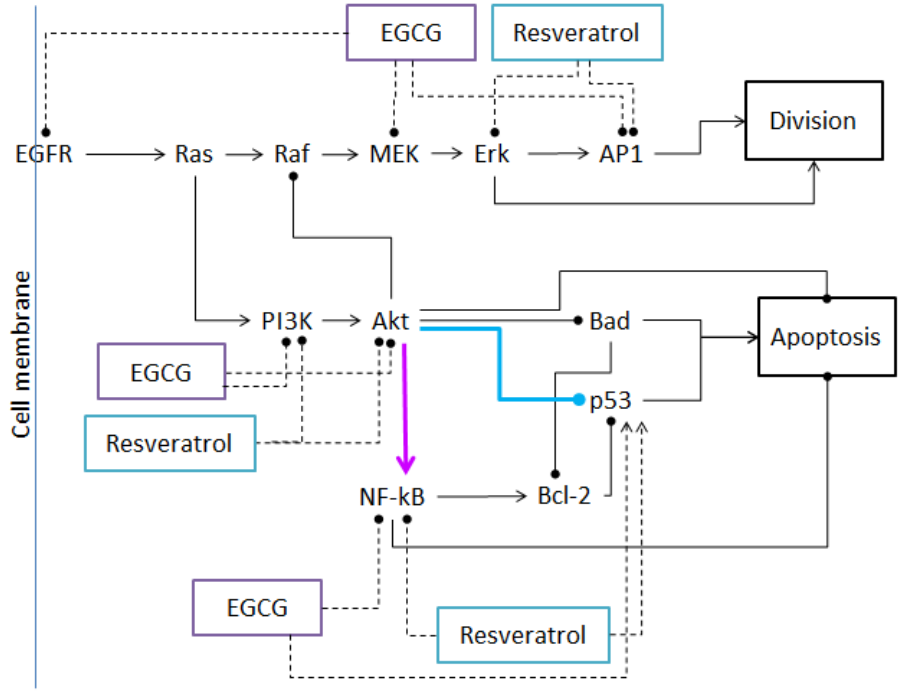


Figure 6.7: (*a*) The predicted apoptotic response from the top perturbation is shown in cyan, and the response for the top two perturbations is shown in magenta. (*b*) The processes within the signaling network targeted by each perturbation are highlighted in cyan and magenta, respectively.

These predictions emphasize the importance of Akt-mediated signaling in HNSCC. For example, the Head and Neck Cancer Tissue Array Initiative has shown that Akt-mTOR signaling is often activated in HNSCC, independently of mutant p53 or EGFR [271]. Moreover, the Akt signaling pathway is a key mechanism by which the cell

can bypass inhibition of EGFR [272, 273]. That Akt-mediated processes were the top two ranked perturbation targets suggests that additional synergistic effects may be observed by combining EGCG and resveratrol with other natural compounds that target Akt signaling, such as curcumin, pomegranate, and lycopene [243]. For example, a recent study demonstrated that combining resveratrol with curcumin induced greater pro-apoptotic effects in several HNSCC cell lines than curcumin alone [274].

6.4.2. Spatial Feedback and Effects of Hypoxia

The ABM was used to investigate how spatial structures and hypoxic effects might affect responses to the resveratrol and EGCG combinations. Figure 6.8 describes this workflow and the interaction between the ODE and the ABM.

The first case involved a uniform random distribution of cells (Figure 6.9(a)),

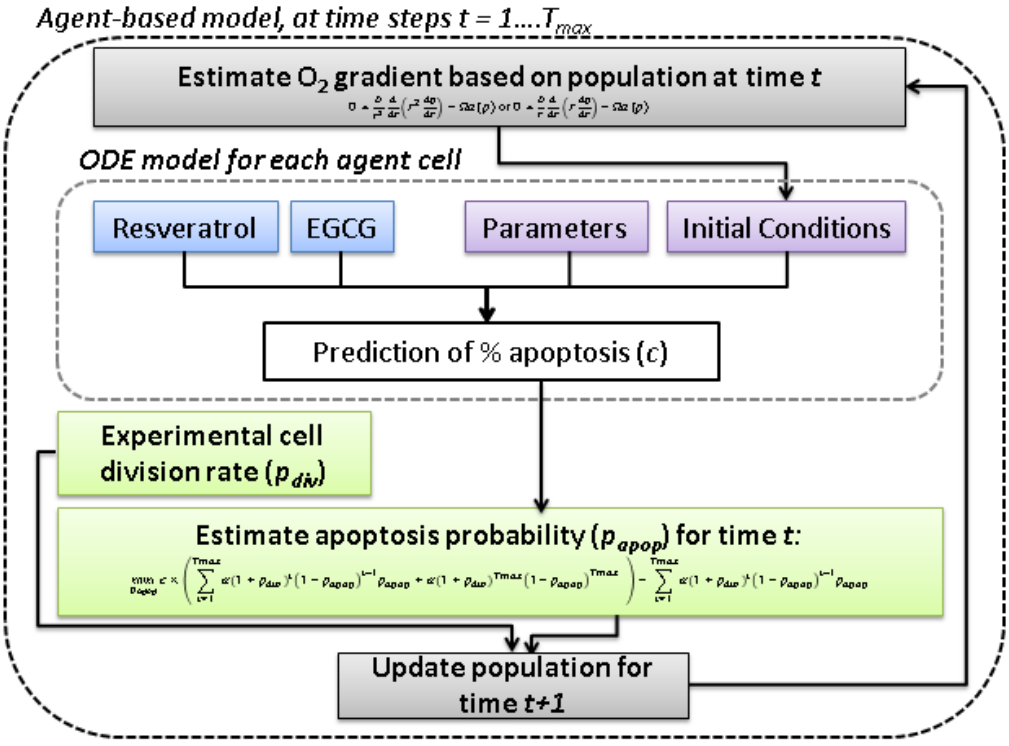


Figure 6.8: Feedback between the ODE and the ABM, based on estimation of the O_2 gradient and resulting hypoxic effects on individual cells.

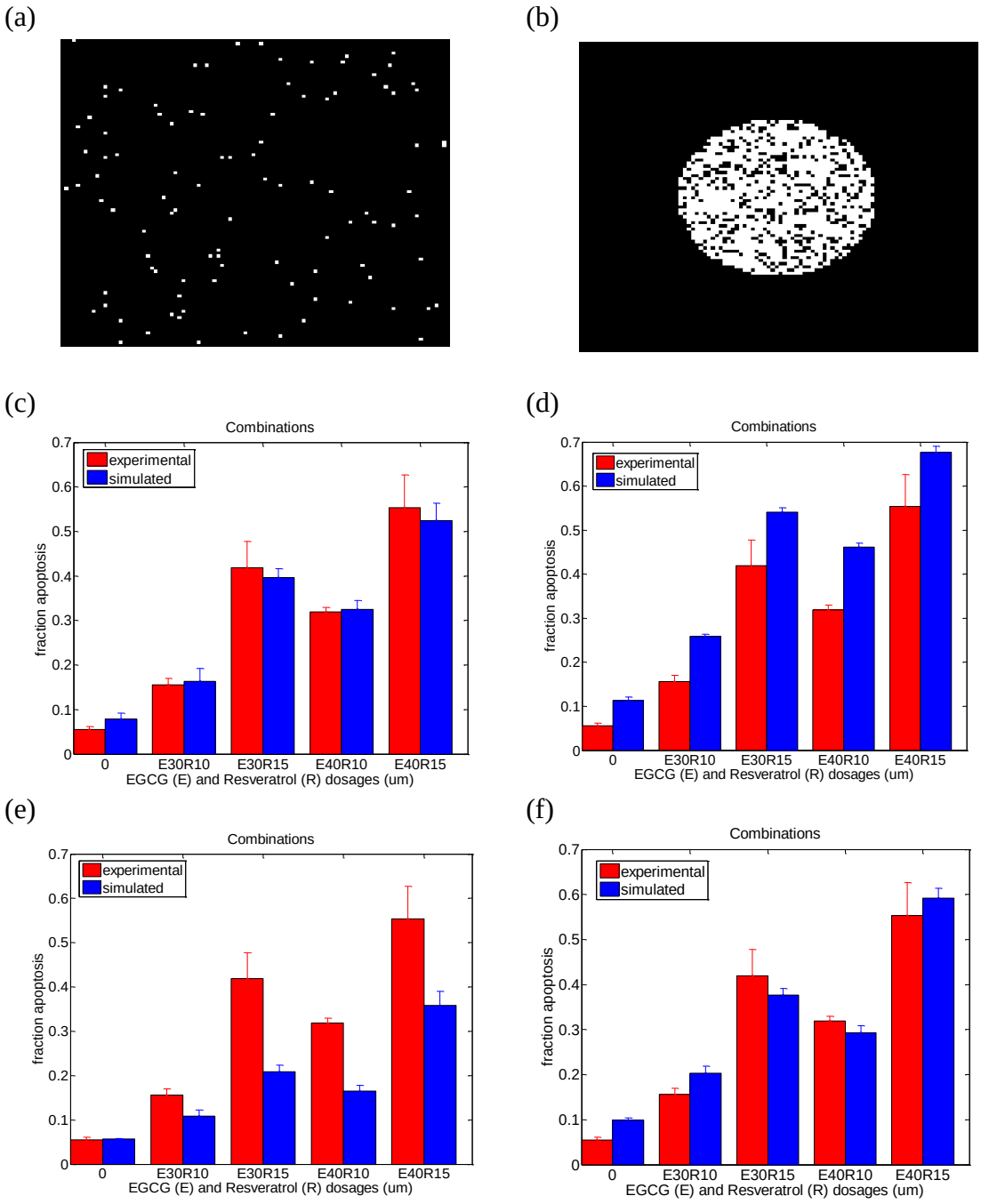


Figure 6.9: (a) uniform randomly distributed cell population; (b) spherical cell distribution. Baseline fraction apoptosis patterns with no O_2-based effects for (c) the uniform population and (d) the spherical population. Apoptosis fraction patterns with O_2-based effects on Akt activity for (e) the uniform population and (f) the spherical population.

similar to the *in vitro* cell culture environment, while the second case involved a spherical distribution of cells (Figure 6.9(b)), similar to a multicellular spheroid or avascular tumor. In the first case, an O_2 gradient was imposed by simulating the presence of a blood vessel along the vertical axis, from which O_2 diffused into the environment. In the second case, the O_2 gradient results from diffusion into the sphere from the ambient environment. In both cases, the O_2 gradients were calculated following the models and parameters in [275, 276]. The effect of hypoxia on Akt activity was modeled as a linear function of the O_2 gradient.

In the case with randomly distributed cells, the fraction apoptosis predicted by the ABM when no hypoxic effects where present matched the experimental observations for the Tu212 cell line, as shown in Figure 6.9(c). In the case with cells arranged in a sphere, the baseline fraction apoptosis, with no O_2 gradient, was predicted to be higher, as shown in Figure 6.9(d). This is reasonable because cell proliferation is restricted in the interior of the sphere due to cell crowding, while apoptosis is not. In both scenarios, the model predicted that even moderate increases in Akt activity (10-50%) could result in notable decreases in the predicted fraction of apoptotic cells. Figure 6.9(e-f) show the predicted response patterns when the O_2 gradient resulted in a maximum 20% increase in Akt activity for hypoxic cells. These predictions are supported by recent findings that the pro-apoptotic activity of curcumin, which affects many of the same pathways as EGCG and resveratrol, was inhibited by overexpression of active Akt [277].

6.5. Gene Expression Analysis

RNAseq analysis was used to implicate additional genes and processes for future expansion and refinement of the model. RNAseq data was available for four samples: no treatment, resveratrol-only, EGCG-only, and one combination (E40R15). Bioconductor packages were used to annotate sorted BAM files using the hg19 reference sequence from UCSC and to count reads. Differential expression analysis was performed using edgeR [192] to compare each treatment case against the no treatment case, with an FDR threshold of 0.05. Functional analysis of differentially expressed gene (DEG) lists was performed using DAVID [202].

The synergistic effect of EGCG and resveratrol is apparent through the pattern of DEGs observed for individual and combination treatments, as shown in Table 6.5. EGCG treatment alone led to 54 DEGs, and resveratrol alone led to 111 DEGs. The combination treatment led to a more than four-fold increase, with 466 DEGs. Among all three lists, 22 genes were in common. The DEGs associated with combination treatment also represented a larger variety of biological processes, as shown by Gene Ontology (GO) mining in Table 6.6. 19 GO terms were implicated with the combination treatment DEG list. Among these, two terms were also implicated in the EGCG DEG list, and four in the resveratrol DEG list. Key processes implicated include regulation of cell proliferation,

Table 6.5: Common DEGs for three treatment cases vs. no treatment (NT)

	NT vs. EGCG	NT vs. Resveratrol	NT vs. Combination
NT vs. EGCG	54	23	48
NT vs. Resveratrol	-	111	81
NT vs. Combination	-	-	466

Table 6.6: Significant Gene Ontology terms associated with the DEG list for the no treatment vs. combination case. Terms also associated with the *EGCG* and resveratrol DEG lists are marked.

GO:0006955~immune response	GO:0043067~regulation of programmed cell death
GO:0009611~response to wounding	*GO:0070482~response to oxygen levels*
GO:0006952~defense response	GO:0010941~regulation of cell death
GO:0010033~response to organic substance	GO:0009615~response to virus
GO:0006954~inflammatory response	GO:0051384~response to glucocorticoid stimulus
GO:0042127~regulation of cell proliferation	GO:0001893~maternal placenta development
GO:0048545~response to steroid hormone stimulus	*GO:0001666~response to hypoxia*
GO:0031960~response to corticosteroid stimulus	GO:0030595~leukocyte chemotaxis
GO:0042981~regulation of apoptosis	GO:0050900~leukocyte migration
GO:0009719~response to endogenous stimulus	

apoptosis, and cell death. Notably, the response to oxygen levels and hypoxia are also included in this list. Overall, the results of RNAseq analysis clearly indicate that the combined effect of EGCG and resveratrol is much more extreme and wide-reaching than the effects of either alone.

6.6. Discussion and Key Innovations

This chapter proposes a multi-scale ODE model for predicting and studying the combination effects of natural compounds for HNSCC chemoprevention. The model successfully predicted the combination effects of EGCG and resveratrol in three HNSCC cell lines. In addition, a multi-scale agent-based model was developed in order to couple the predictions of the ODE with feedback from spatially heterogeneous and complex cellular environments. Case studies applied these models to predict the effects of additional targeted interventions and the effects of microenvironmental hypoxia on cell population response.

Multi-scale models of cancer, which can encompass scales from the atomic to the patient level, are valuable tools for quantitatively predicting outcomes and generating testable hypotheses. Many recent models have focused on specific processes in cancer biology, such as invasion [278, 279], angiogenesis [280], and metastasis [281]. Others have focused on specific cancer types. For example, models of brain cancer and non-small cell lung cancer have related the interaction dynamics of EGFR, TGF-α, PLC-γ, and other molecules to proliferative or migratory cell phenotypes [282-284]. Models which consider drug response range from focusing on conventional chemotherapeutics [285, 286] and radiotherapy [287] to those describing the effects of targeted therapeutics, such as a tyrosine kinase inhibitor against EGFR [282], the anti-angiogenic agent endostatin [280], anti-invasive matrix metalloprotease [279], and antiandrogen therapy [288]. I distinguish the current study from prior art in two ways. First, from a biological and clinical perspective, this model focuses specifically on HNSCC and in particular on chemoprevention, not on conventional therapeutics. Second, from a modeling perspective, this model focuses on the complex effects of natural compounds, which interact with the biochemical system at multiple points, rather than targeted therapeutics.

The current results highlight several key directions for future research. First, RNAseq analysis revealed that the combination effects of EGCG and resveratrol are much more extensive than the individual effects of either. Further research into these effect patterns – particularly at the protein and metabolite levels – will yield greater insight into the mechanisms of these natural compounds, and will indicate how they can be more effectively applied in clinical settings. As additional data – particularly the kinetic parameters governing molecular-level processes – become available, the current

models can be expanded and refined to include more biological details. For example, the molecular pathway modeled here omits some intermediate mechanistic steps, such as the role of IKK in the activation of NF-kB, and that of MDM2 in inhibiting p53. These and other interactions have relevance to how EGCG and resveratrol exert their effects [289-292]. As such, they are important in interpreting model predictions, particularly target predictions. Another motivation for model expansion is that all three HNSCC cell lines modeled here are p53 mutants. As more information is gathered on the effects of mutant p53 losses and gains of function on other components of the signal transduction network, these cell line-specific effects can be incorporated into the model [293, 294]. Next, as the model predictions indicate, the response to EGCG and resveratrol may be dampened in more complex *in vitro* and *in vivo* settings, due to microenvironmental and other factors. This is a multi-faceted challenge, and potential solutions include combination with other natural compounds or targeted agents, as suggested by the first case study, as well as the development of effective drug delivery and cell-targeting strategies.

In the long term, as these challenges are addressed, models for predicting natural compound chemoprevention response could become part of a personalized treatment planning system for HNSCC patients. Such models could take into account patient-specific clinical data and –omic expression signatures in order to predict regimens of effective, non-toxic chemoprevention adjuvants. The current modeling study provides the groundwork for the development of such a system, with the overall goal of preventing recurrence, SPT development, and metastasis, and improving HNSCC patient outcomes.

The Key Innovations of this chapter are:

- Developed first multi-scale models for predicting the combination effects of natural compounds in HNSCC
- Tested multi-scale ODE model on dose response data from three HNSCC cell lines, and extended it to generate a multi-scale ABM
- Demonstrated application of ODE and ABM models for target prediction and prediction of response in complex environments, respectively

CHAPTER 7

CONCLUSION

The concrete goals of this dissertation were to develop mathematical modeling tools for mining –omic datasets and for the analysis of biological system behavior in the context of HNSCC. The specific technical achievements of this dissertation corresponding to the three research objectives are:

1. Development and validation of mathematical modeling tools for knowledge-driven exploratory data mining of transcriptomic, proteomic, and metabolomic datasets, in terms of both explicit (hypergeometric similarity measures, DetectTLC) and implicit (DetectTLC) similarity-based analysis
2. Construction of predictive models using integrated analyses between –omic levels to discriminate between early and advanced HNSCC, and within –omic levels for developing robust predictive models applicable to early disease detection
3. Development and validation of integrated molecular- and cellular-level ordinary differential equation model for predicting the response to natural compound adjuvants in HNSCC cell populations, and extension to an agent-based model for prediction under different microenvironmental conditions

7.1. Concrete Innovation Deliverables

The key innovations of this dissertation, as noted at the closing of each chapter, are summarized below:

- (Chapter 2) Development of binary hypergeometric similarity measure using Fisher's exact test

- (Chapter 2) Development of multivariate hypergeometric similarity measure using the Fisher-Freeman-Halton test
- (Chapter 2) Development of a piecewise approximation algorithm to facilitate application of the multivariate hypergeometric similarity measure to high-dimensional data vectors
- (Chapter 2) Implementation on two HNSCC (transcriptomic and proteomic) and one non-HNSCC (MSI, lipidomic) datasets indicates that proposed multivariate hypergeometric similarity measure makes relevant selections not identified by other similarity measures
- (Chapter 3) Development of the first analytical pipelines using quantitative image features for identifying *m/z* images containing spot-like regions in MSI data
- (Chapter 3) Design, implementation, and validation of the first software tool, DetectTLC, for enabling and accelerating TLC-MSI studies in metabolomics by automatically finding mixture components of potential interest in TLC-MSI datasets
- (Chapter 4) Performed the first supervised modeling study for modeling progression in HNSCC by integrating both proteomic and transcriptomic data
- (Chapter 4) Developed between-omic level integrated ensemble models with significant improvement in performance for predicting HNSCC pathological stage
- (Chapter 5) Performed within–omic level integrative modeling study using microarray and RNAseq data for detection of HNSCC
- (Chapter 5) Translated ensemble models developed for discriminating between HNSCC and paired normal cases to the problem of early HNSCC detection

- (Chapter 5) Implemented tool to facilitate model translation and use of ensemble transcriptomic models in the HNSCC research community
- (Chapter 6) Developed first multi-scale models for predicting the combination effects of natural compounds in HNSCC
- (Chapter 6) Tested multi-scale ODE model on dose response data from three HNSCC cell lines, and extended it to generate a multi-scale ABM
- (Chapter 6) Demonstrated application of ODE and ABM models for target prediction and prediction of response in complex environments, respectively

Figure 7.1 demonstrates how these deliverables map to both clinical challenges and technical challenges in HNSCC research.

		Clinical Challenges		
		Biomarker Discovery	Early Detection and Diagnosis	Intervention Strategies
Technical Challenges	Knowledge-Driven Mining	Similarity Measures (Chapter 2) DetectTLC (Chapter 3)		
	Knowledge Integration	Transcriptomic and Proteomic Models (Chapters 4 and 5)		
	Multi-scale System Analysis			Chemoprevention Response Models (Chapter 6)

Figure 7.1: Mapping Deliverables to Clinical and Technical Challenges

7.2. Concrete Publication Deliverables

The section provides a comprehensive list of publications completed during my years as a Ph.D. student. Those which contribute directly to this dissertation are highlighted in Table 7.1.

Table 7.1: Overview of publications related to dissertation

Specific Aim	Sub-Aim	Citation	Publication Type	Publication Status
1	Binary hypergeometric similarity measure	[C1]	Conference paper	Published
	Multivariate hypergeometric similarity measure	[J2]	Journal paper	Published
	DetectTLC	[J6]	Journal paper	In preparation
2	Early stage vs. normal: microarray and RNAseq data	[C6]	Conference paper	Accepted for Publication
	Early vs. advanced stage: gene and protein expression data	[C4] [J4]	Conference paper Journal paper	Published Under review
3	Single-scale, cellular-level cancer model	[C3]	Conference paper	Published
	Multi-scale chemoprevention model	[J7]	Journal paper	In preparation

Published or Accepted for Publication

Journal Papers

[J1] Quo CF, Kaddi C, Phan JH, Zollanvari A, Xu M, Wang MD and Alterovitz G (2012) Reverse engineering biomolecular systems using –omic data: challenges, progress and opportunities. Briefings in Bioinformatics, 13: pp. 430-445

[J2] Kaddi CD, Parry RM and Wang MD (2013) Multivariate hypergeometric similarity measure. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(6): 1505-1516

[J3] Kaddi CD, Phan JH and Wang MD (2013) Computational Nanomedicine: Modeling of Nanoparticle-Mediated Hyperthermal Cancer Therapy. Nanomedicine, 8(8): 1232-1233

Book Chapters

[B1] Kaddi CD and Wang MD (2015) Computational methods for mass spectrometry imaging: challenges, progress, and opportunities.

Refereed (Peer-Reviewed) Conference Papers

[C1] Kaddi C, Parry RM and Wang MD (2011) Hypergeometric similarity measure for spatial analysis in tissue imaging mass spectrometry. Proceedings of the IEEE International Conference in Bioinformatics & Biomedicine (BIBM), pp. 604-607

[C2] Kells K, Kong KY, White WB, Kaddi C and Wang MD (2012) LED light source for fluorescence endoscopy using quantum dots. Proceedings of IEEE EMBS-Point-of-Care Healthcare Technologies (POCHT) 2013, pp. 9-12

[C3] Kaddi CD and Wang MD (2013) Mathematical model of the effect of inter-cellular cooperative interactions in cancer during drug therapy. Proceedings of Oak Ridge National Lab BSEC 2013, pp. 1-4.

[C4] Kaddi CD and Wang MD (2014) Models for predicting stage in head and neck squamous cell carcinoma using proteomic data. Proceedings of IEEE Engineering in Medicine and Biology Society 2014, pp. 5216-5219

[C5] Sarkari S, Kaddi CD, Bennett RV, Fernández FM and Wang MD (2014) Comparison of clustering pipelines for the analysis of mass spectrometry imaging data. Proceedings of IEEE Engineering in Medicine and Biology Society 2014, pp. 4771-4

[C6] Kaddi CD and Wang MD (2015) Developing Robust Predictive Models for Head and Neck Cancer across Microarray and RNAseq Gene Expression Data. Proceedings of ACM-BCB 2015.

Manuscripts under Review

Journal Papers

[J4] Kaddi CD and Wang MD (2015) Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma using Proteomic and Transcriptomic Data.

[J5] White WB, Oh KJ, Kaddi CD, Okusanya OT, Mohs A, Nie S, Wang MD, Singhal S. (2015) Fluorescent Imaging System for Intraoperative Molecular Imaging during Endoscopic Surgery.

Manuscripts in Preparation

Journal Papers

[J6] Kaddi CD, Bennett RV, Paine MRL, Weber AL, Fernández FM and Wang MD (2015) DetectTLC: A tool for semi-automated region of interest identification of reaction mixture separations on the basis of DESI images.

[J7] Kaddi CD, Amin ARMR, Chen Z, Shin DM, and Wang MD (2015) Multi-scale Modeling to Predict Combination Effects of Natural Compounds for Chemoprevention in Head and Neck Squamous Cell Carcinoma.

[J8] Moffitt RA, Wang X, Hurwitz SJ, Kaddi CD, Fox BM, Sizemore EK, Shin DM, Chen ZG and Wang MD (2015) Combined tumor growth and pharmacodynamic model for studying spatial dynamics and response to paclitaxel.

[J9] Wu PY, [author order not finalized: Cheng CW, Hoffman R, Kaddi CD, Venugopalan J], and Wang MD (2015) Towards the new era of evidence based medicine and healthcare using big data. IEEE Transactions on Biomedical Engineering.

7.3. Directions for Future Research and Concluding Remarks

The models and tools developed in this dissertation are complete and fully functional. However, the critical final step in a research project is to recognize potential future applications and extensions of the current work. Some specific avenues for further inquiry were mentioned in the discussion sections concluding each chapter. Here, I elaborate on opportunities in two dimensions: (1) basic and translational research in HNSCC and (2) the design and development of novel mathematical models.

7.3.1. Basic and Translational Research in HNSCC

One of the major goals of Big Data research in biomedicine is biomarker identification, for specific and practical applications like early diagnosis, patient stratification, and prediction of treatment response. Applying the modeling infrastructure developed in this dissertation to new, more comprehensive –omic datasets can greatly facilitate these tasks:

Early Disease Detection with Transcriptomic, Proteomic, and Metabolomic Data

Because of the differences in HNSCC outcomes according to the stage at which the disease is detected, molecular marker-based systems for early diagnosis of HNSCC could have a large clinical impact. This is particularly important for disease subsites for which early disease symptoms may be limited, like the oropharynx, or for which symptoms may be misattributed, as in the oral cavity [295, 296]. Overlap between the transcriptomic features identified in Chapter 5 and validated salivary mRNA markers for detecting oral cancer [235] is encouraging, and establishes the stage for clinical validation of other transcriptomic features highlighted through the models developed in

this research. In addition, Chapter 4 demonstrated that the integration of proteomic and transcriptomic data can assist in stage prediction. This between-omic level integration may also assist in the problem of early diagnosis, once proteomic data for matched early and normal patient samples becomes available. Incorporating metabolomic data into these models is also worthy of investigation.

Integration of Mass Spectrometry Imaging and Natural Compound Chemoprevention Research

The potential of MSI in HNSCC research is immense, and informative model-driven experiments could immediately follow the acquisition of MSI datasets from HNSCC samples. Possible experimental settings include tumors, xenografts, or spheroidal cultures. MSI enables combined molecular and spatial analysis. This could be particularly informative following treatment with bio-active natural compounds. Chapter 6 developed a multi-scale ABM, which is a spatial model that could be applied to understand and predict spatially heterogeneous molecular expression and cellular-level response patterns observed in MSI data. In addition, the similarity measures developed in Chapter 2 could be applied to assess spatial molecular expression patterns across tissue regions (i.e., tumor, marginal, and surrounding normal), as well as among regions showing different degrees of response to administered natural compounds.

Integration of TLC-MSI and Chemoprevention Research

Another key direction for research is the investigation of lipid and metabolite profiles in HNSCC. The natural compounds being investigated for HNSCC

chemoprevention affect multiple biochemical entities, both directly and indirectly. For example, recent studies have indicated the importance of lipid rafts to the effects of EGCG and resveratrol on downstream signaling [297, 298]. The DetectTLC system developed in Chapter 3 provides a computational framework for investigating the effects of natural compounds on lipids and metabolites, and hence for obtaining a better understanding of their mechanisms of action.

Applications of Similarity Measure in Biomedical Image Analysis

The multivariate hypergeometric similarity measure introduced in Chapter 2 may also be applied to other data types in addition to molecular expression –omic data. One potential application is with wavelets, which are used for signal and image processing in many different application areas. For example, in radiomics, features from wavelet-transformed X-ray computed tomography (CT) images were among a set of image features used for prognostic prediction in HNSCC [299]. In this dissertation, data similarity was assessed based on binned expression levels. Future research could investigate the performance of the proposed similarity measure for comparing images and data in terms of wavelet features.

Another potential application is in tissue imaging using quantum dots (QDs). QDs are fluorescent nanoparticles that can be conjugated to antibodies for targeted visualization of molecular and cellular targets [300, 301]. Compared to fluorescent dyes, QDs are advantageous because of their long-lasting fluorescence, target specificity, and multiplexing capabilities. Clinical applications are currently not possible due to the issue of heavy metal toxicity. However, this may change in the future as an initial trial in non-

human primates showed no toxic effects during the first 90 days after administration [302]. However, QDs remain valuable for research applications. In recent HNSCC research in particular, QDs have been used to investigate the association of aldehyde dehydrogenease 1 with lymph node metastasis [303] and that of caveolin-1 with clinical stage, histological grade, and cancer development [304]. Additionally, QD-based immunohistofluorescence was observed to have greater sensitivity and objectivity compared to immunohistochemistry in an HNSCC application [303]. The multivariate hypergeometric similarity measure developed in Chapter 2 provides a framework for comparing fluorescent images, particularly when using multiplexed QDs. In one scenario, each bin (i.e., the class to which a pixel is assigned) in the reference and query images could represent a QD expression intensity level, enabling two QDs to be compared. In another scenario, each bin could represent an *n*-dimensional vector of expression levels for a group of *n* QDs, thereby enabling similarity assessment of multiplexed QD data.

7.3.2. Design and Development of Novel Mathematical Models

The design and development of new modeling techniques can assist in many biomedical research areas, including HNSCC. In the following section, I identify key directions for building upon and extending the modeling infrastructure developed in this dissertation:

Time-Series Analysis

Chapters 4 and 5 have demonstrated the development of integrated –omics models for predicting clinically relevant endpoints. However, all data currently used is static, obtained at a single time-point. Metabolomics data in particular is highly dynamic,

and has shown potential not only for early diagnosis but also for monitoring of disease status [160, 161, 305]. This reveals an opportunity to develop predictive models which utilize time-series –omic data to track patient risk and prognosis over time. Such models could help clinicians monitor the status of their patients and could serve to improve personalized medicine.

Ensemble Model Construction

Chapters 4 and 5 have also demonstrated how integrated ensemble modeling techniques can improve prediction performance. However, selecting the most appropriate ensemble from among all possible ensembles can be challenging, especially to users from non-computational backgrounds. A second predictive modeling layer could help to address this issue. For example, the input to such a model could be a new dataset of interest. Given historical performance patterns observed for other datasets across various models (as in Chapter 5), a similarity-based approach (as in Chapter 2) could be used to compare dataset properties, and thereby identify corresponding ensemble constructions that are likely to yield good performance on the new dataset.

Systems Models

Chapter 6 developed multi-scale models for predicting the response to natural compounds. While my previous HNSCC system model used parameters based on experimental data [245], parameter estimation was necessary for the more complex multi-scale model. Thus, a key direction for improvement is parameterization of molecular- and cellular-level processes based on experimental measurements. This could be a dynamic

process in itself. For example, when a new cell line or patient becomes available, such a model could accept accompanying time-series data and automatically extract relevant parameters. If no data is available, a similarity-based approach (as in Chapter 2) may be used to identify the most relevant previously examined samples, and adapt experimental parameters from them. This would reduce the number of parameters to be estimated via error minimization. Next, another direction for improvement is to expand the list of processes associated with the multi-scale ABM to include movement and mechanical interactions. This would enable more realistic prediction of cellular-level behaviors, and investigation of cancer-relevant processes like invasion and metastasis. Lastly, nanoparticles – including gold nanoparticles – have been proposed as delivery vehicles to improve bioavailability of natural compounds [54, 306]. In addition, there has been evidence that the combination of natural compounds and hyperthermia can have synergistic effects [307]. Thus, the multi-scale ABM could be integrated with models for nanoparticle-based drug delivery and hyperthermia for cancer treatment, a subject which I have previously reviewed [308].

7.3.3. Concluding Remarks

In this dissertation, I have developed a suite of mathematical modeling tools to address key challenges in HNSCC research. It includes mathematical models for data mining and system dynamics that have been successfully applied to investigate HNSCC molecular characteristics, progression, and chemoprevention response. In the preceding sections, I have also discussed several potential seeds for future investigations, building upon this work. Overall, this dissertation contributes to the research space by accelerating

and enabling the application of large –omics datasets to basic and translational cancer research.

REFERENCES

[1] C. R. Leemans, B. J. Braakhuis, and R. H. Brakenhoff, "The molecular biology of head and neck cancer," *Nat Rev Cancer,* vol. 11, pp. 9-22, Jan 2011.

[2] A. Matta and R. Ralhan, "Overview of current and future biologically based targeted therapies in head and neck squamous cell carcinoma," *Head Neck Oncol,* vol. 1, p. 6, 2009.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA Cancer J Clin,* vol. 65, pp. 5-29, Jan 2015.

[4] K. Strong, C. Mathers, J. Epping-Jordan, S. Resnikoff, and A. Ullrich, "Preventing cancer through tobacco and infection control: how many lives can we save in the next 10 years?," *Eur J Cancer Prev,* vol. 17, pp. 153-61, Apr 2008.

[5] M. Hashibe, P. Brennan, S.-c. Chuang, S. Boccia, X. Castellsague, C. Chen, *et al.*, "Interaction between Tobacco and Alcohol Use and the Risk of Head and Neck Cancer: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium," *Cancer Epidemiology Biomarkers & Prevention,* vol. 18, pp. 541-550, February 1, 2009 2009.

[6] Y. J. Chen, J. T. Chang, C. T. Liao, H. M. Wang, T. C. Yen, C. C. Chiu, *et al.*, "Head and neck cancer in the betel quid chewing area: recent advances in molecular carcinogenesis," *Cancer Sci,* vol. 99, pp. 1507-14, Aug 2008.

[7] S. S. Muttagi, P. Chaturvedi, R. Gaikwad, B. Singh, and P. Pawar, "Head and neck squamous cell carcinoma in chronic areca nut chewing Indian women: Case series and review of literature," *Indian Journal of Medical and Paediatric Oncology : Official Journal of Indian Society of Medical & Paediatric Oncology,* vol. 33, pp. 32-35, Jan-Mar 2012.

[8] C. Dansky Ullmann, L. C. Harlan, V. L. Shavers, and J. L. Stevens, "A population-based study of therapy and survival for patients with head and neck cancer treated in the community," *Cancer,* vol. 118, pp. 4452-4461, 2012.

[9] J. Bernier, C. Domenge, M. Ozsahin, K. Matuszewska, J. L. Lefebvre, R. H. Greiner, *et al.*, "Postoperative irradiation with or without concomitant chemotherapy for locally advanced head and neck cancer," *N Engl J Med,* vol. 350, pp. 1945-52, May 6 2004.

[10] J. S. Cooper, T. F. Pajak, A. A. Forastiere, J. Jacobs, B. H. Campbell, S. B. Saxman, *et al.*, "Postoperative concurrent radiotherapy and chemotherapy for high-risk squamous-cell carcinoma of the head and neck," *N Engl J Med,* vol. 350, pp. 1937-44, May 6 2004.

[11] D. Kiprian, A. Kawecki, A. Jarzabski, W. Michalski, and B. Pawlowska-Sendulka, "The results and toxicity of organ preservation treatment for locoregionally advanced laryngeal and hypopharyngeal cancer," *Otolaryngol Pol,* vol. 65, pp. 363-368, Sep 2011.

[12] M. R. Posner, D. M. Hershock, C. R. Blajman, E. Mickiewicz, E. Winquist, V. Gorbounova, *et al.*, "Cisplatin and fluorouracil alone or with docetaxel in head and neck cancer," *N Engl J Med,* vol. 357, pp. 1705-15, Oct 25 2007.

[13] D. Rades, T. Meyners, N. Kazic, A. Bajrovic, V. Rudat, and S. E. Schild, "Comparison of radiochemotherapy alone to surgery plus radio(chemo)therapy for non-metastatic stage III/IV squamous cell carcinoma of the head and neck: A matched-pair analysis," *Strahlenther Onkol,* vol. 187, pp. 541-7, Sep 2011.
[14] D. Pulte and H. Brenner, "Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis," *Oncologist,* vol. 15, pp. 994-1001, 2010.
[15] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in Scientific Data Infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, 2013, pp. 48-55.
[16] R. Higdon, W. Haynes, L. Stanberry, E. Stewart, G. Yandl, C. Howard, *et al.*, "Unraveling the Complexities of Life Sciences Data," *Big Data,* vol. 1, pp. 42-50, 2013/03/01 2012.
[17] S. M. Hanash, C. S. Baik, and O. Kallioniemi, "Emerging molecular biomarkers--blood-based strategies to detect and monitor cancer," *Nat Rev Clin Oncol,* vol. 8, pp. 142-50, Mar 2011.
[18] C. L. Sawyers, "The cancer biomarker problem," *Nature,* vol. 452, pp. 548-552, 04/03/print 2008.
[19] K. K. W. To, "MicroRNA: a prognostic biomarker and a possible druggable target for circumventing multidrug resistance in cancer chemotherapy," *Journal of Biomedical Science,* vol. 20, pp. 99-99, 12/20
09/15/received
12/16/accepted 2013.
[20] N. S. Nagaraj, "Evolving 'omics' technologies for diagnostics of head and neck cancer," *Briefings in Functional Genomics & Proteomics,* vol. 8, pp. 49-59, January 1, 2009 2009.
[21] L. Sepiashvili, J. P. Bruce, S. H. Huang, B. O'Sullivan, F. F. Liu, and T. Kislinger, "Novel insights into head and neck cancer using next-generation "omic" technologies," *Cancer Res,* vol. 75, pp. 480-6, Feb 1 2015.
[22] E. A. Vucic, K. L. Thu, K. Robison, L. A. Rybaczyk, R. Chari, C. E. Alvarez, *et al.*, "Translating cancer 'omics' to improved outcomes," *Genome Research,* vol. 22, pp. 188-195, 2012.
[23] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science,* vol. 339, pp. 1546-1558, March 29, 2013 2013.
[24] L. W. Ellisen and D. A. Haber, "Basic Principles of Cancer Genetics," in *Principles of Clinical Cancer Genetics: A Handbook from the Massachusetts GEneral Hospital*, D. C. Chung and D. A. Haber, Eds., ed: Springer Science+Business Media, 2010, pp. 1-22.
[25] F. Bunz, *Principles of Cancer Genetics*: Springer Science+Business Media B.V., 2008.
[26] R. Weinberg, *The Biology of Cancer*, 2nd ed. vol. Garland Science, Taylor & Francis Group, LLC, 2014.
[27] D. Hanahan and Robert A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell,* vol. 144, pp. 646-674, 3/4/ 2011.

[28] M. Parfenov, C. S. Pedamallu, N. Gehlenborg, S. S. Freeman, L. Danilova, C. A. Bristow, *et al.*, "Characterization of HPV and host genome interactions in primary head and neck cancers," *Proc Natl Acad Sci U S A,* vol. 111, pp. 15544-9, Oct 28 2014.

[29] N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, *et al.*, "The mutational landscape of head and neck squamous cell carcinoma," *Science,* vol. 333, pp. 1157-60, Aug 26 2011.

[30] TCGA, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature,* vol. 517, pp. 576-582, 01/29/print 2015.

[31] M. Baker, "Metabolomics: from small molecules to big ideas," *Nat Meth,* vol. 8, pp. 117-121, 02//print 2011.

[32] R. Smith, A. D. Mathis, D. Ventura, and J. T. Prince, "Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view," *BMC Bioinformatics,* vol. 15 Suppl 7, p. S9, 2014.

[33] A. Matta, R. Ralhan, L. V. DeSouza, and K. W. Siu, "Mass spectrometry-based clinical proteomics: head-and-neck cancer biomarkers and drug-targets discovery," *Mass Spectrom Rev,* vol. 29, pp. 945-61, Nov-Dec 2010.

[34] S. Tiziani, V. Lopes, and U. L. Gunther, "Early stage diagnosis of oral cancer using 1H NMR-based metabolomics," *Neoplasia,* vol. 11, pp. 269-76, 4p following 269, Mar 2009.

[35] J. Xu, Y. Chen, R. Zhang, Y. Song, J. Cao, N. Bi, *et al.*, "Global and Targeted Metabolomics of Esophageal Squamous Cell Carcinoma Discovers Potential Diagnostic and Therapeutic Biomarkers," *Molecular & Cellular Proteomics,* vol. 12, pp. 1306-1318, May 1, 2013 2013.

[36] K. Yonezawa, S. Nishiumii, J. Kitamoto-Matusda, T. Fujita, K. Morimoto, D. Yamashita, *et al.*, "Serum and Tissue Metabolomics of Head and Neck Cancer," *Cancer Genomics - Proteomics,* vol. 10, pp. 233-238, September 1, 2013 2013.

[37] M. J. Worsham, "Identifying the risk factors for late-stage head and neck cancer," *Expert Rev Anticancer Ther,* vol. 11, pp. 1321-5, Sep 2011.

[38] N. Kondoh, S. Ohkura, M. Arai, A. Hada, T. Ishikawa, Y. Yamazaki, *et al.*, "Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma," *Oral Oncol,* vol. 43, pp. 455-62, May 2007.

[39] P. Saintigny, L. Zhang, Y.-H. Fan, A. K. El-Naggar, V. A. Papadimitrakopoulou, L. Feng, *et al.*, "Gene Expression Profiling Predicts the Development of Oral Cancer," *Cancer Prevention Research,* vol. 4, pp. 218-229, February 1, 2011 2011.

[40] K. Chen, R. Sawhney, M. Khan, M. S. Benninger, Z. Hou, S. Sethi, *et al.*, "Methylation of multiple genes as diagnostic and therapeutic markers in primary head and neck squamous cell carcinoma," *Arch Otolaryngol Head Neck Surg,* vol. 133, pp. 1131-8, Nov 2007.

[41] M. A. Ginos, G. P. Page, B. S. Michalowicz, K. J. Patel, S. E. Volker, S. E. Pambuccian, *et al.*, "Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck," *Cancer Res,* vol. 64, pp. 55-63, Jan 1 2004.

[42] L. Lo Russo, M. Papale, D. Perrone, E. Ranieri, C. Rubini, G. Giannatempo, *et al.*, "Salivary Proteomic Signatures of Oral Squamous Cell Carcinoma," *European Journal of Inflammation,* vol. 10, pp. 61-70, 2012.

[43] E. Mendez, C. Cheng, D. G. Farwell, S. Ricks, S. N. Agoff, N. D. Futran, *et al.*, "Transcriptional expression profiles of oral squamous cell carcinomas," *Cancer,* vol. 95, pp. 1482-94, Oct 1 2002.

[44] M. Pietrowska, J. Polanska, R. Suwinski, M. Widel, T. Rutkowski, M. Marczyk, *et al.*, "Comparison of peptide cancer signatures identified by mass spectrometry in serum of patients with head and neck, lung and colorectal cancers: association with tumor progression," *Int J Oncol,* vol. 40, pp. 148-56, Jan 2012.

[45] O. Saglam, V. Shah, and M. J. Worsham, "Molecular differentiation of early and late stage laryngeal squamous cell carcinoma: an exploratory analysis," *Diagn Mol Pathol,* vol. 16, pp. 218-21, Dec 2007.

[46] C. E. Schmalbach, D. B. Chepeha, T. J. Giordano, M. A. Rubin, T. N. Teknos, C. R. Bradford, *et al.*, "Molecular profiling and the identification of genes associated with metastatic oral cavity/pharynx squamous cell carcinoma," *Arch Otolaryngol Head Neck Surg,* vol. 130, pp. 295-302, Mar 2004.

[47] T. Maier, M. Güell, and L. Serrano, "Correlation of mRNA and protein in complex biological samples," *FEBS Letters,* vol. 583, pp. 3966-3973, 12/17/ 2009.

[48] B. Zahorowska, P. Crowe, and J.-L. Yang, "Combined therapies for cancer: a review of EGFR-targeted monotherapy and combination treatment with other drugs," *Journal of Cancer Research and Clinical Oncology,* vol. 135, pp. 1137-1148, 2009/09/01 2009.

[49] E. Rosenthal, T. Chung, W. Carroll, L. Clemons, R. Desmond, and L. Nabell, "Assessment of Erlotinib as Adjuvant Chemoprevention in High-Risk Head and Neck Cancer Patients," *Annals of Surgical Oncology,* vol. 21, pp. 4263-4269, 2014/12/01 2014.

[50] N. F. Saba, S. J. Hurwitz, S. A. Kono, C. S. Yang, Y. Zhao, Z. Chen, *et al.*, "Chemoprevention of head and neck cancer with celecoxib and erlotinib: results of a phase ib and pharmacokinetic study," *Cancer Prev Res (Phila),* vol. 7, pp. 283-91, Mar 2014.

[51] M. A. Rahman, A. R. M. R. Amin, and D. M. Shin, "Chemopreventive Potential of Natural Compounds in Head and Neck Cancer," *Nutrition and Cancer,* vol. 62, pp. 973-987, 2010/09/23 2010.

[52] N. F. Saba, M. Haigentz Jr, J. B. Vermorken, P. Strojan, P. Bossi, A. Rinaldo, *et al.*, "Prevention of head and neck squamous cell carcinoma: Removing the “chemo” from “chemoprevention”," *Oral Oncology,* vol. 51, pp. 112-118, 2// 2015.

[53] A. R. M. R. Amin, M. A. Rahman, D. Wang, F. R. Khuri, Z. G. Chen, and D. M. Shin, "Abstract 1588: Synergistic apoptosis by combination of natural compound EGCG and resveratrol in head and neck cancer: Potential role for AKT-dependent signaling," *Cancer Research,* vol. 72, p. 1588, April 15, 2012 2012.

[54] D. J. Bharali, I. A. Siddiqui, V. M. Adhami, J. C. Chamcheu, A. M. Aldahmash, H. Mukhtar, *et al.*, "Nanoparticle Delivery of Natural Products in the Prevention

and Treatment of Cancers: Current Status and Future Prospects," *Cancers,* vol. 3, pp. 4024-4045, 2011.

[55] S. Gao and M. Hu, "Bioavailability Challenges Associated with Development of Anti-Cancer Phenolics," *Mini reviews in medicinal chemistry,* vol. 10, pp. 550-567, 2010.

[56] R. Manikandan, M. Beulaja, C. Arulvasu, S. Sellamuthu, D. Dinesh, D. Prabhu, *et al.*, "Synergistic anticancer activity of curcumin and catechin: An in vitro study using human cancer cell lines," *Microscopy Research and Technique,* vol. 75, pp. 112-116, 2012.

[57] R. P. Araujo and D. L. S. McElwain, "A history of the study of solid tumour growth: The contribution of mathematical modelling," *Bulletin of Mathematical Biology,* vol. 66, pp. 1039-1091, 2004/09/01 2004.

[58] H. M. Byrne, "Dissecting cancer through mathematics: from the cell to the animal model," *Nat Rev Cancer,* vol. 10, pp. 221-230, 03//print 2010.

[59] A. R. A. Anderson and V. Quaranta, "Integrative mathematical oncology," *Nat Rev Cancer,* vol. 8, pp. 227-234, 03//print 2008.

[60] T. S. Deisboeck, Z. Wang, P. Macklin, and V. Cristini, "Multiscale Cancer Modeling," *Annual review of biomedical engineering,* vol. 13, pp. 10.1146/annurev-bioeng-071910-124729, 2011.

[61] R. G. Sadygov and J. R. Yates, "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases," *Analytical Chemistry,* vol. 75, pp. 3792-3798, 2003.

[62] I. M. Shih, K. Nakayama, G. Wu, N. Nakayama, J. H. Zhang, and T. L. Wang, "Amplification of the ch19p13.2 NACC1 locus in ovarian high-grade serous carcinoma," *Modern Pathology,* vol. 24, pp. 638-645, 2011.

[63] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, *et al.*, "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches," *PLoS ONE,* vol. 6, Mar 2011.

[64] V. Megalooikonomou, M. Barnathan, D. Kontos, P. R. Bakic, and A. D. A. Maidment, "A Representation and Classification Scheme for Tree-Like Structures in Medical Images: Analyzing the Branching Pattern of Ductal Trees in X-ray Galactograms," *IEEE Transactions on Medical Imaging,* vol. 28, pp. 487-493, 2009.

[65] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, *et al.*, "Predicting human brain activity associated with the meanings of nouns," *Science,* vol. 320, pp. 1191-1195, 2008.

[66] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining Drug and Gene Similarity Measures for Drug-Target Elucidation," *Journal of Computational Biology,* vol. 18, pp. 133-145, 2011.

[67] G. Yona, W. Dirks, S. Rahman, and D. M. Lin, "Effective similarity measures for expression profiles," *Bioinformatics,* vol. 22, pp. 1616-1622, 2006.

[68] S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematical Models and Methods in Applied Sciences,* vol. 1, pp. 300-307, 2007.

[69] X. B. Li and R. C. Dubes, "A Probabilistic Measure of Similarity For Binary Data in Pattern-Recognition," *Pattern Recognition,* vol. 22, pp. 397-409, 1989.
[70] L. MacAleese, J. Stauber, and R. M. A. Heeren, "Perspectives for imaging mass spectrometry in the proteomics landscape," *Proteomics,* vol. 9, pp. 819-834, Feb 2009.
[71] N. Goto-Inoue, T. Hayasaka, N. Zaima, and M. Setou, "Imaging mass spectrometry for lipidomics," *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids,* vol. 1811, pp. 961-969, Nov 2011.
[72] Y. Sugiura and M. Setou, "Imaging Mass Spectrometry for Visualization of Drug and Endogenous Metabolite Distribution: Toward In Situ Pharmacometabolomes," *Journal of Neuroimmune Pharmacology,* vol. 5, pp. 31-43, Mar 2010.
[73] S. A. Patel, A. Barnes, N. Loftus, R. Martin, P. Sloan, N. Thakker, *et al.*, "Imaging mass spectrometry using chemical inkjet printing reveals differential protein expression in human oral squamous cell carcinoma," *Analyst,* vol. 134, pp. 301-307, 2009.
[74] N. Y. R. Agar, J. G. Malcolm, V. Mohan, H. W. Yang, M. D. Johnson, A. Tannenbaum, *et al.*, "Imaging of Meningioma Progression by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry," *Analytical Chemistry,* vol. 82, pp. 2621-2625, Apr 2010.
[75] S. Rauser, C. Marquardt, B. Balluff, S. O. Deininger, C. Albers, E. Belau, *et al.*, "Classification of HER2 Receptor Status in Breast Cancer Tissues by MALDI Imaging Mass Spectrometry," *Journal of Proteome Research,* vol. 9, pp. 1854-1863, Apr 2010.
[76] S. R. Oppenheimer, D. M. Mi, M. E. Sanders, and R. M. Caprioli, "Molecular Analysis of Tumor Margins by MALDI Mass Spectrometry in Renal Carcinoma," *Journal of Proteome Research,* vol. 9, pp. 2182-2190, May 2010.
[77] Y. Morita, K. Ikegami, N. Goto-Inoue, T. Hayasaka, N. Zaima, H. Tanaka, *et al.*, "Imaging mass spectrometry of gastric carcinoma in formalin-fixed paraffin-embedded tissue microarray," *Cancer Science,* vol. 101, pp. 267-273, Jan 2010.
[78] L. H. Cazares, D. Troyer, S. Mendrinos, R. A. Lance, J. O. Nyalwidhe, H. A. Beydoun, *et al.*, "Imaging Mass Spectrometry of a Specific Fragment of Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase Kinase Kinase 2 Discriminates Cancer from Uninvolved Prostate Tissue," *Clinical Cancer Research,* vol. 15, pp. 5541-5551, Sep 2009.
[79] J. W. Park, H. K. Shon, B. C. Yoo, I. H. Kim, D. W. Moon, and T. G. Lee, "Differentiation between human normal colon mucosa and colon cancer tissue using ToF-SIMS imaging technique and principal component analysis," *Applied Surface Science,* vol. 255, pp. 1119-1122, Dec 2008.
[80] M. C. Djidja, E. Claude, M. F. Snel, P. Scriven, S. Francese, V. Carolan, *et al.*, "MALDI-Ion Mobility Separation-Mass Spectrometry Imaging of Glucose-Regulated Protein 78 kDa (Grp78) in Human Formalin-Fixed, Paraffin-Embedded Pancreatic Adenocarcinoma Tissue Sections," *Journal of Proteome Research,* vol. 8, pp. 4876-4884, Oct 2009.
[81] A. L. Dill, D. R. Ifa, N. E. Manicke, A. B. Costa, J. A. Ramos-Vara, D. W. Knapp, *et al.*, "Lipid Profiles of Canine Invasive Transitional Cell Carcinoma of

the Urinary Bladder and Adjacent Normal Tissue by Desorption Electrospray Ionization Imaging Mass Spectrometry," *Analytical Chemistry,* vol. 81, pp. 8758-8764, Nov 2009.

[82] Y. F. Chen, J. Allegood, Y. Liu, E. Wang, B. Cachon-Gonzalez, T. M. Cox, *et al.*, "Imaging MALDI mass spectrometry using an oscillating capillary nebulizer matrix coating system and its application to analysis of lipids in brain from a mouse model of Tay-Sachs/Sandhoff disease," *Analytical Chemistry,* vol. 80, pp. 2780-2788, Apr 2008.

[83] M. Aranyosiova, M. Michalka, M. Kopani, B. Rychly, J. Jakubovsky, and D. Velic, "Microscopy and chemical imaging of Behcet brain tissue," *Applied Surface Science,* vol. 255, pp. 1584-1587, Dec 2008.

[84] D. Hare, B. Reedy, R. Grimm, S. Wilkins, I. Volitakis, J. L. George, *et al.*, "Quantitative elemental bio-imaging of Mn, Fe, Cu and Zn in 6-hydroxydopamine induced Parkinsonism mouse models," *Metallomics,* vol. 1, pp. 53-58, Jan 2009.

[85] K. Skold, M. Svensson, A. Nilsson, X. Q. Zhang, K. Nydahl, R. M. Caprioli, *et al.*, "Decreased striatal levels of PEP-19 following MPTP lesion in the mouse," *Journal of Proteome Research,* vol. 5, pp. 262-269, Feb 2006.

[86] R. W. Hutchinson, A. G. Cox, C. W. McLeod, P. S. Marshall, A. Harper, E. L. Dawson, *et al.*, "Imaging and spatial distribution of beta-amyloid peptide and metal ions in Alzheimer's plaques by laser ablation-inductively coupled plasma-mass spectrometry," *Analytical Biochemistry,* vol. 346, pp. 225-233, Nov 2005.

[87] N. Tahallah, A. Brunelle, S. De La Porte, and O. Laprevote, "Lipid mapping in human dystrophic muscle by cluster-time-of-flight secondary ion mass spectrometry imaging," *Journal of Lipid Research,* vol. 49, pp. 438-454, Feb 2008.

[88] D. Touboul, A. Brunelle, F. Halgand, S. De La Porte, and O. Laprevote, "Lipid imaging by gold cluster time-of-flight secondary ion mass spectrometry: application to Duchenne muscular dystrophy," *Journal of Lipid Research,* vol. 46, pp. 1388-1395, Jul 2005.

[89] D. Touboul, S. Roy, D. P. Germain, P. Chaminade, A. Brunelle, and O. Laprevote, "MALDI-TOF and cluster-TOF-SIMS imaging of Fabry disease biomarkers," *International Journal of Mass Spectrometry,* vol. 260, pp. 158-165, Feb 2007.

[90] N. E. Manicke, M. Nefliu, C. Wu, J. W. Woods, V. Reiser, R. C. Hendrickson, *et al.*, "Imaging of Lipids in Atheroma by Desorption Electrospray Ionization Mass Spectrometry," *Analytical Chemistry,* vol. 18, pp. 8702-8707, 2009.

[91] J. H. Kim, B. J. Ahn, J. H. Park, H. K. Shon, Y. S. Yu, D. W. Moon, *et al.*, "Label-free calcium imaging in ischemic retinal tissue by TOF-SIMS," *Biophysical Journal,* vol. 94, pp. 4095-4102, May 2008.

[92] S. Koizumi, S. Yamamoto, T. Hayasaka, Y. Konishi, M. Yamaguchi-Okada, N. Goto-Inoue, *et al.*, "IMAGING MASS SPECTROMETRY REVEALED THE PRODUCTION OF LYSO-PHOSPHATIDYLCHOLINE IN THE INJURED ISCHEMIC RAT BRAIN," *Neuroscience,* vol. 168, pp. 219-225, Jun 2010.

[93] S. X. Jiang, S. Whitehead, A. Aylsworth, J. Slinn, B. Zurakowski, K. Chan, *et al.*, "Neuropilin 1 Directly Interacts with Fer Kinase to Mediate Semaphorin 3A-

induced Death of Cortical Neurons," *Journal of Biological Chemistry,* vol. 285, pp. 9908-9918, 2010.

[94] C. Eriksson, K. Borner, H. Nygren, K. Ohlson, U. Bexell, N. Billerdahl, *et al.*, "Studies by imaging TOF-SIMS of bone mineralization on porous titanium implants after 1 week in bone," 2006, pp. 6757-6760.

[95] H. Nygren, C. Eriksson, K. Hederstierna, and P. Malmberg, "TOF-SIMS analysis of the interface between bone and titanium implants-Effect of porosity and magnesium coating," 2008, pp. 1092-1095.

[96] E. Acquadro, C. Cabella, S. Ghiani, L. Miragoli, E. M. Bucci, and D. Corpillo, "Matrix-Assisted laser Desorption Ionization Imaging Mass Spectrometry Detection of a Magnetic Resonance Imaging Contrast Agent in Mouse liver," *Analytical Chemistry,* vol. 81, pp. 2779-2784, Apr 2009.

[97] S. J. Atkinson, P. M. Loadman, C. Sutton, L. H. Patterson, and M. R. Clench, "Examination of the distribution of the bioreductive drug AQ4N and its active metabolite AQ4 in solid tumours by imaging matrix-assisted laser desorption/ionisation mass spectrometry," *Rapid Communications in Mass Spectrometry,* vol. 21, pp. 1271-1276, 2007.

[98] L. Signor, E. Varesio, R. F. Staack, V. Starke, W. F. Richter, and G. Hopfgartner, "Analysis of erlotinib and its metabolites in rat tissue sections by MALDI quadrupole time-of-flight mass spectrometry," *Journal of Mass Spectrometry,* vol. 42, pp. 900-909, Jul 2007.

[99] P. J. Trim, C. M. Henson, J. L. Avery, A. McEwen, M. F. Snel, E. Claude, *et al.*, "Matrix-Assisted Laser Desorption/Ionization-Ion Mobility Separation-Mass Spectrometry Imaging of Vinblastine in Whole Body Tissue Sections," *Analytical Chemistry,* vol. 80, pp. 8628-8634, Nov 2008.

[100] J. M. Wiseman, D. R. Ifa, Y. X. Zhu, C. B. Kissinger, N. E. Manicke, P. T. Kissinger, *et al.*, "Desorption electrospray ionization mass spectrometry: Imaging drugs and metabolites in tissues," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 105, pp. 18120-18125, Nov 2008.

[101] M. Zoriy, A. Matusch, T. Spruss, and J. S. Becker, "Laser ablation inductively coupled plasma mass spectrometry for imaging of copper, zinc, and platinum in thin sections of a kidney from a mouse treated with cis-platin," *International Journal of Mass Spectrometry,* vol. 260, pp. 102-106, Feb 2007.

[102] C. D. Kaddi and M. D. Wang, "Computational Methods for Mass Spectrometry Imaging: Challenges, Progress, and Opportunities," in *Health Informatics Data Analysis: Methods and Examples*, Y. Zhang, Ed., ed, 2015.

[103] R. K. Curtis, M. Orešič, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *Trends in Biotechnology,* vol. 23, pp. 429-435, 2005.

[104] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics,* vol. 81, pp. 98-104, 2002.

[105] R. G. Sadygov and J. R. Yates, III., "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases," *Analytical Chemistry,* vol. 75, pp. 3792-3798, 2003.

[106] Z. B. Alfassi, "Vector analysis of multi-measurements identification," *Journal of Radioanalytical and Nuclear Chemistry,* vol. 266, pp. 245-250, 2005.

[107] H. Hong, Y. Dragan, J. Epstein, C. Teitel, B. Chen, and Q. e. a. Xie, "Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) mass spectrometry (MS)," *BMC Bioinformatics,* vol. 6(Suppl2), 2004.

[108] L. A. McDonnell, A. von Remoortere, R. J. M. van Zeijl, and A. M. Deelder, "Mass spectrometry image correlation: quantifying colocalization," *Journal of Proteome Research,* vol. 7, pp. 3619-3627, 2008.

[109] S. E. Stein and D. R. Scott, "Optimization and testing of mass spectral library search algorithms for compound identification," *Journal of the American Society for Mass Spectrometry,* vol. 5, pp. 859-866, 1994.

[110] R. Van de Plas, K. Pelckmans, B. De Moor, and E. Waelkens, "Spatial querying of imaging mass spectrometry data: a nonnegative least squares approach," presented at the Neural Information Processing Systems Workshop on Machine Learning in Computational Biology 2007.

[111] X. Li and R. C. Dubes, "A probabilistic measure of similarity for binary data in pattern recognition," *Pattern Recognition,* vol. 22, pp. 397-409, 1989.

[112] V. Chvátal, "The tail of the hypergeometric distribution," *Discrete Mathematics,* vol. 25, pp. 285-287, 1979.

[113] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association,* vol. 58, pp. 13-30, 1963.

[114] O. Lund, M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak, *Immunological Bioinformatics*. Cambridge, MA: The MIT Press, 2005.

[115] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed.: Prentice Hall, 2002.

[116] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics,* vol. 18, pp. S231-S240, Oct 2002.

[117] G. H. Freeman and J. H. Halton, "Note on an Exact Treatment of Contingency, Goodness of Fit and Other Problems of Significance," *Biometrika,* vol. 38, pp. 141-149, 1951.

[118] P. Sprent and N. C. Smeeton, *Applied nonparametric statistical methods*. Boca Raton, FL: Chapman & Hall/CRC, 2001.

[119] C. Kaddi, R. M. Parry, and M. D. Wang, "Hypergeometric similarity measure for spatial analysis in tissue imaging mass spectrometry," *Proceedings of IEEE BIBM 2011* pp. 604-607 2011.

[120] A. Verbeek and P. M. Kroonenberg, "A Survey of Algorithms for Exact Distributions of Test Statistics in R X C Contingency Tables With Fixed Margins," *Computational Statistics & Data Analysis,* vol. 3, pp. 159-185, 1985.

[121] F. Greselin, "Counting and enumerating frequency tables with given margins," *Statistica & Applicazioni,* vol. 1, pp. 87-104, 2003.

[122] M. Gail and N. Mantel, "Counting the Number of r x c Contingency Tables with Fixed Margins," *Journal of the American Statistical Association,* vol. 72, pp. 859-862, 1977.

[123] G. B. Nath and P. V. K. Iyer, "Note on the combinatorial formula for nHr," *Journal of the Australian Mathematical Society,* vol. 14, pp. 264-268, 1972.

[124] H. Anand, V. C. Dumir, and H. Gupta, "A combinatorial distribution problem," *Duke Mathematical Journal,* vol. 33, pp. 757-769, 1966.
[125] H. Ye, T. Yu, S. Temam, B. Ziober, J. Wang, J. Schwartz, *et al.*, "Transcriptomic dissection of tongue squamous cell carcinoma," *BMC Genomics,* vol. 9, pp. 1-11, 2008/02/06 2008.
[126] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, *et al.*, "TCPA: a resource for cancer functional proteomics data," *Nat Methods,* vol. 10, pp. 1046-7, Nov 2013.
[127] K. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "A multiscale and multiparametric approach for modeling the progression of oral cancer," *BMC Med Inform Decis Mak,* vol. 12, p. 136, 2012.
[128] X. Zhang, H. Yang, J. J. Lee, E. Kim, S. M. Lippman, F. R. Khuri, *et al.*, "MicroRNA-related Genetic Variations as Predictors for Risk of Second Primary Tumor and/or Recurrence in Patients with Early-Stage Head and Neck Cancer," *Carcinogenesis,* September 5, 2010 2010.
[129] L.-J. Ma, W. Li, X. Zhang, D.-H. Huang, H. Zhang, J.-Y. Xiao, *et al.*, "Differential Gene Expression Profiling of Laryngeal Squamous Cell Carcinoma by Laser Capture Microdissection and Complementary DNA Microarrays," *Archives of Medical Research,* vol. 40, pp. 114-123, 2// 2009.
[130] K. Uchida, A. Oga, M. Nakao, T. Mano, M. Mihara, S. Kawauchi, *et al.*, "Loss of 3p26.3 is an independent prognostic factor in patients with oral squamous cell carcinoma," *Oncol Rep,* vol. 26, pp. 463-9, Aug 2011.
[131] C. Xu, Y. Liu, P. Wang, W. Fan, T. C. Rue, M. P. Upton, *et al.*, "Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival," *Mol Cancer,* vol. 9, p. 143, 2010.
[132] S. C. Cheong, G. V. R. Chandramouli, A. Saleh, R. B. Zain, S. H. Lau, S. Sivakumaren, *et al.*, "Gene expression in human oral squamous cell carcinoma is influenced by risk factor exposure," *Oral Oncology,* vol. 45, pp. 712-719, 8// 2009.
[133] B. Bojovic and D. L. Crowe, "Telomere dysfunction promotes metastasis in a TERC null mouse model of head and neck cancer," *Mol Cancer Res,* vol. 9, pp. 901-13, Jul 2011.
[134] A. A. Saeed, A. H. Sims, S. S. Prime, I. Paterson, P. G. Murray, and V. R. Lopes, "Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas," *Oral Oncology,* vol. 51, pp. 237-246, 3// 2015.
[135] C. Perisanidis, B. Savarese-Brenner, T. Würger, F. Wrba, A. Huynh, C. Schopper, *et al.*, "HCRP1 expression status is a significant prognostic marker in oral and oropharyngeal cancer," *Oral Diseases,* vol. 19, pp. 206-211, 2013.
[136] Y. Kuribayashi, K.-i. Morita, H. Tomioka, M. Uekusa, D. Ito, and K. Omura, "Gene expression analysis by oligonucleotide microarray in oral leukoplakia," *Journal of Oral Pathology & Medicine,* vol. 38, pp. 356-361, 2009.
[137] Z.-w. Yu, L.-p. Zhong, T. Ji, P. Zhang, W.-t. Chen, and C.-p. Zhang, "MicroRNAs contribute to the chemoresistance of cisplatin in tongue squamous cell carcinoma lines," *Oral Oncology,* vol. 46, pp. 317-322, 4// 2010.

[138] P. D. Vermeer, P. L. Colbert, B. G. Wieking, D. W. Vermeer, and J. H. Lee, "Targeting ERBB Receptors Shifts Their Partners and Triggers Persistent ERK Signaling through a Novel ERBB/EFNB1 Complex," *Cancer Research,* vol. 73, pp. 5787-5797, September 15, 2013 2013.

[139] A. Ghosh, S. Ghosh, G. P. Maiti, M. G. Sabbir, N. Alam, N. Sikdar, *et al.*, "SH3GL2 and CDKN2A/2B loci are independently altered in early dysplastic lesions of head and neck: correlation with HPV infection and tobacco habit," *The Journal of Pathology,* vol. 217, pp. 408-419, 2009.

[140] N. J. Silveira, L. Varuzza, A. Machado-Lima, M. S. Lauretto, D. G. Pinheiro, R. V. Rodrigues, *et al.*, "Searching for molecular markers in head and neck squamous cell carcinomas (HNSCC) by statistical and bioinformatic analysis of larynx-derived SAGE libraries," *BMC Med Genomics,* vol. 1, p. 56, 2008.

[141] G. Zheng, M. Zhou, X. Ou, B. Peng, Y. Yu, F. Kong, *et al.*, "Identification of carbonic anhydrase 9 as a contributor to pingyangmycin-induced drug resistance in human tongue cancer cells," *FEBS Journal,* vol. 277, pp. 4506-4518, 2010.

[142] S. Iwasawa, Y. Yamano, Y. Takiguchi, H. Tanzawa, K. Tatsumi, and K. Uzawa, "Upregulation of thioredoxin reductase 1 in human oral squamous cell carcinoma," *Oncol Rep,* vol. 25, pp. 637-44, Mar 2011.

[143] S. Zhang, X. L. Feng, L. Shi, C. J. Gong, Z. J. He, H. J. Wu, *et al.*, "Genome-wide analysis of DNA methylation in tongue squamous cell carcinoma," *Oncol Rep,* vol. 29, pp. 1819-26, May 2013.

[144] L. Ruan, X.-H. Li, X.-X. Wan, H. Yi, C. Li, M.-Y. Li, *et al.*, "Analysis of EGFR signaling pathway in nasopharyngeal carcinoma cells by quantitative phosphoproteomics," *Proteome Science,* vol. 9, pp. 35-35, 2011.

[145] M. P. Wong, M. Cheang, E. Yorida, A. Coldman, C. B. Gilks, D. Huntsman, *et al.*, "Loss of desmoglein 1 expression associated with worse prognosis in head and neck squamous cell carcinoma patients," *Pathology,* vol. 40, pp. 611-616, 2008/01/01 2008.

[146] I. Takeda, S.-i. Maruya, T. Shirasaki, H. Mizukami, T. Takahata, J. N. Myers, *et al.*, "Simvastatin inactivates β1-integrin and extracellular signal-related kinase signaling and inhibits cell proliferation in head and neck squamous cell carcinoma cells," *Cancer Science,* vol. 98, pp. 890-899, 2007.

[147] H. Mirghani, N. Ugolin, C. Ory, M. Lefèvre, S. Baulande, P. Hofman, *et al.*, "A predictive transcriptomic signature of oropharyngeal cancer according to HPV16 status exclusively," *Oral Oncology,* vol. 50, pp. 1025-1034, 11// 2014.

[148] R. B. Pai, S. B. Pai, R. M. Lalitha, S. V. Kumaraswamy, N. Lalitha, R. N. Johnston, *et al.*, "Over-expression of c-Myc oncoprotein in oral squamous cell carcinoma in the South Indian population," *ecancermedicalscience,* vol. 3, p. 128, 2009.

[149] E. Vairaktaris, S. Loukeri, S. Vassiliou, E. Nkenke, S. Spyridonidou, A. Vylliotis, *et al.*, "EGFR and c-Jun exhibit the same pattern of expression and increase gradually during the progress of oral oncogenesis," *In Vivo,* vol. 21, pp. 791-6, Sep-Oct 2007.

[150] N. I. Pollock and J. R. Grandis, "HER2 as a Therapeutic Target in Head and Neck Squamous Cell Carcinoma," *Clinical Cancer Research,* November 25, 2014 2014.

[151] J.-S. Kim, H. S. Yun, H.-D. Um, J. K. Park, K.-H. Lee, C.-M. Kang, *et al.*, "Identification of inositol polyphosphate 4-phosphatase type II as a novel tumor resistance biomarker in human laryngeal cancer HEp-2 cells," *Cancer Biology & Therapy,* vol. 13, pp. 1307-1318, 2012.

[152] J. W.-F. Yuen, G. T.-Y. Chung, S. W.-M. Lun, C. C.-M. Cheung, K.-F. To, and K.-W. Lo, "Epigenetic Inactivation of Inositol polyphosphate 4-phosphatase B (INPP4B), a Regulator of PI3K/AKT Signaling Pathway in EBV-Associated Nasopharyngeal Carcinoma," *PLoS ONE,* vol. 9, p. e105163, 2014.

[153] Y.-W. Su, Y.-H. Lin, M.-H. Pai, A.-C. Lo, Y.-C. Lee, I. C. Fang, *et al.*, "Association between Phosphorylated AMP-Activated Protein Kinase and Acetyl-CoA Carboxylase Expression and Outcome in Patients with Squamous Cell Carcinoma of the Head and Neck," *PLoS ONE,* vol. 9, p. e96183, 2014.

[154] R. Mishra, "Glycogen synthase kinase 3 beta: can it be a target for oral cancer," *Molecular Cancer,* vol. 9, pp. 144-144, 2010.

[155] B. J. Moeller, J. S. Yordy, M. D. Williams, U. Giri, U. Raju, D. P. Molkentine, *et al.*, "DNA repair biomarker profiling of head and neck cancer: Ku80 expression predicts locoregional failure and death following radiotherapy," *Clin Cancer Res,* vol. 17, pp. 2035-43, Apr 1 2011.

[156] C. Hebert, K. Norris, P. Parashar, R. A. Ord, N. G. Nikitakis, and J. J. Sauk, "Hypoxia-inducible factor-1alpha polymorphisms and TSC1/2 mutations are complementary in head and neck cancers," *Mol Cancer,* vol. 5, p. 3, 2006.

[157] I. Irigoien and C. Arenas, "INCA: New statistic for estimating the number of clusters and identifying atypical units," *Statistics in Medicine,* vol. 27, pp. 2948-2973, 2008.

[158] J. B. German, L. A. Gillies, J. T. Smilowitz, A. M. Zivkovic, and S. M. Watkins, "Lipidomics and lipid profiling in metabolomics," *Curr Opin Lipidol,* vol. 18, pp. 66-71, Feb 2007.

[159] P. Hunter, "Reading the metabolic fine print. The application of metabolomics to diagnostics, drug research and nutrition might be integral to improved health and personalized medicine," *EMBO Reports,* vol. 10, pp. 20-23, 2009.

[160] G. N. Gowda, S. Zhang, H. Gu, V. Asiago, N. Shanaiah, and D. Raftery, "Metabolomics-based methods for early disease diagnostics," *Expert Review of Molecular Diagnostics,* vol. 8, pp. 617-633, 2008.

[161] A. Zhang, H. Sun, and X. Wang, "Saliva Metabolomics Opens Door to Biomarker Discovery, Disease Diagnosis, and Treatment," *Applied Biochemistry and Biotechnology,* vol. 168, pp. 1718-1727, 2012/11/01 2012.

[162] T. Hyotylainen and S. Wiedmer, *Chromatographic Methods in Metabolomics*: Royal Society of Chemistry, 2013.

[163] E. M. Lenz and I. D. Wilson, "Analytical strategies in metabonomics," *J Proteome Res,* vol. 6, pp. 443-58, Feb 2007.

[164] C. F.Poole and S. K. Poole, "Instrumental Thin-Layer Chromatography," *Analytical Chemistry,* vol. 66, pp. 27A-37A, 1994/01/01 1994.

[165] T. Ferenci and R. Maharjan, "Comparative Metabolome Profiling Using Two Dimensional Thin Layer Chromatography (2DTLC)," in *Metabolome Analyses: Strategies for Systems Biology*, S. Vaidyanathan, G. Harrigan, and R. Goodacre, Eds., ed: Springer US, 2005, pp. 63-81.

[166] R. Prasad Maharjan and T. Ferenci, "Global metabolite analysis: the influence of extraction methodology on metabolome profiles of Escherichia coli," *Analytical Biochemistry,* vol. 313, pp. 145-154, 2/1/ 2003.
[167] K. L. Busch, "Mass spectrometric detection for thin-layer chromatographic separations," *TRAC Trends in Analytical Chemistry,* vol. 11, pp. 314-324, 1992.
[168] S. E. Unger, A. Vincze, R. G. Cooks, R. Chrisman, and L. D. Rothman, "Identification of quaternary alkaloids in mushroom by chromatography/secondary ion mass spectrometry," *Analytical Chemistry,* vol. 53, pp. 976-981, 1981/06/01 1981.
[169] S.-C. Cheng, M.-Z. Huang, and J. Shiea, "Thin layer chromatography/mass spectrometry," *Journal of Chromatography A,* vol. 1218, pp. 2700-2711, 2011.
[170] G. Morlock and W. Schwack, "Coupling of planar chromatography to mass spectrometry," *TRAC Trends in Analytical Chemistry,* vol. 29, pp. 1157-1171, 2010.
[171] A. I. Gusev, O. J. Vasseur, A. Proctor, A. G. Sharkey, and D. M. Hercules, "Imaging of thin-layer chromatograms using matrix-assisted laser desorption/ionization mass spectrometry," *Analytical Chemistry,* vol. 67, pp. 4565-4570, 1995/12/01 1995.
[172] G. J. Van Berkel, M. J. Ford, and M. A. Deibel, "Thin-Layer Chromatography and Mass Spectrometry Coupled Using Desorption Electrospray Ionization," *Analytical Chemistry,* vol. 77, pp. 1207-1215, 2005/03/01 2005.
[173] G. J. Van Berkel and V. Kertesz, "Automated Sampling and Imaging of Analytes Separated on Thin-Layer Chromatography Plates Using Desorption Electrospray Ionization Mass Spectrometry," *Analytical Chemistry,* vol. 78, pp. 4938-4944, 2006/07/01 2006.
[174] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 9, pp. 62-66, 1979.
[175] C. Kaddi, R. M. Parry, and M. D. Wang, "Hypergeometric Similarity Measure for Spatial Analysis in Tissue Imaging Mass Spectrometry," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on,* 2011, pp. 604-607.
[176] C. D. Kaddi, R. M. Parry, and M. D. Wang, "Multivariate hypergeometric similarity measure," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 10, pp. 1505-16, Nov-Dec 2013.
[177] C. D. Kaddi, R. V. Bennett, M. R. L. Paine, A. L. Weber, F. M. Fernandez, and M. D. Wang, "DetectTLC: A Tool for Turnkey Reaction Mixture Screening on the Basis of Desorption Electrospray Ionization Images," *Manuscript In Preparation,* 2015.
[178] X. Gu, X. Song, Y. Dong, H. Cai, E. Walters, R. Zhang, *et al.*, "Vitamin E succinate induces ceramide-mediated apoptosis in head and neck squamous cell carcinoma in vitro and in vivo," *Clin Cancer Res,* vol. 14, pp. 1840-8, Mar 15 2008.
[179] Y. L. Yang, C. Ji, Z. G. Bi, C. C. Lu, R. Wang, B. Gu, *et al.*, "Deguelin induces both apoptosis and autophagy in cultured head and neck squamous cell carcinoma cells," *PLoS One,* vol. 8, p. e54736, 2013.
[180] M. J. Frederick, A. J. VanMeter, M. A. Gadhikar, Y. C. Henderson, H. Yao, C. C. Pickering, *et al.*, "Phosphoproteomic analysis of signaling pathways in head and

neck squamous cell carcinoma patient samples," *Am J Pathol*, vol. 178, pp. 548-71, Feb 2011.

[181] B. Hong, V. W. Lui, E. P. Hui, Y. Lu, H. S. Leung, E. Y. Wong, *et al.*, "Reverse phase protein array identifies novel anti-invasion mechanisms of YC-1," *Biochem Pharmacol*, vol. 79, pp. 842-52, Mar 15 2010.

[182] A. M. Gonzalez-Angulo, B. T. Hennessy, F. Meric-Bernstam, A. Sahin, W. Liu, Z. Ju, *et al.*, "Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer," *Clin Proteomics*, vol. 8, p. 11, 2011.

[183] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, *et al.*, "Modeling precision treatment of breast cancer," *Genome Biol*, vol. 14, p. R110, 2013.

[184] J. Sonntag, C. Bender, Z. Soons, S. v. der Heyde, R. König, S. Wiemann, *et al.*, "Reverse phase protein array based tumor profiling identifies a biomarker signature for risk classification of hormone receptor-positive breast cancer," *Translational Proteomics*, vol. 2, pp. 52-59, 3// 2014.

[185] M. S. Carey, R. Agarwal, B. Gilks, K. Swenerton, S. Kalloger, J. Santos, *et al.*, "Functional proteomic analysis of advanced serous ovarian cancer using reverse phase protein array: TGF-beta pathway signaling indicates response to primary chemotherapy," *Clin Cancer Res*, vol. 16, pp. 2852-60, May 15 2010.

[186] R. Ummanni, H. A. Mannsperger, J. Sonntag, M. Oswald, A. K. Sharma, R. Konig, *et al.*, "Evaluation of reverse phase protein array (RPPA)-based pathway-activation profiling in 84 non-small cell lung cancer (NSCLC) cell lines as platform for cancer proteomics and biomarker discovery," *Biochim Biophys Acta*, Dec 19 2013.

[187] C. G. A. R. Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061-8, Oct 23 2008.

[188] D. K. Gascoigne, S. W. Cheetham, P. B. Cattenoz, M. B. Clark, P. P. Amaral, R. J. Taft, *et al.*, "Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes," *Bioinformatics*, vol. 28, pp. 3042-3050, December 1, 2012 2012.

[189] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, pp. 493-500, Feb 15 2010.

[190] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, *et al.*, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Res*, vol. 38, p. e178, Oct 2010.

[191] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, *et al.*, "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*, vol. 29, pp. 1035-1043, April 15, 2013 2013.

[192] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139-40, Jan 1 2010.

[193] G. Brown, A. Pocock, Z. Ming-Jie, and M. Luján, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research,* vol. 13, pp. 27-66, 2012.
[194] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol,* vol. 3, pp. 185-205, Apr 2005.
[195] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans Pattern Anal Mach Intell,* vol. 27, pp. 1226-38, Aug 2005.
[196] D. Bollegala, "Dynamic Feature Scaling for Online Learning of Binary Classifiers," *arXiv:1407.7584v1,* 2014.
[197] L. Wu, S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, *et al.*, "Variation and genetic control of protein abundance in humans," *Nature,* vol. 499, pp. 79-82, 07/04/print 2013.
[198] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nat Rev Genet,* vol. 13, pp. 227-232, 04//print 2012.
[199] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, *et al.*, "Global quantification of mammalian gene expression control," *Nature,* vol. 473, pp. 337-342, 05/19/print 2011.
[200] R. d. S. Abreu, L. O. Penalva, E. M. Marcotte, and C. Vogel, "Global signatures of protein and mRNA expression levels," *Molecular bioSystems,* vol. 5, pp. 1512-1526, 10/01 2009.
[201] S. Haider and R. Pal, "Integrated Analysis of Transcriptomic and Proteomic Data," *Current Genomics,* vol. 14, pp. 91-110, 2013.
[202] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc,* vol. 4, pp. 44-57, 2009.
[203] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res,* vol. 37, pp. 1-13, Jan 2009.
[204] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, *et al.*, "The Reactome pathway knowledgebase," *Nucleic Acids Res,* vol. 42, pp. D472-7, Jan 2014.
[205] A. K. Goulioumis, J. Varakis, P. Goumas, and H. Papadaki, "Androgen Receptor in Laryngeal Carcinoma: Could There Be an Androgen-Refractory Tumor?," *ISRN Oncology,* vol. 2011, p. 5, 2011.
[206] C. Elser, L. L. Siu, E. Winquist, M. Agulnik, G. R. Pond, S. F. Chin, *et al.*, "Phase II Trial of Sorafenib in Patients With Recurrent or Metastatic Squamous Cell Carcinoma of the Head and Neck or Nasopharyngeal Carcinoma," *Journal of Clinical Oncology,* vol. 25, pp. 3766-3773, August 20, 2007 2007.
[207] J. T. Chang, H.-M. Wang, K.-W. Chang, W.-H. Chen, M.-C. Wen, Y.-M. Hsu, *et al.*, "Identification of differentially expressed genes in oral squamous cell carcinoma (OSCC): Overexpression of NPM, CDK1 and NDRG1 and underexpression of CHES1," *International Journal of Cancer,* vol. 114, pp. 942-949, 2005.

[208] Y. Song, C. Zhao, L. Dong, M. Fu, L. Xue, Z. Huang, *et al.*, "Overexpression of cyclin B1 in human esophageal squamous cell carcinoma cells induces tumor cell invasive growth and metastasis," *Carcinogenesis,* vol. 29, pp. 307-315, February 1, 2008 2008.

[209] S. W. Pyo, M. Hashimoto, Y. S. Kim, C. H. Kim, S. H. Lee, K. R. Johnson, *et al.*, "Expression of E-cadherin, P-cadherin and N-cadherin in oral squamous cell carcinoma: correlation with the clinicopathologic features and patient outcome," *J Craniomaxillofac Surg,* vol. 35, pp. 1-9, Jan 2007.

[210] P. Amornphimoltham, V. Sriuranpong, V. Patel, F. Benavides, C. J. Conti, J. Sauk, *et al.*, "Persistent activation of the Akt pathway in head and neck squamous cell carcinoma: a potential target for UCN-01," *Clin Cancer Res,* vol. 10, pp. 4029-37, Jun 15 2004.

[211] C. Freudlsperger, J. R. Burnett, J. A. Friedman, V. R. Kannabiran, Z. Chen, and C. Van Waes, "EGFR–PI3K–AKT–mTOR signaling in head and neck squamous cell carcinomas: attractive targets for molecular-oriented therapy," *Expert Opinion on Therapeutic Targets,* vol. 15, pp. 63-74, 2011.

[212] K. A. Gold, H.-Y. Lee, and E. S. Kim, "Targeted therapies in squamous cell carcinoma of the head and neck," *Cancer,* vol. 115, pp. 922-935, 2009.

[213] A. Weber, U. R. Hengge, I. Stricker, I. Tischoff, A. Markwart, K. Anhalt, *et al.*, "Protein microarrays for the detection of biomarkers in head and neck squamous cell carcinomas," *Human Pathology,* vol. 38, pp. 228-238, 2// 2007.

[214] P. Lothaire, E. de Azambuja, D. Dequanter, Y. Lalami, C. Sotiriou, G. Andry, *et al.*, "Molecular markers of head and neck squamous cell carcinoma: Promising signs in need of prospective evaluation," *Head & Neck,* vol. 28, pp. 256-269, 2006.

[215] X.-H. Tang and L. J. Gudas, "Retinoids, Retinoic Acid Receptors, and Cancer," *Annual Review of Pathology: Mechanisms of Disease,* vol. 6, pp. 345-364, 2011/02/28 2011.

[216] E. De Corso, S. Baroni, S. Agostino, G. Cammarota, G. Mascagna, A. Mannocci, *et al.*, "Bile Acids and Total Bilirubin Detection in Saliva of Patients Submitted to Gastric Surgery and in Particular to Subtotal Billroth II Resection," *Annals of Surgery,* vol. 245, pp. 880-885, 2007.

[217] M.-W. Sung, J.-L. Roh, B. J. Park, S. W. Park, T.-K. Kwon, S. J. Lee, *et al.*, "Bile Acid Induces Cyclo-Oxygenase-2 Expression in Cultured Human Pharyngeal Cells: A Possible Mechanism of Carcinogenesis in the Upper Aerodigestive Tract by Laryngopharyngeal Reflux," *The Laryngoscope,* vol. 113, pp. 1059-1063, 2003.

[218] M. R. Junttila, R. Ala-aho, T. Jokilehto, J. Peltonen, M. Kallajoki, R. Grenman, *et al.*, "p38[alpha] and p38[delta] mitogen-activated protein kinase isoforms regulate invasion and growth of head and neck squamous carcinoma cells," *Oncogene,* vol. 26, pp. 5267-5279, 03/05/online 2007.

[219] J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological Bulletin,* vol. 87, pp. 245-251, 1980.

[220] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms,* 1st ed.: Chapman and Hall/CRC, 2012.

[221] T. C. G. A. Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, pp. 576-82, Jan 29 2015.
[222] Y.-H. Yu, H.-K. Kuo, and K.-W. Chang, "The Evolving Transcriptome of Head and Neck Squamous Cell Carcinoma: A Systematic Review," *PLoS ONE*, vol. 3, p. e3215, 2008.
[223] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews. Genetics*, vol. 10, pp. 57-63, 2009.
[224] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, "Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets," *PLoS Medicine*, vol. 5, p. e184, 09/02 2008.
[225] F. Clatot, S. Gouérant, S. Mareschal, M. Cornic, A. Berghian, O. Choussy, *et al.*, "The gene expression profile of inflammatory, hypoxic and metabolic genes predicts the metastatic spread of human head and neck squamous cell carcinoma," *Oral Oncology*, vol. 50, pp. 200-207, 3// 2014.
[226] P. Roepman, L. F. Wessels, N. Kettelarij, P. Kemmeren, A. J. Miles, P. Lijnzaad, *et al.*, "An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas," *Nat Genet*, vol. 37, pp. 182-6, Feb 2005.
[227] L. De Cecco, P. Bossi, L. Locati, S. Canevari, and L. Licitra, "Comprehensive gene expression meta-analysis of head and neck squamous cell carcinoma microarray data defines a robust survival predictor," *Annals of Oncology*, May 14, 2014 2014.
[228] E. Fountzilas, K. Markou, K. Vlachtsis, A. Nikolaou, P. Arapantoni-Dadioti, E. Ntoula, *et al.*, "Identification and validation of gene expression models that predict clinical outcome in patients with early-stage laryngeal cancer," *Annals of Oncology*, vol. 23, pp. 2146-2153, August 1, 2012 2012.
[229] M. P. Rethman, W. Carpenter, E. E. Cohen, J. Epstein, C. A. Evans, C. M. Flaitz, *et al.*, "Evidence-based clinical recommendations regarding screening for oral squamous cell carcinomas," *J Am Dent Assoc*, vol. 141, pp. 509-20, May 2010.
[230] D. Pyeon, M. A. Newton, P. F. Lambert, J. A. den Boon, S. Sengupta, C. J. Marsit, *et al.*, "Fundamental Differences in Cell Cycle Deregulation in Human Papillomavirus–Positive and Human Papillomavirus–Negative Head/Neck and Cervical Cancers," *Cancer research*, vol. 67, pp. 4605-4619, 2007.
[231] M. Pavón, M. Parreño, M. Téllez-Gabriel, F. Sancho, M. López, M. Céspedes, *et al.*, "Gene expression signatures and molecular markers associated with clinical outcome in locally advanced head and neck carcinoma," *Carcinogenesis*, vol. 33, pp. 1707-1716, September 1, 2012 2012.
[232] G. Ohmura, T. Tsujikawa, T. Yaguchi, N. Kawamura, S. Mikami, J. Sugiyama, *et al.*, "Aberrant Myosin 1b Expression Promotes Cell Migration and Lymph Node Metastasis of HNSCC," *Mol Cancer Res*, vol. 13, pp. 721-31, Apr 2015.
[233] A. Stanam, L. Love-Homan, T. S. Joseph, M. Espinosa-Cotton, and A. L. Simons, "Upregulated interleukin-6 expression contributes to erlotinib resistance in head and neck squamous cell carcinoma," *Molecular Oncology*.
[234] C. Zhang, X. Song, M. Zhu, S. Shi, M. Li, L. Jin, *et al.*, "Association between MMP1 -1607 1G>2G polymorphism and head and neck cancer risk: a meta-analysis," *PLoS One*, vol. 8, p. e56294, 2013.

[235] D. Elashoff, H. Zhou, J. Reiss, J. Wang, H. Xiao, B. Henson, *et al.*, "Prevalidation of Salivary Biomarkers for Oral Cancer Detection," *Cancer Epidemiology Biomarkers & Prevention,* vol. 21, pp. 664-672, April 1, 2012 2012.

[236] E. J. Franzmann, E. P. Reategui, L. H. M. Pereira, F. Pedroso, D. Joseph, G. O. Allen, *et al.*, "Salivary protein and solCD44 levels as a potential screening tool for early detection of head and neck squamous cell carcinoma," *Head & Neck,* vol. 34, pp. 687-695, 2012.

[237] S. Ying, D. Wei, Z. Chunguang, Z. You, C. Zhongbo, T. Yuan, *et al.*, "A Computational Method for Prediction of Saliva-Secretory Proteins and Its Application to Identification of Head and Neck Cancer Biomarkers for Salivary Diagnosis," *NanoBioscience, IEEE Transactions on,* vol. 14, pp. 167-174, 2015.

[238] C. D. Kaddi and M. D. Wang, "Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma using Proteomic and Transcriptomic Data," *Manuscript Under Review,* 2015.

[239] S. Marur, G. D'Souza, W. H. Westra, and A. A. Forastiere, "HPV-associated head and neck cancer: a virus-related cancer epidemic," *The Lancet Oncology,* vol. 11, pp. 781-789, 8// 2010.

[240] D. M. Shin, "Oral cancer prevention advances with a translational trial of green tea," *Cancer Prev Res (Phila),* vol. 2, pp. 919-21, Nov 2009.

[241] N. Li, Z. Sun, C. Han, and J. Chen, "The Chemopreventive Effects of Tea on Human Oral Precancerous Mucosa Lesions," *Experimental Biology and Medicine,* vol. 220, pp. 218-224, 1999.

[242] A. S. Tsao, D. Liu, J. Martin, X.-m. Tang, J. J. Lee, A. K. El-Naggar, *et al.*, "Phase II Randomized, Placebo-Controlled Trial of Green Tea Extract in Patients with High-Risk Oral Premalignant Lesions," *Cancer Prevention Research,* vol. 2, pp. 931-941, November 1, 2009 2009.

[243] A. R. Amin, O. Kucuk, F. R. Khuri, and D. M. Shin, "Perspectives for cancer prevention with natural compounds," *J Clin Oncol,* vol. 27, pp. 2712-25, Jun 1 2009.

[244] A. R. Amin, D. Wang, H. Zhang, S. Peng, H. J. Shin, J. C. Brandes, *et al.*, "Enhanced anti-tumor activity by the combination of the natural compounds (-)-epigallocatechin-3-gallate and luteolin: potential role of p53," *J Biol Chem,* vol. 285, pp. 34557-65, Nov 5 2010.

[245] C. D. Kaddi and M. D. Wang, "Mathematical model of the effect of intercellular cooperative interactions in cancer during drug therapy," in *Biomedical Sciences and Engineering Conference (BSEC), 2013*, 2013, pp. 1-4.

[246] W. M. Harriss-Phillips, E. Bezak, and E. K. Yeoh, "Monte Carlo radiotherapy simulations of accelerated repopulation and reoxygenation for hypoxic head and neck cancer," *Br J Radiol,* vol. 84, pp. 903-18, Oct 2011.

[247] W. Tuckwell, E. Bezak, E. Yeoh, and L. Marcu, "Efficient Monte Carlo modelling of individual tumour cell propagation for hypoxic head and neck cancer," *Phys Med Biol,* vol. 53, pp. 4489-507, Sep 7 2008.

[248] A. V. Chvetsov, "Tumor response parameters for head and neck cancer derived from tumor-volume variation during radiation therapy," *Med Phys,* vol. 40, p. 034101, Mar 2013.

[249] A. V. Chvetsov, L. Dong, J. R. Palta, and R. J. Amdur, "Tumor-volume simulation during radiotherapy for head-and-neck cancer using a four-level cell population model," *Int J Radiat Oncol Biol Phys,* vol. 75, pp. 595-602, Oct 1 2009.

[250] B. Titz and R. Jeraj, "An imaging-based tumour growth and treatment response model: investigating the effect of tumour oxygenation on radiation therapy response," *Phys Med Biol,* vol. 53, pp. 4471-88, Sep 7 2008.

[251] L. Marcu, E. Bezak, and I. Olver, "Scheduling cisplatin and radiotherapy in the treatment of squamous cell carcinomas of the head and neck: a modelling approach," *Physics in Medicine and Biology,* vol. 51, pp. 3625-3637, Aug 2006.

[252] L. G. Marcu and E. Bezak, "Neoadjuvant cisplatin for head and neck cancer: Simulation of a novel schedule for improved therapeutic ratio," *J Theor Biol,* vol. 297, pp. 41-7, Mar 21 2012.

[253] I. Sorrell, R. J. Shipley, V. Hearnden, H. E. Colley, M. H. Thornhill, C. Murdoch, *et al.*, "Combined mathematical modelling and experimentation to predict polymersome uptake by oral cancer cells," *Nanomedicine,* vol. 10, pp. 339-48, Feb 2014.

[254] S. R. Earnshaw, A. P. Brogan, and C. L. McDade, "Model-based cost-effectiveness analyses for prostate cancer chemoprevention : a review and summary of challenges," *Pharmacoeconomics,* vol. 31, pp. 289-304, Apr 2013.

[255] C. Hur, N. S. Nishioka, and G. S. Gazelle, "Cost-Effectiveness of Aspirin Chemoprevention for Barrett's Esophagus," *Journal of the National Cancer Institute,* vol. 96, pp. 316-325, 2004.

[256] I. Shureiqi, P. Reddy, and D. E. Brenner, "Chemoprevention: general perspective," *Critical Reviews in Oncology/Hematology,* vol. 33, pp. 157-167, 3// 2000.

[257] T. C. Chou, "Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies," *Pharmacol Rev,* vol. 58, pp. 621-81, Sep 2006.

[258] Y. Koizumi and S. Iwami, "Mathematical modeling of multi-drugs therapy: a challenge for determining the optimal combinations of antiviral drugs," *Theor Biol Med Model,* vol. 11, p. 41, 2014.

[259] B. B. Aggarwal, A. Bhardwaj, R. S. Aggarwal, N. P. Seeram, S. Shishodia, and Y. Takada, "Role of resveratrol in prevention and therapy of cancer: preclinical and clinical studies," *Anticancer Res,* vol. 24, pp. 2783-840, Sep-Oct 2004.

[260] J. K. Kundu and Y.-J. Surh, "Molecular basis of chemoprevention by resveratrol: NF-κB and AP-1 as potential targets," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* vol. 555, pp. 65-80, 11/2/ 2004.

[261] H. Jiang, X. Shang, H. Wu, S. C. Gautam, S. Al-Holou, C. Li, *et al.*, "Resveratrol downregulates PI3K/Akt/mTOR signaling pathways in human U251 glioma cells," *Journal of experimental therapeutics & oncology,* vol. 8, pp. 25-33, 2009.

[262] P. Roy, N. Kalra, S. Prasad, J. George, and Y. Shukla, "Chemopreventive Potential of Resveratrol in Mouse Skin Tumors Through Regulation of Mitochondrial and PI3K/AKT Signaling Pathways," *Pharmaceutical Research,* vol. 26, pp. 211-217, 2009/01/01 2009.

[263] G. S. Van Aller, J. D. Carson, W. Tang, H. Peng, L. Zhao, R. A. Copeland, *et al.*, "Epigallocatechin gallate (EGCG), a major component of green tea, is a dual phosphoinositide-3-kinase/mTOR inhibitor," *Biochemical and Biophysical Research Communications,* vol. 406, pp. 194-199, 3/11/ 2011.

[264] J. Y. Chung, J. O. Park, H. Phyu, Z. Dong, and C. S. Yang, "Mechanisms of inhibition of the Ras-MAP kinase signaling pathway in 30.7b Ras 12 cells by tea polyphenols (-)-epigallocatechin-3-gallate and theaflavin-3,3'-digallate," *Faseb j,* vol. 15, pp. 2022-4, Sep 2001.

[265] S. Fulda and K.-M. Debatin, "Resveratrol modulation of signal transduction in apoptosis and cell survival: A mini-review," *Cancer Detection and Prevention,* vol. 30, pp. 217-223, // 2006.

[266] M. Miloso, A. A. E. Bertelli, G. Nicolini, and G. Tredici, "Resveratrol-induced activation of the mitogen-activated protein kinases, ERK1 and ERK2, in human neuroblastoma SH-SY5Y cells," *Neuroscience Letters,* vol. 264, pp. 141-144, 4/2/ 1999.

[267] J. Bussink, A. J. van der Kogel, and J. H. Kaanders, "Activation of the PI3-K/AKT pathway and implications for radioresistance mechanisms in head and neck cancer," *Lancet Oncol,* vol. 9, pp. 288-96, Mar 2008.

[268] A. Franovic, L. Gunaratnam, K. Smith, I. Robert, D. Patten, and S. Lee, "Translational up-regulation of the EGFR by tumor hypoxia provides a nonmutational explanation for its overexpression in human cancer," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 104, pp. 13092-13097, 2007.

[269] W. R. Wilson and M. P. Hay, "Targeting hypoxia in cancer therapy," *Nat Rev Cancer,* vol. 11, pp. 393-410, 06//print 2011.

[270] H. Stegeman, J. H. Kaanders, D. L. Wheeler, A. J. van der Kogel, M. M. Verheijen, S. J. Waaijer, *et al.*, "Activation of AKT by hypoxia: a potential target for hypoxic tumors of the head and neck," *BMC Cancer,* vol. 12, pp. 463-463, 2012.

[271] A. A. Molinolo, S. M. Hewitt, P. Amornphimoltham, S. Keelawat, S. Rangdaeng, A. Meneses Garcia, *et al.*, "Dissecting the Akt/mammalian target of rapamycin signaling network: emerging results from the head and neck cancer tissue array initiative," *Clin Cancer Res,* vol. 13, pp. 4964-73, Sep 1 2007.

[272] Y. Baba, M. Fujii, Y. Tokumaru, and Y. Kato, "Present and Future of EGFR Inhibitors for Head and Neck Squamous Cell Cancer," *Journal of Oncology,* vol. 2012, p. 9, 2012.

[273] M. Rebucci, P. Peixoto, A. Dewitte, N. Wattez, M. A. De Nuncques, N. Rezvoy, *et al.*, "Mechanisms underlying resistance to cetuximab in the HNSCC cell line: role of AKT inhibition in bypassing this resistance," *Int J Oncol,* vol. 38, pp. 189-200, Jan 2011.

[274] L. Masuelli, E. D. Stefano, M. Fantini, R. Mattera, M. Benvenuto, L. Marzocchella, *et al.*, "Resveratrol potentiates the in vitro and in vivo anti-tumoral effects of curcumin in head and neck carcinomas," *Oncotarget,* vol. 5, pp. 10745-10762, 2014.

[275] D. R. Grimes, A. G. Fletcher, and M. Partridge, *Oxygen consumption dynamics in steady-state tumour models* vol. 1, 2014.

[276] D. R. Grimes, C. Kelly, K. Bloch, and M. Partridge, *A method for estimating the oxygen consumption rate in multicellular tumour spheroids* vol. 11, 2014.
[277] A. R. Amin, A. Haque, M. A. Rahman, Z. G. Chen, F. R. Khuri, and D. M. Shin, "Curcumin induces apoptosis of upper aerodigestive tract cancer cells by targeting multiple pathways," *PLoS One,* vol. 10, p. e0124218, 2015.
[278] V. Quaranta, K. A. Rejniak, P. Gerlee, and A. R. A. Anderson, "Invasion emerges from cancer cell adaptation to competitive microenvironments: Quantitative predictions from multiscale mathematical models," *Seminars in Cancer Biology,* vol. 18, pp. 338-348, 10// 2008.
[279] B. Ribba, O. Saut, T. Colin, D. Bresch, E. Grenier, and J. P. Boissel, "A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents," *Journal of Theoretical Biology,* vol. 243, pp. 532-541, 12/21/ 2006.
[280] F. Billy, B. Ribba, O. Saut, H. Morre-Trouilhet, T. Colin, D. Bresch, *et al.*, "A pharmacologically based multiscale mathematical model of angiogenesis and its use in investigating the efficacy of a new cancer treatment strategy," *Journal of Theoretical Biology,* vol. 260, pp. 545-562, 10/21/ 2009.
[281] R.-C. Ignacio, A. J. C. Mark, R. A. A. Alexander, and D. Dirk, "Multi-scale modelling of cancer cell intravasation: the role of cadherins in metastasis," *Physical Biology,* vol. 6, p. 016008, 2009.
[282] X. Sun, L. Zhang, H. Tan, J. Bao, C. Strouthos, and X. Zhou, "Multi-scale agent-based brain cancer modeling and prediction of TKI treatment response: incorporating EGFR signaling pathway and angiogenesis," *BMC Bioinformatics,* vol. 13, p. 218, 2012.
[283] Z. Wang, L. Zhang, J. Sagotsky, and T. S. Deisboeck, "Simulating non-small cell lung cancer with a multiscale agent-based model," *Theor Biol Med Model,* vol. 4, p. 50, 2007.
[284] L. Zhang, C. A. Athale, and T. S. Deisboeck, "Development of a three-dimensional multiscale agent-based tumor model: Simulating gene-protein interaction profiles, cell phenotypes and multicellular patterns in brain cancer," *Journal of Theoretical Biology,* vol. 244, pp. 96-107, 1/7/ 2007.
[285] S. Benzekry, G. Chapuisat, J. Ciccolini, A. Erlinger, and F. Hubert, "A new mathematical model for optimizing the combination between antiangiogenic and cytotoxic drugs in oncology," *Comptes Rendus Mathematique,* vol. 350, pp. 23-28, 1// 2012.
[286] H. B. Frieboes, M. E. Edgerton, J. P. Fruehauf, F. R. A. J. Rose, L. K. Worrall, R. A. Gatenby, *et al.*, "Prediction of Drug Response in Breast Cancer Using Integrative Experimental/Computational Modeling," *Cancer Research,* vol. 69, pp. 4484-4492, 2009.
[287] B. Ribba, T. Colin, and S. Schnell, "A multiscale mathematical model of cancer, and its use in analyzing irradiation therapies," *Theoretical Biology & Medical Modelling,* vol. 3, pp. 7-7, 2006.
[288] H. V. Jain, S. K. Clinton, A. Bhinder, and A. Friedman, "Mathematical modeling of prostate cancer progression in response to androgen ablation therapy," *Proceedings of the National Academy of Sciences,* vol. 108, pp. 19701-19706, 2011.

[289] F. Yang, H. S. Oz, S. Barve, W. J. de Villiers, C. J. McClain, and G. W. Varilek, "The green tea polyphenol (-)-epigallocatechin-3-gallate blocks nuclear factor-kappa B activation by inhibiting I kappa B kinase activity in the intestinal epithelial cell line IEC-6," *Mol Pharmacol,* vol. 60, pp. 528-33, Sep 2001.

[290] S. K. Manna, A. Mukhopadhyay, and B. B. Aggarwal, "Resveratrol suppresses TNF-induced activation of nuclear transcription factors NF-kappa B, activator protein-1, and apoptosis: potential role of reactive oxygen intermediates and lipid peroxidation," *J Immunol,* vol. 164, pp. 6509-19, Jun 15 2000.

[291] L. F. Young and K. R. Martin, "Time-dependent resveratrol-mediated mRNA and protein expression associated with cell cycle in WR-21 cells containing mutated human c-Ha-Ras," *Mol Nutr Food Res,* vol. 50, pp. 70-7, Jan 2006.

[292] Q.-B. She, A. M. Bode, W.-Y. Ma, N.-Y. Chen, and Z. Dong, "Resveratrol-induced Activation of p53 and Apoptosis Is Mediated by Extracellular- Signal-regulated Protein Kinases and p38 Kinase," *Cancer Research,* vol. 61, pp. 1604-1610, February 2, 2001 2001.

[293] W. A. Freed-Pastor and C. Prives, "Mutant p53: one name, many proteins," *Genes & Development,* vol. 26, pp. 1268-1286, June 15, 2012 2012.

[294] Y. C. Henderson, E. Wang, and G. L. Clayman, "Genotypic analysis of tumor suppressor genes PTEN/MMAC1 and p53 in head and neck squamous cell carcinomas," *The Laryngoscope,* vol. 108, pp. 1553-1556, 1998.

[295] D. Stepnick and D. Gilpin, "Head and Neck Cancer: An Overview," *Seminars in Plastic Surgery,* vol. 24, pp. 107-116, 2010.

[296] S. E. Scott, E. A. Grunfeld, J. Main, and M. McGurk, "Patient delay in oral cancer: a qualitative study of patients' experiences," *Psycho-Oncology,* vol. 15, pp. 474-485, 2006.

[297] D. Colin, E. Limagne, S. Jeanningros, A. Jacquel, G. Lizard, A. Athias, *et al.*, "Endocytosis of resveratrol via lipid rafts and activation of downstream signaling pathways in cancer cells," *Cancer Prev Res (Phila),* vol. 4, pp. 1095-106, Jul 2011.

[298] M. Masuda, T. Wakasaki, S. Toh, M. Shimizu, and S. Adachi, "Chemoprevention of Head and Neck Cancer by Green Tea Extract: EGCG-The Role of EGFR Signaling and & "Lipid Raft"," *Journal of Oncology,* vol. 2011, p. 7, 2011.

[299] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat Commun,* vol. 5, 06/03/online 2014.

[300] X. Gao, Y. Cui, R. M. Levenson, L. W. K. Chung, and S. Nie, "In vivo cancer targeting and imaging with semiconductor quantum dots," *Nat Biotech,* vol. 22, pp. 969-976, 08//print 2004.

[301] N. Portney and M. Ozkan, "Nano-oncology: drug delivery, imaging, and sensing," *Analytical and Bioanalytical Chemistry,* vol. 384, pp. 620-630, 2006/02/01 2006.

[302] L. Ye, K.-T. Yong, L. Liu, I. Roy, R. Hu, J. Zhu, *et al.*, "A pilot study in non-human primates shows no adverse response to intravenous injection of quantum dots," *Nat Nano,* vol. 7, pp. 453-458, 07//print 2012.

[303] J. Xu, S. Müller, S. Nannapaneni, L. Pan, Y. Wang, X. Peng, *et al.*, "Comparison of Quantum Dot Technology with Conventional Immunohistochemistry in Examining Aldehyde Dehydrogenase 1A1 as a Potential Biomarker for Lymph

Node Metastasis of Head and Neck Cancer," *European Journal of Cancer*, vol. 48, pp. 1682-1691, 02/15 2012.

[304] J. Xue, H. Chen, L. Diao, X. Chen, and D. Xia, "Expression of caveolin-1 in tongue squamous cell carcinoma by quantum dots," *Eur J Histochem*, vol. 54, p. e20, 2010.

[305] L. Stanberry, G. Mias, W. Haynes, R. Higdon, M. Snyder, and E. Kolker, "Integrative Analysis of Longitudinal Metabolomics Data from a Personal Multi-Omics Profile," *Metabolites*, vol. 3, pp. 741-760, 2013.

[306] R. Shukla, N. Chanda, A. Zambre, A. Upendran, K. Katti, R. R. Kulkarni, *et al.*, "Laminin receptor specific therapeutic gold nanoparticles (198AuNP-EGCg) show efficacy in treating prostate cancer," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 12426-12431, July 31, 2012 2012.

[307] J. C. Tang, H. S. Shi, L. Q. Wan, Y. S. Wang, and Y. Q. Wei, "Enhanced antitumor effect of curcumin liposomes with local hyperthermia in the LL/2 model," *Asian Pac J Cancer Prev*, vol. 14, pp. 2307-10, 2013.

[308] C. D. Kaddi, J. H. Phan, and M. D. Wang, "Computational nanomedicine: modeling of nanoparticle-mediated hyperthermal cancer therapy," *Nanomedicine (Lond)*, vol. 8, pp. 1323-33, Aug 2013.

VITA

Chanchala D. Kaddi

Chanchala was born in Jackson, Mississippi. She graduated high school in Louisville, Kentucky before coming to Georgia Tech for both undergraduate and graduate studies. She received the B.S. in Biomedical Engineering in 2008 and the M.S. in Electrical and Computer Engineering in 2014, and defended her Ph.D. dissertation in Bioengineering in 2015. When she is not working on her research, she enjoys spending time with her family and friends, reading (both fiction and non-fiction), and playing the piano.