BAYESIAN FRAMEWORKS FOR PARSIMONIOUS MODELING OF MOLECULAR CANCER DATA

by

Arturo López Pineda

B. Sc. in Computer Science, Tecnológico de Monterrey, 2006

M. Sc. in Intelligent Systems, Tecnológico de Monterrey, 2008

M. Sc. in Biomedical Informatics, University of Pittsburgh, 2012

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Arturo López Pineda

It was defended on

December 1, 2015

and approved by

Vanathi Gopalakrishnan, Ph.D.
Associate Professor of Biomedical Informatics, University of Pittsburgh

Shyam Visweswaran, M.D., Ph.D.
Associate Professor of Biomedical Informatics, University of Pittsburgh

Gregory F. Cooper, M.D., Ph.D.
Professor of Biomedical Informatics, University of Pittsburgh

Claudia Rangel Escareño, Ph.D.
Professor of Computational Genomics, National Institute of Genomic Medicine (Mexico)

Dissertation Advisor: Vanathi Gopalakrishnan, Ph.D.

BAYESIAN FRAMEWORKS FOR PARSIMONIOUS MODELING OF MOLECULAR CANCER DATA

Arturo López Pineda, M.S.

University of Pittsburgh, 2015

BAYESIAN FRAMEWORKS FOR PARSIMONIOUS MODELING OF MOLECULAR CANCER DATA

Arturo López Pineda, M.S.

University of Pittsburgh, 2015

In this era of precision medicine, clinicians and researchers critically need the assistance of computational models that can accurately predict various clinical events and outcomes (e.g,, diagnosis of disease, determining the stage of the disease, or molecular subtyping). Typically, statistics and machine learning are applied to 'omic' datasets, yielding computational models that can be used for prediction. In cancer research there is still a critical need for computational models that have high classification performance but are also parsimonious in the number of variables they use. Some models are very good at performing their intended classification task, but are too complex for human researchers and clinicians to understand, due to the large number of variables they use. In contrast, some models are specifically built with a small number of variables, but may lack excellent predictive performance.

This dissertation proposes a novel framework, called Junction to Knowledge (J2K), for the construction of parsimonious computational models. The J2K framework consists of four steps: filtering (discretization and variable selection), Bayesian network generation, Junction tree generation, and clique evaluation. The outcome of applying J2K to a particular dataset is a parsimonious Bayesian network model with high predictive performance, but also that is composed of a small number of variables. Not only does J2K find parsimonious gene cliques, but also provides the ability to create multi-omic models that can further improve the classification performance. These multi-omic models have the potential to accelerate biomedical discovery, followed by translation of their results into clinical practice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ALGORITHMS

ACKNOWLEDGEMENTS

Dedicated to my parents, Puri and Serafin. Their infinite love and support allowed me to pursue my dreams. Also to my brother Serafin, for his encouragement in pursuing my career in the United States.

I express my deepest gratitude to my advisor, Dr. Vanathi Gopalakrishnan. Her advice and support guided me through the realms of research. She has provided the tools and skills for starting a career as an independent researcher.

I thank my committee members, Dr. Shyam Visweswaran, and Dr. Gregory F. Cooper from whom I appreciate their scholarly support and academic advice; and Dr. Claudia Rangel Escareño, who kindly introduced me to the genomics community in Mexico through her laboratory.

I thank all my professors, especially Dr. Rich Tsui, my master's research advisor. His admirable patience in teaching was instrumental when I started in this field. I am also thankful to Dr. Gerry P. Douglas, from whom I learned the value of global health informatics.

I cannot thank Ms. Toni Porterfield enough. It is because of her affectionate and caring support that I had an amazing time in the University of Pittsburgh.

A special recognition to my friends and classmates for the long hours of discussion and their valuable ideas. Particularly, I thank Dr. Fernando Suárez Obando, Dr. Charalampos S.

Floudas, Mr. Lucas Santana dos Santos, Mr. Jeya Balasubramanian, Mr. Henry Ato Ogoe, Mr. Victor Ruiz Herrera, Dr. Sergio Castro Díaz, and Mr. Jose Posada Aguilar.

I acknowledge the financial support of The International Fulbright Science and Technology Award. I especially thank Mr. Vincent Pickett, Ms. Sarah Boeving, Mr. Justin Van Ness, Ms. Jordana Berres Paul, and all the Fulbright S&T fellows. Also, the support from the U.S.–Mexico Commission for Educational and Cultural Exchange, in particular from Dr. Arturo Borja Tamayo, Ms. Michelle Ceballos García, and Ms. Marcela Cruz Caballero. I am very grateful for the financial support provided by from The National Council of Science and Technology of Mexico, and the University of Pittsburgh's Specialized Program of Research Excellence in Lung Cancer.

Finally, I want to acknowledge the support of my friends and family in Mexico, as well as the friends that I have made in the United States. They have been a valuable source of learning and provided a welcoming and nurturing atmosphere. I will always keep them in my heart.

1.0 INTRODUCTION

In medicine, cancer is one of the leading causes of morbidity and mortality worldwide. According to the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO), in 2012 there were approximately 14,000,000 new cancer cases, and 8,200,000 cancer related deaths (World Health Organization 2012). The WHO World Cancer Report 2014 (Stewart & Wild 2014) highlights the importance of cancer control programs given the high burden of disease that cancer represents. For example, one in four deaths in the United States is due to cancer (R. Siegel et al. 2014). Globally, breast cancer is the leading cause of cancer death in females; similarly, lung cancers are the leading cause of cancer death in males. Cancer incidence rate in developed countries is double of that in developing countries, and survival tends to be poorer in developing countries (Jemal et al. 2011). By 2030, the WHO estimates that the global number of cancer cases will rise 69% to 21,000,000, and the number of cancer deaths will rise 72% to 13,000,000 (Zarocostas 2010), primarily because of population growth. However, the American Cancer Society (ACS) also estimates that there has been a decline in cancer deaths since 1991, mainly due to early detection and improved treatments.

Early detection of cancer, also known as screening, has been shown to be associated with reduction in mortality. For example, in breast cancer the reduction in mortality is approximately 20% (Myers et al. 2015) when following the screening recommendations of the ACS. These guidelines include non-invasive methods like self-examination, clinical imaging (X-ray, CT-

scan, MRI). However, each method has its own limitations that make it difficult to identify occult lesions. For example, in a prospective study [3] with 1,909 women, the sensitivity and specificity of screening methods for breast cancer was assessed in the task of tumor identification. In this study, the sensitivity for self-examination was 17.9%, for mammography 33%, and for MRI 79.5%, while the specificity was 98.1%, 95%, and 89.8% respectively. In a different prospective study [4] with 649 women, the sensitivity for mammography was 40%, and for MRI was 77%, while the specificity was 93% and 81% respectively. Overall, the specificity of all these methods is much higher than the sensitivity; hence, the adverse consequences for the individuals that are falsely classified need to be addressed, and new diagnostic methods are needed to improve both the sensitivity and specificity of diagnosis.

Patients with positive screening test results would be given a diagnostic test. These tests are used to determine the presence and severity of cancer. Diagnostic tests also are used to gather more information about the cancer to guide decisions about treatment. Diagnosis of cancer is done with the use of invasive methods, like biopsy or surgery, where a small sample of tissue is extracted from the patient to be analyzed by a pathology laboratory. The analysis includes the inspection of the tissue under a microscope to search for differential cellular structures. Other studies such as immunohistochemistry (IHC) can be used to detect protein receptors in the cells of a tissue, i.e., in breast cancer ER, PR, and HER2.

Clinicians and researchers could benefit from precise diagnostic capabilities using computational analysis of genomic, proteomic, and epigenomic data [13]. For example, the clinical impact of genomic testing in cancer was recognized by the American Society of Clinical Oncology (ASCO) in 2010, when it updated its policy on diagnostic testing to highlight the importance of assessing the presence of BRCA1 mutation [9] for breast cancer diagnosis. In

recent years, high-throughout “omic” methods have allowed the identification of groups of molecular biomarkers for breast cancer, leading to the development of genomic assays that are used in the clinical practice. For example, PAM50 (Parker et al. 2009) is a 50-gene classifier that improves significantly the subtype prediction of breast cancer subtypes, leading to better prognosis. Oncotype DX® (Lyman et al. 2007) is a 21-gene classifier for the risk of recurrence in estrogen receptor-positive women with early-stage breast cancer receiving tamoxifen.

Although previous genomic assays have been shown to be highly sensitive and specific when searching for the intrinsic subtypes of cancer, the development of these assays is still a lengthy process (Bastien et al. 2012). Identifying these classifiers from high-throughput data is still an open problem. In 2001, the value of network thinking was already recognized to be essential to science, since there is a struggle to interpret the data from genomics at the time (Strogatz 2001). More recently, gene networks have been used to explain the interactions of genes in cancer-related research questions. Some examples of these complex networks can be seen in Figure 1 (Kairov et al. 2012), (Correia et al. 2014), (Qabaja et al. 2014), (J. Liu et al. 2015). Many molecular cancer studies deliver results in the form of a ranked list of gene names, but there is an unsolved problem of using this information in downstream analysis to create a diagnostic or predictive gene signature for a disease (Kairov et al. 2012). The use of networks is a promising tool to represent the association between gene expression and disease (Qabaja et al. 2014). The example networks in Figure 1 are not classification models, because they only represent relationships between molecular elements, but they do not depict the disease-class directly.

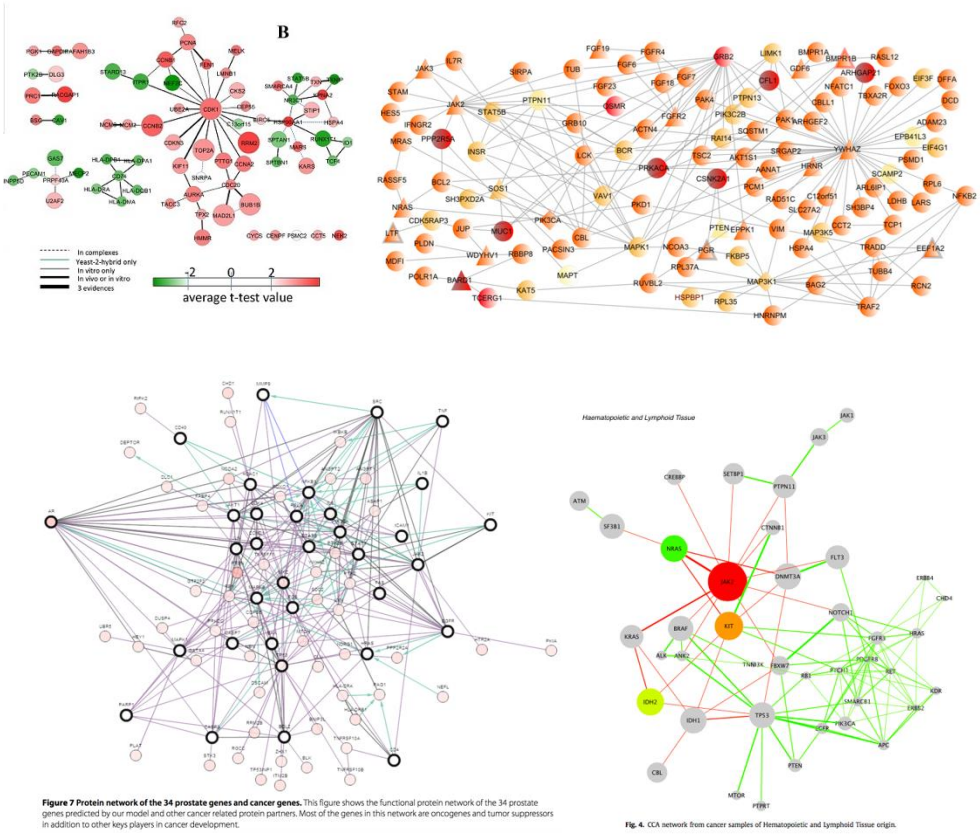

Figure 1. Examples of complex networks.
In top-left Kairov et al. (2012), in top-right Correia et al. (2014), in bottom-left Quabaja (2014), and in bottom-right Liu et al. (2015).

1.1 THE PROBLEM

There is a critical need to build disease models that are useful in clinical practice for screening, diagnosis, monitoring, or prognosis of cancer patients. However, there are two main challenges in building those models: 1) finding an efficient method that can build accurate models from molecular cancer data, and 2) interpreting those models (functional analysis and validation of results). The first challenge deals with classification performance of the models, where models should be able to predict the disease state of the patient. The second challenge

deals with the parsimony of the models (number of variables in the model), so that these models can be interpreted and used in downstream translational research.

1.1.1 Modeling molecular cancer data

Analysis of molecular cancer data produced by high-throughput technologies involves the following challenges: 1) dimensionality of data, 2) missing data, 3) discretization, 4) feature selection, and 5) model building. This section explores some of the problems when dealing with these challenges in molecular cancer studies.

Dimensionality of data. Typically, molecular cancer datasets are composed of large number of variables that are on the order of thousands (e.g., Illumina's Infinium Human DNA Methylation 27K has 27,578 variables) to millions of variables (e.g., next-generation sequencing). In contrast, the number of cases and controls in those datasets is relatively small, ranging from tens of samples to a few hundred. The problem of sample availability is usually restricted by the clinical problem being analyzed (i.e., prevalence of disease), economical factors regarding the study (e.g., cost of enrolling patients), and quality of the data (e.g., batch effect or missing clinical information). Given the abundance of variables for each sample, the selection of an appropriate classification method is of critical importance in modeling molecular data from cancer samples.

Missing data. Imputation methods can infer the most probable value for missing information. The effect of imputation results in an slight improvement to the classification in genomic datasets (De Souto et al. 2015). Another approach consists of building computational models that can handle missing data without imputation. Bayesian classifiers have been used to represent the presence or absence of data, leading to information value on its own (J. H. Lin &

Haug 2008). However, in molecular cancer high-throughput technologies, the problem of missing data is becoming less frequent as new sequencing machines are being developed. Nevertheless, in clinical care missing data remains an important challenge to address.

Discretization. The use of an efficient discretization strategy has been shown to have the capacity of improve performance of machine learning classifiers (Dougherty et al. 1995). However, one challenge in genomic datasets is the identification of the appropriate discretization strategy that might later have a downstream effect in the interpretability of the models resulting from this discretization. Modeling an efficient discretization schema will be of critical importance in the selection of a computational model that describes the data.

Feature selection. Most machine learning algorithms would like to avoid some of the problems associated with learning from large numbers of irrelevant features or variables (Saeys et al. 2007). The use of a feature selection method can lead to: a) avoiding overfitting and improving model performance, and b) providing faster and more accurate methods. The challenge of finding an appropriate feature selection mechanism for genomic data is important, since there was a considerable effort taken to obtain the data in the first place. Typically, in genomic data analysis selecting the differentially expressed genes would be used as the default feature selection mechanism. This is a good strategy if the intent is to find those genes that have the largest variations between cases and controls. However, in computational modeling sometimes there is a need to find the group of genes that jointly provide the best separability of the class, and not just a group of genes that individually separate well. Finding the group of genes that performs this task well can be challenging.

Model building. Building computational methods for disease classification of cancer is a task that has accompanied the development of high-throughput technologies. For example

Aliferis et al. (Aliferis et al. 2002) compared several machine learning methods in the task of classifying between lung cancer subtypes. Although the results were promising, achieving an accuracy of up to 89%, the number of variables was not a critical factor for this study (ranging in the order of 388 genes used in the machine learning models). More recently, the same investigators (Statnikov et al. 2013) compared several modern machine learning models using high-throughput data to achieve higher classification performances. Nevertheless, the main objective of both studies was not the number of features used for the high classification.

There is big potential to keep exploring computational models that accelerate translational research, facilitating improved diagnosis and personalized treatment options for patients. For example, a study by Chang and Ramoni (Chang & Ramoni 2009), yielded a 25-gene classifier that can distinguish between lung cancer subtypes with 95% accuracy.

1.1.2 Interpreting the models

Often, computational models built from data are very complex. In most cases it is hard for human researchers to interpret the large number of nodes and connections that those models have. With complex genomic datasets, Bayesian networks create computational models that are difficult to visualize and interpret (Cossalter et al. 2011). Computational models are often very good at performing their intended classification task, but they are far too complex for human researchers to understand. In the genomic data analysis described in Section 1.1, it is illustrated that the computational models created will be further investigated by functional analysis and then validated via experimental validation (in the wet lab). It is a big burden for human researchers to validate a few interesting genes, which involves a significant amount of time, effort and money to perform a validation study of those genes.

Parsimony is a concept that deals with providing the same results with feweer resources. In the case of computational models, it would mean providing the same classification performance as a big model, with fewer variables. There is a critical need to develop models that are (a) computationally efficient as the larger complex models, (b) have less cognitive burden for human users.

1.2 THE APPROACH

This dissertation explores a novel framework to build parsimonious computational models that can be used for translational research in cancer. This framework builds upon various previously developed algorithms, using them in a way that creates a novel data pipeline. The resulting data-driven model cannot be achieved with the use of only one or some of the components. The overall goal is to develop an automated way to create computational models that can perform classification for a specific biomedical task in cancer research. These models should have two important characteristics: 1) be able to achieve high classification performance, and 2) have small number of variables, i.e., be parsimonious.

In particular, this work develops a **post-classification framework called "Junction to Knowledge" (J2K)** that is composed of four elements: a) filtering, b) Bayesian network generation, c) Junction tree generation, and d) clique evaluation.

In the filtering step, all variables are discretized into one or more intervals using Fayyad and Irani's minimum description length principle cut (MDLPC) (Fayyad & Irani 1993), and removing those features that have a single interval. Then, feature selection is performed using the multivariate ReliefF algorithm [23], where the nearest neighbor samples (weighted by distance)

are used to calculate the feature contribution to class separability. The top scoring features are selected from this step.

In the model building step, an augmented naïve Bayesian model (ANB) is built. ANB is a specific type of Bayesian classifier where most variables (nodes) are conditionally dependent (children) of the target node (class variable), but there are also conditional dependencies between the nodes. The Efficient Bayesian Multivariate Classifier (EBMC) [24] is used to build an ANB model. EBMC greedily searches over the subspace of Bayesian networks that best predict the target node. To make the search efficient, it initially starts with an empty model and it identifies the set of nodes that are independent parents of the target and predicts it well. Then, EBMC transforms the temporary network into its statistically equivalent network where the parent nodes become children of the target with arcs between them. It then iterates the search for a new set of parents given the current structure. Finally, it greedily eliminates arcs between the children nodes.

In the Junction tree generation step, the directed Bayesian network from the previous step is used to create an undirected graph. First, the graph is moralized, which adds a connection between every two nodes that have a common children node. Then, the directionality of the network is removed. Triangulation of the network is a process that normally would be part of the Junction tree generation step, but it is not necessary since the network is already triangulated given that the seeded Bayesian network is an ANB. Later, the cliques of the network are found and a network of cliques is built (Junction tree).

In the clique evaluation step, each clique is successively evaluated in the original Bayesian network for classification performance. Each clique that contains the target node is extracted as a sub-network, where the classification performance can be tested for the smaller

network, similarly as it would be done in a complete network. The best performing clique is selected from this step.

Finally, parsimonious models build from J2K can be used to create a multi-omic data integration model. For some cancer-related classification tasks, the use of single-omic datasets does not achieve high classification performance, as for example determining the stage of cancer in a tumor. Hence, the use multi-omic datasets can improve performance.

1.2.1 Thesis

The central thesis of this dissertation is that the J2K framework produces parsimonious computational models with two properties: high classification performance and small number of variables. The parsimonious groups of variables are nodes in a Bayesian network, which represents genes of interest for a particular biomedical question.

Based on the experiments performed on publicly available cancer datasets, the following specific conjectures are investigated:

Claim 1. The components in the J2K framework, and the specific sequence of use, provide a mechanism for the identification of a parsimonious Bayesian network model with high classification performance.

- **a**. MDLPC discretization finds intervals in gene expression or methylation data that can be used to build a parsimonious model, effectively reducing the high dimensionality of data by removing variables with single intervals.
- **b**. ReliefF feature selection facilitates finding groups of genes or methylation sites that discriminate between disease states.

c. EBMC facilitates the search of a Bayesian network that provides a parsimonious classifier.

d. Post classification processing improves the parsimony of a classifier.

Claim 2. The single-omic parsimonious models created by J2K can be used to create multi-omic models.

1.3 SIGNIFICANCE

This section discusses the significance that would follow if the above hypothesis and conjectures are supported by the experimental results. From an informatics perspective, the J2K framework uses existing algorithms to successively transform ‘omic’ data into a computational model. This data-driven approach creates a computationally efficient Bayesian network, which J2K transforms into a parsimonious network with the same computational efficiency as the complete network. The novelty of this approach consists in the way Junction-trees are used to extract a parsimonious model from a larger model. Unfortunately, current frameworks and pipelines produce models that are computationally efficient for a particular classification task, but do not focus on reducing the number of variables that the model uses.

The J2K framework builds Bayesian networks (BNs) (Neapolitan 2012), which traditionally have been used in other domains to perform probabilistic inference; and Junction trees (JTs) (Lauritzen & Spiegelhalter 1988), which have been widely applied to propagate belief over a network and compute exact posterior probabilities (Serang 2014). While there are many computational algorithms that can assist in the creation of Junction trees, their application to finding a parsimonious model is new.

From a biomedical perspective, the immediate impact of a parsimonious model generated by J2K will be the identification of a computational classifier that can be used for translational research. The gene regulatory networks, extracted from a data-driven analysis using the J2K framework, could lead to new findings in cancer research. Three translational research classification tasks can benefit from this research: 1) distinction between tumor and tumor-adjacent normal samples, 2) molecular subtyping, and 3) cancer stage prediction. Building parsimonious models for each of these tasks can facilitate biological understanding by clinicians and researchers. A parsimonious model of genes can be investigated through experimental research in a laboratory and diagnostic tests with small number of genes could be a direct consequence of this study.

1.4 DISSERTATION OVERVIEW

The remainder of this dissertation is organized as follows. Chapter 2 provides relevant background information on the use of parsimonious models for genomic data analysis, with special emphasis in the use of Bayesian networks and Junction trees. Chapter 3 describes the J2K framework in detail, including an extension for application to multi-omic data integration, and provides descriptions of the datasets used. Chapter 4 provides an annotated example using J2K applied to breast cancer data. Chapter 5 presents experiments done to evaluate the J2K framework, that include comparisons to alternative algorithms in each step of the framework. Chapter 6 presents conclusions and discusses future plans.

2.0 BACKGROUND

In this chapter, background on what constitutes a parsimonious model is provided, including some examples of most common methods that are described in the literature. Later, use of Junction trees in biomedical datasets is discussed, to support the idea of using J2K as a tool for finding parsimonious data models. In addition, background on functional modules is provided. Finally, these methods can be applied to the integrative multi-omic approach, which has been recently explored in the literature.

2.1 ANALYSIS WORKFLOW OF MOLECULAR DATA

Allison et al. (Allison et al. 2006) reported a typical workflow for the analysis of microarray data. Their workflow includes five steps: design, preprocessing, inference, classification, and validation of findings. A more recent model by Braun (Braun 2014) depicted a process with six steps: generating high-throughput data, experimental design, gene-level statistical analysis, identifying functional modules, dimension reduction, and pathway analysis. Another example is the process proposed by Karimpour-Fard et al. (Karimpour-Fard et al. 2015) with four steps: observe data and quality control, traditional statistics, dimension reduction with machine learning, and pathway analysis. A combination of the previous genomic analysis workflows can be distilled in the workflow shown in Figure 2, which involves five steps: data

generation (experimental design), preprocessing, or low-level analysis, modeling, functional analysis (pathway analysis), and validation of findings.

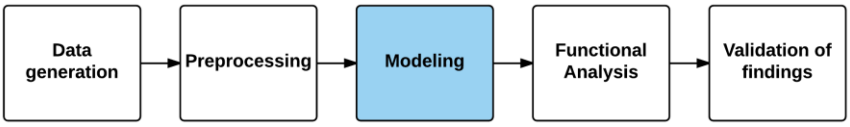

Figure 2. Genomic data analysis workflow

Data generation is the step of the workflow where high-throughput sequencing machines are used to analyze biopsies, and obtain genomic data for analysis. The internal validity of the experiment needs to consider the following concerns: 1) The use of optimal number of replicates to increase stability of the microarray measurements (Zakharkin et al. 2006). Biological replicates are used to address biological variance, and technical replicates are used to avoid measurement error of the assay (Kerr 2003); 2) Increasing the sample size to improve the power of an stratified experiment (C. Wei et al. 2004); 3) Pooling biological samples to increase power (W. Zhang et al. 2007); and 4) Avoiding confounding factors (i.e., patients under different treatment plans).

The **preprocessing** (or low-level analysis) step uses statistical methods to remove systematic variation. These quality-control measures are critically important in any high-throughput study to provide certainty about the results (X. Wang et al. 2003). Robust multiarray analysis (RMA) is the most widely used preprocessing algorithm for Affymetrix and Nimblegen gene expression microarrays (McCall et al. 2012). RMA performs background correction, normalization, and summarization in a modular way. Meanwhile, common methods for RNA-seq normalization include: upper quartile scaling (UQ) (Bullard et al. 2010), trimmed mean of M values (TMM) (Robinson & Oshlack 2010), reads per kilobase of exon model (RPKM) (Mortazavi et al. 2008), remove unwanted variation (RUV) (Risso et al. 2014). A variety of

factors or ‘batch effects’ can contribute unwanted variation to the data, dominating the signal of interest. The choice of normalization method affects the power and reproducibility of the results (Peixoto et al. 2015). Furthermore, there is a need to account for the specific cell composition, given that it has been shown to influence specific high-throughput technologies, e.g., DNA methylation (Maksimovic et al. 2015).

The **modeling** step involves the use of computational approaches to attempt a division of the samples into classes with two approaches: 1) Unsupervised classification (class discovery), where the goal is the identification of novel groups of samples on the basis of their molecular profiles (Hastie et al. 2009); and 2) Supervised classification (class prediction), where the goal is the identification of a minimal set of genes that can be used to categorize well a new sample into one of several known types, based on its molecular profile. This dissertation focuses on supervised classification modeling.

In the **functional analysis** step the objective is to find biological processes of the differentially expressed genes or variables found in the modeling step. There are many tools available that can be used for the biological interpretation of gene lists (Huang et al. 2009). Enrichment analysis is a computational method that determines whether an a priori defined set of genes shows statistically significant differences between two phenotypes. The tools used for enrichment annotation of genes use statistical correlation between the genes and the knowledge-based ontologies available. The annotation can have 1) annotation for individual genes, 2) annotation for the entire list of genes, and/or 3) annotation of modules of genes along with their inter-relationships. Some of the popular knowledge-based functional analysis bioinformatics tools include National Center for Biotechnology Information’s “Database for Annotation, Visualization and Integrated Discovery” DAVID (Huang et al. 2007), Ingenuity Pathway

Analysis IPA®, 3) Illumina®'s NextBio®, and 4) the "Kyoto Encyclopedia of Genes and Genomes" (Ogata et al. 1999).

In the **validation of findings** step, the discoveries and conclusions made in the previous phases can be confirmed by testing them in the laboratory or clinical setting.

2.2 PARSIMONIOUS DATA MODELS

The principle of parsimony, also known as "Ockham's razor" was first introduced by William of Ockham, an English friar, in the 14th century (Guyon et al. 2010) stating in Latin as "*Plurilitas non est ponenda sin necessitate*" (plurality should not be posited without necessity). This is a heuristic approach that searches for a minimal explanation. The term razor refers to distinguishing between two hypotheses either by "shaving away" unnecessary assumptions or cutting apart two similar conclusions.

In machine learning, the parsimony principle states that if two models can adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data (Seasholtz & Kowalski 1993). This property opens new possibilities to search for minimum cost models that can still be meaningful to the data (Goemans & Bertsimas 1993). A compelling argument in favor of parsimony is to reduce the over-fitting of models (Guyon et al. 2010).

The need for parsimonious modeling has been addressed in various fields. In education, parsimonious models are used as tools to help students understand complex concepts by presenting simple models that still hold the truth of the phenomenon being studied (Nelson & A. F. Siegel 1987). In retail, parsimony is used to forecast the revenue of certain products (Sawhney

& Eliashberg 1996); or estimate the an equity index (Arnett et al. 2003). In finance, parsimonious models are built to estimate asset pricing (Guvenen 2009). In telecommunications, parsimonious models are used to facilitate controlling mobile networks (Piorkowski et al. 2009). In industrial engineering, parsimonious models help in determination of levels of merchandise for display in a store (Ho & Chong 2003). In biology, parsimony helps in organism classification (Carpenter 1988). Particularly in genomics, parsimonious modeling has been used to learn gene regulatory networks using biclustering methods (Bonneau et al. 2006).

The application of parsimonious models often seeks to reduce overfitting. However, the parsimonious models in the examples described above also exhibit a common characteristic: the models created are meant for human use or interaction. This human component often means that there will be some interpretation of the model's, that the models will be used in real time applications, or that the final computation of a process requires a decision by a human agent to keep processing. In cognitive psychology, the amount of elements of information that a human can process is typically framed by the heuristic rule of seven plus minus two (G. A. Miller 1956). Therefore, the human interpretation of parsimonious models would likely fall in the size of Miller's law of cognitive information processing, where models would have seven plus minus two variables.

In Machine Learning, the problem of learning a parsimonious model can be done with combinations of: a) preprocessing and feature selection, b) model building, and, c) post-processing. The dimensionality of the variable space is typically reduced by feature selection methods (e.g., ReliefF), many machine learning algorithms also have feature selection embedded (e.g., EBMC), or they have a mechanism to reduce the number of variables (e.g., tree pruning,

backward elimination, regularization). Other alternatives include the use of wrappers or ensemble methods, but these are outside the scope of this dissertation.

2.3 JUNCTION TREES FOR BIOMEDICAL DATA

A Junction tree (Lauritzen & Spiegelhalter, 1988) is a tree-structured undirected graph, whose nodes correspond to cliques of variables, and whose links connect pairs of cliques that have variables in common. A clique is a subset of nodes in an undirected graph where any two nodes are connected by an edge. A Junction tree can serve as the computational structure for belief propagation in a Bayesian network (Pakzad & Anantharam 2005). Graph decomposition is a way to solve inference in large and complex Bayesian networks (Olesen & Madsen 2002), which is the task of updating the probabilities while evidence is being acquired. Message passing is a way to make an efficient inference in Bayesian network (Madsen 2004). This technique creates a secondary structure (Junction trees) that propagates all possible dependence relations. Belief propagation by message passing can converge to optimal solutions in polynomial time (McAuley et al. 2008), although in the worst case scenario this is not true.

The use of Junction trees in biomedical datasets is an emerging field that has yet to be explored. A search in PubMed (MEDLINE) with the keyword ("junction tree/s") retrieved a list of 12 articles, but only half of them deal with biomedical datasets.

Totir et al. (Totir et al. 2009) used Junction trees to efficiently calculate the posterior probabilities of the genotype in cattle pedigree (26 individuals). A monogenic recessive disease affects some members of this pedigree. From the original cattle (3 individuals), the breeding patterns are known for the available three generations, with loops in the breeding pattern.

Junction trees solved the issue of dealing with loops, while facilitating the inference. Similarly, Slooten (Slooten 2011), also used Junction trees for calculating posterior probabilities of human pedigree for recognition of remains in a disaster. Slooten compared the DNA profiles of unidentified individuals with surviving relatives. First, a Bayesian network is constructed to capture the dependencies in the data, and then a Junction tree is built to compute the posterior probabilities given the available data.

Serang and Noble (Serang & Noble 2012) used Junction trees to efficiently identify proteins in a mixture of tandem mass spectrometry. They compared their results with sampling and marginalization techniques, and found that the use of Junction trees is more efficient in time because it increases convergence in the message passing. Three protein datasets (C. elegans, H. influenza, S. cerevisiae) were used in their experiments. Protein mixtures are digested into peptides and separated by hydrophobicity. Each peptide population is fragmented into a tandem mass spectrum. The spectra are matched to a known database to score them for protein inference.

Prutenau-Malinici et al. (Pruteanu-Malinici et al. 2013) describe a graphical model to label gene expression time series images of the Drosophila embryonic development. Junction trees are used to facilitate the inference of annotations in the images. This method, annotate all images at once, instead of individually.

Martini et al. (Martini et al. 2013) analyzed gene sets of chronic myeloid leukemia from the biological pathway Kyoto Encyclopedia of Genes and Genomes (KEGG). First, they selected the differentially expressed genes of the pathway and constructed a Junction tree. Then, they define sub-paths using the structure of cliques and junctions given by the Junction tree. Later, the sub-paths are compared in biological correlation to the disease, and the most relevant is selected.

They validated the experiments with an acute lymphocytic leukemia dataset. The finding of a relevant sub-path was possible due to the use of Junction trees.

2.4 MULTI-OMIC DATA INTEGRATION

Clinicians and researchers could benefit from precise diagnostic capabilities of 'omic' technologies (Coughlin 2014). Integrating multiple 'omic' data types from the same cohort of patients is referred to as multi-omic data integration (Mason et al. 2014). Integrating information from multiple molecular elements of the cell to identify novel targets has the potential to improve the clinical management of cancer (W. Wang, Baladandayuthapani, Morris, et al. 2013). Multi-omic data integration pose challenges of time and effort required for analysis (Palsson & Zengler 2010), i.e., the human genome is estimated to have around 20,000 genes (Ezkurdia et al. 2013), which represent an increasingly large number of variables to be analyzed. Integrating multi-omic datasets increases the time and effort required of analysis and data processing (Palsson & Zengler 2010), but also provides clues for new research topics and has the potential for transforming the biological insight (Payne 2015). Nevertheless, preliminary results show that the integrative approach can offer a more complete picture that can be used for biomarker discovery in cancer (Y. Liu et al. 2013).

Multi-omic data integration could provide a complete picture of the underlying biology of a disease, where researchers need to understand the relationships between different data types merged in a unified model. For example, a recent whole-cell computational model was developed for a human pathogen including its molecular components and interactions (Karr et al. 2012). This model was possible because of the simplicity of the organism, a parasite containing

525 genes. In contrast, the human genome is estimated to have around 20,000 genes (Ezkurdia et al. 2013). There is a need to integrate the phenotype with the various layers of molecular data from the genome to enable personalized medicine. Analyzing the resulting data from sequencing technologies is a research bottleneck in clinical studies.

Multi-omic data integration aims to enable personalized medicine by using information from various molecular elements. Multi-omic data integration occurs when multiple 'omic' technologies are applied to samples from the same cohort of patients, with the purpose being the discovery of novel interactions between distinct molecules, or to improve the diagnostic capabilities of a model. This dissertation focuses on vertical genomic integration, where previous efforts in this area can be divided into two groups: a) mapping of molecular elements to provide a graphical representation of their functional involvement in a phenotype that adds to interactome knowledge, and b) supervised classification to find molecular elements that describe a phenotype. In the first group, multi-omic data integration is achieved by creating a network of interconnected gene-gene interactions from multi-omic data (Moulos et al. 2011); a network of interconnected protein-protein interactions from multi-omic data (Tieri et al. 2014); or a network of interconnected molecular elements based on functional enrichment analysis (Y. Liu et al. 2013; Balbin et al. 2013). In the second group, multi-omic data integration is achieved by vertically merging multi-omic datasets and building machine learning algorithms (Stetson et al. 2014; Jayawardana et al. 2015) or statistical regression (D. Lin et al. 2014) models for classification. All of these previous efforts have shown evidence that the use of multi-omic data integration models is a feasible way to improve phenotype classification. However, our proposed method takes a different approach from previous studies by creating networks of interconnected molecular elements that can be used for classification.

3.0 METHODS

This section describes the datasets that are used in this dissertation, which are from The Cancer Genome Atlas. Also, this section describes the "Junction to Knowledge" (J2K) framework. Every J2K component is described, which are: discretization, feature selection, Bayesian network generation, Junction tree generation, and clique evaluation. Finally, the "Multi-Omic Data Integration" (MODI) framework is described as one potential implementation of J2K.

3.1 DESCRIPTION OF DATA

This section describes how gene expression and methylation data is acquired, which are the focus of this dissertation. The Cancer Genome Atlas is the main source of data for all the experiments in this dissertation. All the subsequent sections of this dissertation use the datasets and platforms mentioned in this section.

3.1.1 Microarray Technology

An emerging diagnostic technology is the use of DNA microarrays to measure specific characteristics of the DNA, analyzing thousands of genes at the same time. Typical analyses

include gene expression and more recently DNA methylation. Since DNA methylation plays a significant role in the regulation of gene expression (Phillips 2008), there is an added value of investigating both data types.

Gene Expression

Gene Expression is a measurement of the abundance of the transcripts (mRNA) of genes in the DNA. This measurement helps understand what cells can do, because genes can encode proteins and dictate cell function. The protein production starts when one strand of DNA is transcribed into RNA, which will later be translated into proteins. The protein-coding regions of the RNA (exons) are spliced together to produce messenger RNA (mRNA). The abundance and types of mRNA molecules reflect the function of a particular cell.

Gene expression microarrays have a collection of microscopic spots that can probe specific DNA locations. Each probe tries to hybridize (or bind) with its corresponding complementary RNA (cRNA). A fluoroluminescent solution is used to label probes depending on the specific probe. Then, the microarray will be scanned to determine the abundance of mRNA that hybridized in that particular probe. If a gene is very active, it will produce more color-labeled molecules of mRNA; while genes that are less active produce fewer labeled molecules; if there are no labeled molecules the gene is inactive.

There are two main vendors that offer gene expression microarray platforms: 1) Affymetrix HT HG-U133A (12,042 variables), and 2) Agilent 244k (17,814 variables). Both platforms are considered in this dissertation, but only one is used at any given experiment. A meta-analysis of cohorts analyzed from different platforms is outside the scope of this dissertation.

Gene expression microarray data has incredible value for computational genomics experiments. For example, in a retrospective study gene expression data was used as a classifier between lung carcinomas. Differentially expressed genes were found between adenocarcinoma (ADC) and squamous cell carcinoma (SCC) yielding a good classification performance (Sanchez-Palencia et al. 2011). This result confirms that the molecular gene expression mechanisms of ADC and SCC are considerably different, and they are involved in immune response, cell signal transduction, metabolism, cell division, and cell proliferation (J. Liu et al. 2014).

DNA Methylation

Methylation is a molecular modification of the DNA that denotes the addition of a methyl group in specific locations of the DNA, typically in cytosine-phosphate-guanine (CpG) sites. DNA methylation status of CpG islands is crucial to understand the epigenetic regulation of genes. It has been observed that hypermethylation of normally unmethylated gene promoter regions has the potential to silence gene transcription (Baylin et al. 2001). At the same time, hypomethylation has also been observed to have a role in the regulation of cancer growth and metastasis (Pufulete et al. 2005).

There main vendor of methylation microarray technology is Illumina, which has two microarray platforms: 1) Infinium HumanMethylation27k Bead Chip (27,578 variables), and 2) Infinium HumanMethylation450k Bead Chip (485,577 variables). The first was deprecated in favor of the second one, however a lot of information is still available in the first platform. Both of them cover 96% of CpG islands, their shores and the regions flanking them. The 450k platform further covers CpG sites outside of CpG islands, non-CpG methylated sites identified in

human stem cells, differentially methylated sites identified in tumor versus normal (multiple forms of cancer) and across several tissue types, CpG islands outside of coding regions, and miRNA promoter regions. Although the 450k-methylation platform is more comprehensive of the regions that it is covering, it can always be scaled down to the 27k-methylation variables covered in the original platform. The rationale for doing this is to be able to jointly analyze the samples processed by both platforms, which means that the number of patients with methylation information can be summed up.

Illumina methylation microarray uses probes to target specific sequences to measure the methylation intensity of a particular DNA site. Using a bisulphite solution the cytosine is converted to uracil, leaving the rest of the residues unaffected. This bisulphite treatment introduces specific changes to the DNA sequence depending on the methylation status of a segment of DNA. Then, two distinct fluorescent dye colors are used to recognize the bisulphite-converted sequences. The microarray scanner averages the signal for methylated and unmethylated dyes to calculate the methylation intensity for a particular probeset. This is referenced as the β-value. A β-value of 0 represents an unmethylated CpG site, while a β-value of 1 represents a fully methylated CpG site.

Methylation microarray data is a relatively new technology that has been shown to has incredible value for finding cancer biomarkers. For example, in lung cancer it is able to classify between normal and tumor samples (Rauch et al. 2012). Also, in breast cancer it has shown to have good classification performance (Szyf 2012). It has been suggested that DNA methylation signatures of cancer should be considered as a potential diagnostic or prognostic biomarker of the disease (Pfeifer & Rauch 2009).

3.1.2 The Cancer Genome Atlas

The United States's National Cancer Institute (NCI) created The Cancer Genome Atlas (TCGA) Research Network with the purpose of collecting and studying information on various human cancers to make them available for facilitating the discovery of molecular signatures. The TCGA website has a data portal (http://tcga-data.nci.nih.gov) to make available the datasets to researchers, clinicians, or any person with an email address.

The TCGA Data Portal is the official storage and access point for all TCGA data and analysis tools. The TCGA's strategy follows standard procedures for the analysis of molecular technologies, including gene expression, DNA methylation, protein levels, as well as the clinical information of the patients involved in the study. Through the TCGA data portal it is possible to download these files and analyze them for specific research questions. Each data type (i.e., DNA methylation) has one file per sample containing the results of the molecular technology (i.e., Illumina microarray) for that sample. Separately, there is another file containing the clinical information of the patients from whom those samples were acquired. A description of the TCGA's data portal and the 'omic platforms used is provided in Appendix I.

Multiple centers and institutions participate in the creation of data in TCGA. Tissue and fluid samples are taken from patients in different locations and then contributed to the TCGA Research Network for cancer diagnosis and analysis.

Tissue Source Sites (TSS). It is an institution that collects samples (tissue, cell or blood) and clinical metadata. There are 1052 TSSs currently contributing samples to the TCGA. Tissue samples are made from 30-100mg (preferably 100mg) of tissue, while blood derived samples are made from 2mls of whole blood, buffy coat, or 15ug DNA.

Biospecimen Core Resource (BCR). It receives the frozen samples from the TSSs. A BCR is a center that ensures sample quality through a standard methodology before performing molecular analysis. The BCR sends the samples to a sequencing center

Genome Sequencing Center (GSC). It is a TCGA center that uses high-throughput methods to identify changes to DNA sequences that are associated with specific cancer types. GSCs receive plated DNA analytes, and corresponding aliquot barcodes.

Genome Characterization Center (GCC). It is a TCGA center that uses high-throughput technologies to analyze genomic changes involved in cancer. The genomic changes that are identified will be further studied by the GSCs. GCCs receive plated DNA/RNA analytes, and corresponding aliquot barcodes. There are 32 GCCs currently processing samples.

Data Coordinating Center (DCC). It is the central provider of TCGA data. The DCC standardizes data formats and validates submitted data. The DCC receives participant information, biospecimen data, clinical pathology data, corresponding TCGA barcodes (across all biospecimen data levels), and tissue slide images. Ultimately, the DCC is responsible for posting the data to the Data Portal.

The TCGA Data Portal is the official storage and access point for all TCGA data and analysis tools. The TCGA's strategy follows standard procedures for the analysis of molecular technologies, including gene expression, DNA methylation, or protein levels, as well as the clinical information of the patients involved in the study. Through the TCGA data portal it is possible to download these files and analyze them for specific research questions. Each data type (i.e., DNA methylation) has one file per sample containing the results of the molecular technology (i.e., Illumina microarray) for that sample. Separately, there is another file containing the clinical information of the patients from which those samples were taken from.

Data Extraction from the TCGA

In data mining, an extraction, transformation, and load (ETL) process is required to build a repository that can be further used to infer new knowledge. ETL has commonly been used in domains with large volumes of information, like patent mining (Diaz Prado et al. 2010); and recently it has been used in the extraction of biomedical data (Saleem et al. 2013). The extraction process from the TCGA is shown in Figure 3.

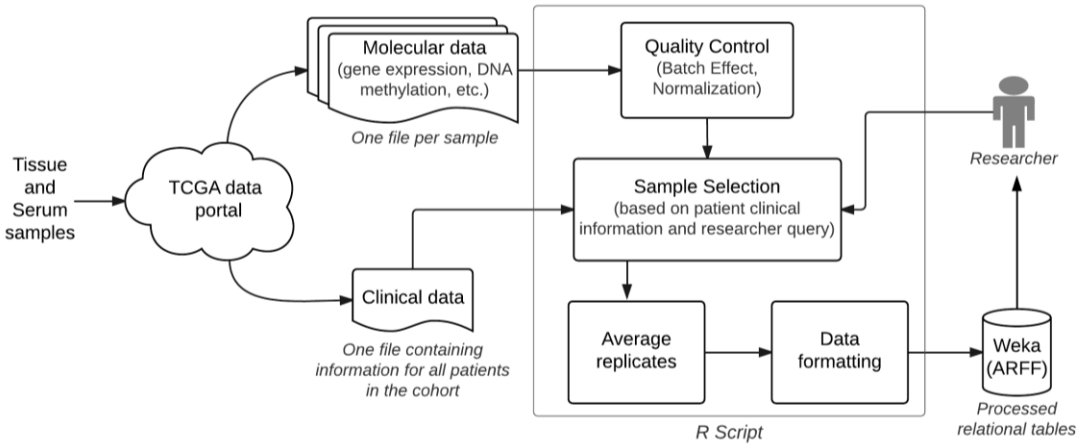

Figure 3. Extraction process from TCGA

Extracting the data. Data extraction, transformation, and load from the TCGA are non-trivial tasks that require reconciling clinical and genomic information. This ETL process has to be customized to the characteristics of the data. The TCGA data portal has a simple interface that requires some basic knowledge of the type of information a person is looking, e.g., the information of gene expression from a cohort of all tumor samples from breast cancer patients. However, one could not specify in the TCGA portal specific details about the patient cohort of interest, e.g., female patients older than 45 years old. In the extraction phase, the TCGA data portal (https://tcga-data.nci.nih.gov) was used. This step only requires an email authentication,

preferably from an academic institution. The selected patient and sample documents can be downloaded from a secure FTP channel.

Transforming the data. In this phase, an R-language script was developed to process the extracted information. This script reconciles the information from the clinical files and 'omic' technologies, and creates a single file with a class label. First, it performs quality control on the data, normalizing across samples, batches and variables. However, this step was omitted since TCGA's level 3 data already covers for data quality. Then, given the specific query of the researcher, the script selects the samples based on the clinical information criteria to have a selection of samples. One potential query could be normal samples and tumor samples; others could be the subtyping of tumor samples, early staging and late staging, benign and invasive tumors. Finally, the script finds the replicates in the cohort and averages them.

Loading the data. In this phase, a script formats the information to become an ARFF file that can be read by the Waikato Environment for Knowledge Analysis (WEKA (Hall et al. 2009)). In this format, it is relatively easy for most researchers to apply machine-learning (ML) algorithms for supervised and/or unsupervised learning. To be able to analyze the 'omic' information for translational research, it is necessary to reconcile it with the clinical information. Patient barcodes are the reference that helps ensuring this task. For example, the reported histological type of breast cancer patients can be "Infiltrating Ductal Carcinoma", "Infiltrating Lobular Carcinoma", "Mucinous Carcinoma", "Medullary Carcinoma", or any of the not specified/available categories. Selecting the appropriate clinical parameters is of great importance. Another example of clinical feature in breast cancer is the selection of sex of the patient. For example, in the TCGA's breast cancer dataset there are 11 male patients out of 1043 total patients. The list of clinical features in the TCGA data portal depends greatly on the cancer

dataset that is being analyzed. It is not the purpose of this dissertation to investigate the selection strategies of those features.

3.1.3 TCGA Datasets

This dissertation uses five cancer datasets from the TCGA: 1) lung adenocarcinoma (The Cancer Genome Atlas Research Network 2014) (LUAD), 2) lung squamous cell carcinoma (The Cancer Genome Atlas Research Network 2012a) (LUSC), 3) breast invasive carcinoma (The Cancer Genome Atlas Research Network, Getz, Saksena, Park, et al. 2012) (BRCA), 4) ovarian carcinoma (The Cancer Genome Atlas Research Network 2011) (OV), and 5) colon adenocarcinoma (The Cancer Genome Atlas Research Network 2012b). The samples for these datasets were contributed from several tissue source sites (TSS), but only specific genome coordination centers (GCC) processed the samples.

In the TCGA pipeline, the University of North Carolina (UNC) is the GCC that analyzed samples with gene expression microarray technologies; while Johns Hopkins and the University of Southern California (JHU_USC) are the GCCs that analyzed samples with DNA methylation microarray technologies. Table 1 shows a breakdown of the number of samples in each dataset, the classification task that is used, and the type of 'omic' technology (gene expression G, or methylation M).

Table 1. TCGA's cancer datasets.

#	Database	TCGA cancer type	G/M	Classification Task	Cases	Controls
A	luad-m-tn	Lung adenocarcinoma	M	Tumor vs Normal	65	24
B	lusc-m-tn	Lung squamous cell carcinoma	M	Tumor vs Normal	132	27
C	lung-g-adsq	Lung carcinomas	G	ADC vs SCC	32	153
D	lung-m-adsq	Lung carcinomas	M	ADC vs SCC	65	132
E	brca-g-tn	Breast invasive carcinoma	G	Tumor vs Normal	1,065	124
F	brca-m-tn	Breast invasive carcinoma	M	Tumor vs Normal	1,065	123
G	brca-g-stage	Breast invasive carcinoma	G	Stage 0-I vs Stage II-IV	92	417
H	brca-m-stage	Breast invasive carcinoma	M	Stage 0-I vs Stage II-IV	184	862
I	ov-g-tn	Ovarian carcinoma	G	Tumor vs Normal	590	8
J	ov-m-tn	Ovarian carcinoma	M	Tumor vs Normal	60	12
K	coad-g-tn	Colon adenocarcinoma	G	Tumor vs Normal	155	19
L	coad-m-tn	Colon adenocarcinoma	M	Tumor vs Normal	166	37

Lung cancer. Lung cancer is the leading cause of human cancer death in the US, with an estimated of over 160 thousand yearly deaths (R. Siegel et al. 2014). Lung cancers can be divided into two major groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC); the latter is further divided into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma (Alberg et al. 2007). Despite extensive research, the mechanisms that lead to these different types of lung cancer remain uncertain. Adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the most common histological subtypes among all lung cancers. Both of them are a form of cancer that develops in the epithelial cells (carcinoma), and belong to the category of non-small cell lung cancer. Lung ADC develops in the glands that secrete products into the bloodstream or some other cavity in the body – the mucus secreting glands in the lungs. Most lung ADC arise in the outer, or peripheral, areas of the lung (College of American Pathologists 2011a). In contrast, lung SCC develops in flat surface covering cells. Squamous cells allow trans-membrane movement, like filtration and diffusion, for example the exchange of air in the alveoli of lungs. Squamous cells can also serve as boundary and protection of various organs. Most lung squamous cell cancers frequently arise in the central chest area in the bronchi

(College of American Pathologists 2011b). Samples in the TCGA data portal are of two types: adenocarcinoma (in LUAD), and squamous cell carcinoma (in LUSC).

Breast cancer. In the TCGA, the available patient samples are included in the Breast Invasive Carcinoma dataset (BRCA). Breast cancer is a global public health concern with 400,000 estimated yearly deaths (Jemal et al. 2011). The clinical impact of genomic testing in breast cancer was recognized by the American Society of Clinical Oncology (ASCO) in 2010, when it updated its policy on diagnostic testing to highlight the importance of assessing the presence of BRCA1 mutation (Robson et al. 2010). In recent years, whole-genome analysis has allowed the identification of groups of molecular biomarkers for breast cancer, leading to the development of genomic assays that are used in the clinical practice. For example, PAM50 (Parker et al. 2009) is a 50-gene classifier that improves significantly the subtype prediction of breast cancer subtypes, leading to better prognosis. Oncotype DX® (Lyman et al. 2007) is a 21-gene classifier for the risk of recurrence in estrogen receptor-positive (ER positive) women with early-stage breast cancer receiving tamoxifen. Although these new genomic assays have shown to be highly sensitive and specific when testing form mutations in cancer (Bastien et al. 2012) there are still open problems to be analyzed.

The information obtained from pathology testing of the excised tissues, from blood tests and imaging studies is used to determine the "stage" of the tumor, according to the 'TNM Staging System' (Compton et al. 2012), which is the most commonly used staging system for breast cancer (E. Miller et al. 2014). This system takes into account the tumor size (T), the presence and the number of positive lymph nodes (N), and the presence of distant metastasis (M), resulting in a staging scale from 0 to IV (Bagaria et al. 2014). The stage of the disease determines the necessity for further treatment and the type thereof: lymph node-negative early

stages of breast cancer (stages 0 and I) may require additional radiotherapy to reduce recurrence, whereas node-positive early stages (some stage II) are commonly treated with prophylactic chemotherapy and radiation to prevent recurrence and metastasis. Locally advanced cancers (stage III and some stage II) and metastatic cancers (stage IV) require invariably the use of chemotherapy. However, the invasive procedures required for accurate breast cancer staging have consequences such as upper limb lymphedema following axillary lymph node dissection. Therefore, researchers have been trying to use less invasive methods, and it is hoped that molecular information will contribute towards this goal (Cyr 2015) and also improve the classification performance of the TNM staging system (Orucevic et al. 2015).

Ovarian cancer. In the TCGA, the available patient samples are included in the Ovarian cancer dataset (OV). Ovarian cancer is one of the major causes of cancer death among women in the United States, accounting for 14,270 estimated deaths in 2014 only (R. Siegel et al. 2014). The standard treatment for ovarian cancer patients include aggressive surgery followed by platinum-taxane chemotherapy. However 25% of platinum-resistant patients will have recurrence (D. S. Miller et al. 2009) of the disease. There is still controversy on the extent of the surgery, both because of recurrence and fertility concerns (Seong et al. 2015). Despite ongoing efforts to develop an effective screening strategy, only 20% of ovarian cancers are diagnosed while they are still limited to the ovaries (Bast et al. 2009). This might be due to the fact that most women report almost no symptom, or symptoms that are similar to gastrointestinal or genitourinary symptoms.

Ovarian cancer develops in the epithelium (cells covering the ovaries), germ line cells (cells that produce the ova) or stromal cells (structural tissue of the ovaries). The majority of ovarian cancers develop in the epithelium. These epithelial ovarian carcinomas have been

suggested to have three distinct phenotypic groups, according to the expression of epidermal growth factor receptor (EGFR), estrogen receptor (ER), progesterone (PR), and human epidermal growth factor receptor 2 (HER2) (Demir et al. 2014). These genes and proteins are also associated with breast cancer subtyping. Nevertheless, additional work is still needed to evaluate the diagnostic capabilities of other genes. One example is the use of Bayesian networks, that have been used to create gene-gene interaction graphs to predict the overall survival of ovarian cancer patients (Q. Zhang et al. 2014).

Colorectal cancer. In the TCGA, the available patient samples are included in the colorectal adenocarcinoma dataset (COAD). Colorectal cancer is the second cause of cancer-related deaths in the United States, with estimated deaths of 50,310 in 2014 alone (R. Siegel et al. 2014). The difference between colon cancer and rectal cancer is primarily the anatomical location, having patient management implications. Colon cancer develops in the large intestine, while rectal cancer develops only in the last centimeters of the colon. Colon adenocarcinoma develops in the inner lining of the intestine, specifically in the cells that secrete mucus to lubricate the colon to facilitate movement. Most colon cancers begin as small polyps in the intestine, and are detected by a colonoscopy. For each detected polyp, an assessment is made as to whether a resection is needed or not, depending on the size of the polyp, optical testing, and patient history (Rex et al. 2011). There are multiple questions that are not yet answered by modern practice, for example what to do with small polyps, the time to next colonoscopy, or the risk of recurrence (Takeuchi et al. 2015). Molecular diagnosis could help understand the biology of the disease to improve its diagnosis. For example, the use of Bayesian models have been shown to identify genes that have the potential to cause colon cancer (Fu et al. 2012). However,

further study on this area is needed to provide computational models that can help in the diagnosis of the disease.

3.2 THE J2K FRAMEWORK

In this dissertation, a novel framework, called "Junction to Knowledge (J2K)", is proposed for the extraction of knowledge from The Cancer Genome Atlas. Its goal is to address the challenges described above and provide a novel application of currently existing algorithms. The J2K framework builds directed acyclic graphs, Bayesian networks, and then it transforms them into undirected graphs, Junction trees. The J2K framework provides a novel way of interpreting computational models to discover biological knowledge.

An analysis of the genomic information from cancer patients can provide new knowledge about the group of genes involved in the disease. This analysis can be performed using publicly available datasets, such as the TCGA, to provide a simpler visualization that allows new hypotheses about the biological functions of the genes in a particular cancer cohort.

The J2K framework, shown in Figure 4, allows the successive manipulation of data into the creation of a human readable graphical model. First, it extracts the TCGA's data and creates processed data. Then, it filters the information by discretizing and feature selecting variables, to reduce dimensionality and create a Bayesian network. Using the directed structure from the Bayesian network it creates an undirected graphical structure in the form of a Junction tree, which is a graphical representation of a network of gene cliques. Each individual clique can be further investigated for accuracy in the classification, and potentially interpreted as functional

modules in the data. An experimental researcher can then use the best ranking clique to test new biological hypothesis.

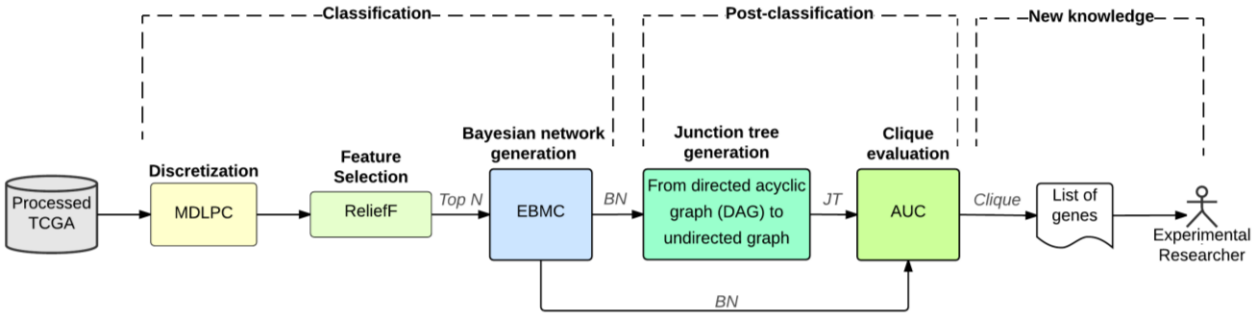

Figure 4. The J2K Framework

3.2.1 Discretization

Genomic microarray data analysis provides information on genes in a continuous manner. Gene expression platforms quantify the concentration of a gene's mRNA transcript in a cell at a given time. The amount of transcripts ranges from zero, when the gene was not expressed, to a few hundred or a few hundreds of thousands, depending on the gene that is being investigated. This variability can have a big impact on the classification performance. A normalization step is often used to reduce the variability in the data using a single reference gene (de Kok et al. 2005). Similarly, DNA methylation microarray technologies measure the amount of methyl groups that can be found at a specific location of the DNA. The technology uses colorization between two probes (green and red) to determine the level of methyl groups that exists in that location. The measurement reflects the intensity at which an optical probe can read the amount of methylation that is present. This value ranges from 0 to 1.

Although continuous values can provide great detail of each microarray measurement, there are some classifiers that require the use of discrete values. Bayesian classifiers use discrete

values to compute the posterior probability of an event to occur given the evidence provided. Even in those cases where Bayesian classifiers are built assuming Gaussian (or other) distributions, there still the need to have cut-points to stablish the occurrence of an event in the Bayes theorem. Partitioning a continuous variable into two or more intervals typically improves classification efforts.

There are some advantages to the use of a supervised discretization method. For example, methylation values ranging from 0 to 1, can be discretized using three potential strategies. In the first strategy, a fix cut-point is determined arbitrarily (i.e., > 0.5 for methylated, and < 0.5 for non-methylated). In a second strategy, an expert-based discretization is made for all variables (i.e., non-methylated < 0.1, partially methylated between 0.1 and 0.8, and methylated > 0.8 (Capra & Kostka 2014)). In the third strategy, a supervised discretization creates independent cut-points for each variable. For the first and second strategies, the same discretization scheme (i.e., same number of intervals or cut-points) is used for all variables. However, this approach is suboptimal for a classification task. For instance, when using MDLPC it is observed that the methylation site cg19782598 was discretized into two categories: methylated (> 0.86) and unmethylated (≤ 0.86); while methylation site cg11693019 was discretized into three categories: methylated (> 0.76), partially methylated (between 0.76 and 0.47), and unmethylated (< 0.47). Thus, supervised discretization could help identify appropriate cut-points for each variable, as opposed to the others, which naïvely assume the same cut-points for variables.

The minimum description length (MDL) principle tries to find a model that facilitates the shortest description of the original data. The length of this description takes into account the description of the model itself and the description of the data using the model.

3.2.1.1 Minimum Description Length with Principle Cut (MDLPC)

MDLPC (Fayyad & Irani 1993) consists of a greedy search method that recursively discretizes each partition. The selected cut point is the one that minimizes the joint entropy of the two resulting subintervals until a stopping criterion based on the minimum description length is met. It is desirable to discard those variables where a cut point was not selected. The minimum description length of an object is described as the minimum number of bits required to uniquely specify that object out of the universe of all objects. Thus, one of the main problems is the selection of cut points. The cut point is a value between two examples of different classes in the sequence of sorted examples. The fact that the algorithm selects cut points that are in the boundary between classes makes the creation of a multi-interval discretization much faster.

3.2.1.2 Minimum Description Length with Kononenko Criteria

Kononenko (Kononenko 1995) developed an algorithm for discretization based on the minimum description length (MDL) principle (M. Li & Vitanyi 2013), and a measure derived from the Relief algorithm (Kira & Rendell 1992). The Relief algorithm estimates the quality of attributes by efficiently dealing with strongly independent attributes. This algorithm searches for the nearest instances from the same class and the nearest instances from different classes.

3.2.2 Feature Selection

Feature selection is the process of identifying the most relevant features (variable-value pairs) for classification. Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. The main objectives of using feature selection include avoiding overfitting, improving prediction performance, providing faster and cost-

effective models, and gaining insight into the processes that generated the data (Saeys et al. 2007).

3.2.2.1 ReliefF Algorithm

ReliefF (Kononenko et al. 1997) is a multivariate filter algorithm that estimates how well a given variable can distinguish the target class given the instances that are near to each other. A matrix filled with zeros (one for each variable) is initially created to represent each variable's score. Then, the algorithm sequentially updates this score and selects the top scoring variables. To update the score, first it selects a random instance (in our configuration all instances are being considered). For this instance, it finds the H nearest hits and the M nearest miss (set to H=M=10 as default value) for each class. The difference between the hits and the misses is being subtracted from the variable's score, and the process is repeated until all instances have being used.

Algorithm 1. Kononenko's Relief

```
1. set all weights W[A] := 0.0;
2. for i := 1 to n do
3. begin
4.    randomly select an instance R;
5.    find nearest hit H and nearest miss M;
6.    for A:= to #all_attributes do
7.          W[A] := W[A] - diff(A, R, H)/n + diff(A, R, M)/n;
8. end
```

3.2.2.2 Information Gain

Information Gain (Quinlan 1986) is a univariate filtering method. It is the amount of information that is lost when a variable X is used to approximate the class variable, defined operationally as the expected extra number of bits required to code samples from the class variable using a code optimized for variable X rather than the code optimized for the class. In

feature selection, a good rule of thumb would seem to be to choose those attributes on which gains the most information.

Equation 1. Information Gain

$$Information\ Gain(class, X) = H(class) - H(class|X_1)$$

Equation 2. Entropy

$$H(X) = -\sum P(X_i) \log_b P(X_i)$$

3.2.2.3 Limma Algorithm

As a way to compare with the feature selection done by ReliefF, the popular bioinformatics tools *limma* (Smyth 2004) was used. It is a tool for gene set analysis (GSEA) that is part of the Bioconductor repository. *Limma* includes functions to fit linear models for each gene given a series of arrays (lmFit) and then compute the log2 differential expression by empirical Bayes moderation of the standard errors towards a common value (eBayes). Although *limma* is not a feature selection algorithm, it has the property of ranking genes based on the eBayes score that each one obtains. The most differentially expressed genes can be selected using this ranking.

In this dissertation, when using features selection, the maximum number of features selected was 30. This number has previously been reported to have a good trade-off between the model complexity and biological relevance of the features chosen (Dudoit et al. 2002).

3.2.3 Building Bayesian Networks

A Bayesian network (BN) (Neapolitan 2012) is a probabilistic graphical model that explains a given set of discrete data. The Bayesian network structure is a directed acyclic graph (DAG) that a set of nodes (random variables) and arcs (probabilistic dependencies). The BN parameters define joint probability distribution over the variables. In the well-known Bayes theorem, the occurrence of an event given some observation can be calculated by the probability of occurrence of that observation given the event, times the probability of the event, and divided by the probability of the observation.

Equation 3. Bayes' Theorem

$$conditional\ probability = \frac{joint\ probability}{marginal\ probability}$$

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Let $X = \{X_1, \ldots, X_n\}$ be a set of discrete random variables. A Bayesian network $B =< G, \Theta >$ is defined by a directed acyclic graph $G =< N, U >$ where N represents the set of nodes (one node for each variable) and U the set of edges, and parameters $\Theta = \{\theta_{ijk}\}$ be the set of conditional probability tables of each node X_i knowing its parents' state P_i.

As an example, consider the Equations shown below (Equation 4) where the event is a patient having the adenocarcinoma (ad) subtype and the observation is a hypothetical $gene_A$ being upregulated. To solve the equations and calculate the posterior probability, various elements would have to be known: 1) the probability of a patient having adenocarcinoma, 2) the probabilities of $gene_A$ being upregulated ($\uparrow_{reg}$) given that the subtype is adenocarcinoma, and 3) the probability that $gene_A$ is upregulated given that the subtype is not adenocarcinoma (in this

case, squamous cell carcinoma). All of these a priori probabilities can be easily derived from the training data.

Equation 4. Bayes Example

$$P(subtype = ad | gene_A = \uparrow_{reg}) = \frac{P(subtype = ad \cap gene_A = \uparrow_{reg})}{P(gene_A = \uparrow_{reg})}$$

$$= \frac{P(gene_A = \uparrow_{reg} | subtype = ad) \cdot P(subtype = ad)}{P(gene_A = \uparrow_{reg} | subtype = ad) \cdot P(subtype = ad) + P(gene_A = \uparrow_{reg} | subtype = sq) \cdot P(subtype = sq)}$$

3.2.3.1 Naïve Bayes

In a naïve Bayes (NB) classifier structure, strong independence between the variables is assumed. In a NB structure, the target node is the parent for all other features, and there are no arcs among those children nodes.

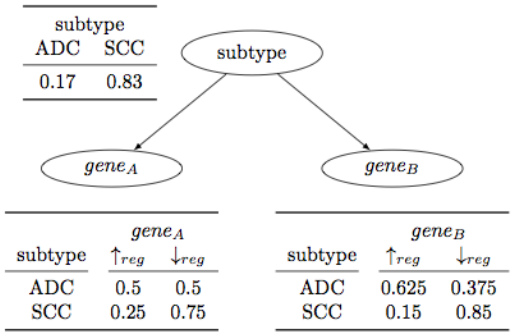

Figure 5. Naïve Bayes network example

Figure 5 shows an NB where two hypothetical genes have been incorporated into the structure. The probability of the maximum likelihood estimate of the subtype being ADC is equal to the number of ADC samples divided by the total number of samples in the dataset. For simplicity of this example, the TCGA's gene expression dataset was used. Then, $P(subtype = ad) = 32/(32 + 154) = 0.17$. Similarly, $P(subtype = sq) = 1 - P(subtype = ad) = 1 - 0.17 = 0.832/(32 + 154) = 0.17$. Next, the a priori probability of the gene being upregulated given the subtype is ad has to be calculated. Let us assume that for the 32 samples that are

adenocarcinoma, $gene_A$ is upregulated in 16 of them $P(gene_A = \uparrow_{reg} | subtype = ad) = \frac{16}{32} = 0.5$, while $gene_B$ is upregulated in 20 of them $P(gene_B = \uparrow_{reg} | subtype = ad) = \frac{20}{32} = 0.625$. That implies that the remaining samples are downregulated ($\downarrow_{reg}$). Lastly, let us assume the a priori probability of $gene_A$ and $gene_B$ being upregulated given the subtype is squamous (sq) to be 0.25 and 0.15 respectively, with the corresponding complementary probabilities for down regulation.

The posterior probability for a new sample of being of subtype adenocarcinoma would be equal to the calculation in the Bayes theorem. Let us assume that for a new sample, both $gene_A$ and $gene_B$ are downregulated, and then the posterior would be equal to the calculations in Eq. 5. In a NB structure the children nodes are independent given the parent, which facilitates the calculation by substituting the joint probability with the product of both probabilities. In this example the probability of the new sample to be of adenocarcinoma subtype is 0.057, which means that probability of it being squamous is 0.943. The NB classifier would call it then a squamous cell carcinoma sample.

Equation 5. Bayes Example Solved

$$P(subtype = ad \mid gene_A = \downarrow_{reg} \cap gene_B = \downarrow_{reg}) = \frac{joint}{marginal}$$

$$= \frac{0.031}{0.561}$$

$$= 0.057$$

$$joint = P(gene_A = \downarrow_{reg} \cap gene_B = \downarrow_{reg} \mid subtype = ad)$$

$$= P(gene_A = \downarrow_{reg} \mid subtype = ad) \cdot P(gene_B = \downarrow_{reg} \mid subtype = ad) \cdot P(gsubtype = ad)$$

$$= 0.5 \cdot 0.375 \cdot 0.17$$

$$= 0.031$$

$$marginal = P(gene_A = \downarrow_{reg} \cap gene_B = \downarrow_{reg} \mid subtype = ad)$$
$$\cdot P(gene_A = \downarrow_{reg} \cap gene_B = \downarrow_{reg} \mid subtype = sq)$$
$$= 0.5 \cdot 0.375 \cdot 0.17 + 0.75 + 0.85 + 0.83$$
$$= 0.561$$

3.2.3.2 Tree Augmented Naïve Bayes

A tree augmented naïve Bayes (TAN) (Friedman et al. 1997) approximates the interactions between attributes by using a tree structure imposed on the naive Bayesian structure. Although naïve Bayesian models have shown to have excellent performance in many datasets, NB has the underlying heavy independence assumption. Augmented Naive Bayes (ANB) classifier appear as a natural extension to the Naive Bayes classifier. It allows relaxing the assumption of independence of attributes given the class variable. TANs are a restricted family of ANBs in which the class variable has no parent and each other attribute has as parents the class variable and at most one other attribute.

3.2.3.3 Efficient Bayesian Multivariate Classifier

Learning the structure of a Bayesian network that explains a given set of data is a difficult task, since the number of possible DAGs for a given number of nodes makes the search task an NP-hard problem (Daly et al. 2011). A heuristic search to find efficient Bayesian network structures is considered a viable alternative. The Efficient Bayesian Multivariate Classifier (EBMC) (Cooper et al. 2010) is a classifier that greedily searches in a subspace of BNs to find the one that best predicts a target node. EBMC efficiently creates Bayesian networks that performs well in high dimensional discrete datasets (Jiang et al. 2014). It initially starts with an empty model and then it identifies a set of nodes that are parents of the target and predicts it well.

EBMC then transforms the temporary network structure into a statistically equivalent one where the parents of the target become children of the target with arcs among them. Then, it greedily eliminates arcs among these children that improve the prediction of the target. It then iterates the whole process until no set of parents (which we can view as a "probabilistic rule" can be added to the target node to improve the prediction of it.

Algorithm 2. EBMC Algorithm: Function EBMC_learn

```
1. var_set S, A; DAG_model Model; Bayesian_network B; Boolean
2. flag;
3. S := set of all variables;
4. flag := true;
5. Model := Ø;
6. while flag
7.    A := Ø;
8.    FindPredictors(Model, A, S, flag, T); //finds a cluster of
      predictors A
9.    if flag
10.         InvertAndPrune(Model, A, T);
11. endwhile;
12. let B be the Bayesian network obtained by parameterizing Model
    using database D
13. return B;
```

Algorithm 3. EBMC Algorithm: Function FindPredictors

```
1. var Boolean flag2;
2. flag2 := true;
3. while flag2
4.    if adding any variable in S as a parent of T in Model
      increases Score(Model, D, T)
5.          Var := Variable that increases Score(Model, D, T) the
            most;
6.          add Var as a parent of T in Model;
7.          add Var to A;
8.    else flag2 := false;
9. endwhile
10. S := S - A;
11. if A := Ø
12.   flag :=false
13. endif
```

Algorithm 4. EBMC Algorithm: Function InvertAndPrune

```
1. var Boolean flag3; node Y; arc Z;
2. remove the Variable in A as being the parents of T in Model;
3. make all Variables in A be children of T in Model;
4. create a saturated set of arcs in Model among the Variables in
   A and let U denote this set of arcs;
6. for each Variable Y in A do
7.    ParentsY := the parents of Y in Model;
8.    flag3 := true;
9.    while flag3
10.         if removing any arc from ParentsY to Y increases
            Score(Model, D, T)
11.               Z := the arc that when removed increases
                  Score(Model, D, T) the most;
12.               remove Z from Model;
13.         else flag3 = false;
14.   endwhile;
15. endfor;
```

This dissertation uses the Waikato Environment for Knowledge Analysis (WEKA) (Hall et al. 2009). I implemented EBMC in the Java programming language to become a WEKA add-on module. The advantages of having all algorithms running in the same framework include the possibility of incorporating other methods in the context of an internal cross-validation, such as discretization (MDLPC) and feature selection (ReliefF).

3.2.4 Creating Junction Trees

A Junction tree (Lauritzen & Spiegelhalter, 1988) is a tree-structured undirected graph, whose nodes correspond to cliques of variables, and whose links connect pairs of cliques that have variables in common. A clique is a subset of nodes in an undirected graph where any two nodes are connected by an edge. Consider the Bayesian network $BN = (G = (V, E), P)$, where G represents a graph with vertices V and edges E, and P represents the set of conditional probability distributions. The vertices V of G correspond to the variables of P. The

transformation of a Bayesian network into a Junction tree $JT = (C, J)$ requires three steps, where C represents the cliques and J represents the junctions:

1. Moralization. The moralization step removes the directionality of the arcs and connects parents with common children. The moral graph G^m of G is obtained by adding undirected edges between all pairs of nodes with a common child, and dropping the direction on all directed nodes. Figure 6 shows an example BN that was moralized. The nodes *C* and *E* are both parents of node *E*, but are not directly connected; therefore, in the moral graph they get a link between them (red line in the graph).

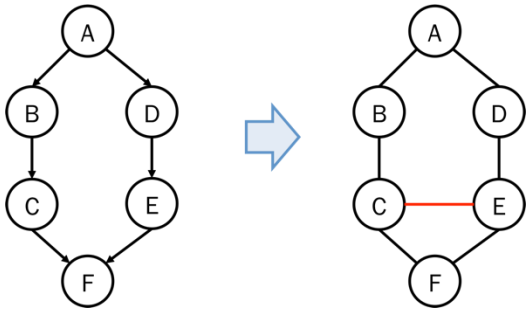

Figure 6. Moralization

2. Triangulation. In the triangulation step the moral graph G^m is triangulated to obtain G^t. A graph is triangulated if every cycle of length greater than 3 has a chord. The process of triangulation is an NP-hard task. This problem can be simplified by selecting an order of elimination. In this dissertation, the order chosen is given by the ranking of features given by ReliefF. A graph is chordal if and only if it has a perfect elimination ordering. A perfect elimination ordering in a graph is an ordering of the vertices of the graph such that, for each vertex, and the neighbors of that vertex in the order form a clique. The ordering from ReliefF might not be an optimal elimination ordering, nevertheless the triangulation solutions chosen follows this ordering.

In Figure 7, an example of the triangulation step is illustrated. The ordering of the nodes follows the alphabet. Originally, there is one cycle of length 5 given by $G^m = \{A, B, C, E, D\}$. For this example, there are 5 possible triangulation solutions; however, the chosen solution (marked in red in the graph) follows an iterative search given the order. The first subset of three nodes $G^m = \{A, B, C\}$ are not originally connected, and therefore they get a new connection to form a triangle. The remaining graph now has a cycle of size 4 given by $G^m = \{A, C, E, D\}$, from which the ordering selected (alphabetical) determines that the new connection to be made should be $G^m = \{A, C, D\}$. The remaining graph is then completely triangulated G^t.

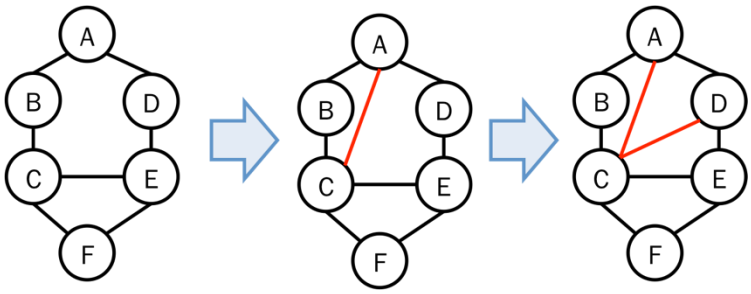

Figure 7. Triangulation

3. Construction of the Junction tree. In the last phase, a junction tree JT is constructed with nodes corresponding to the cliques of G^t, which corresponds to a maximal complete subgraph of G^t. Junctions (separators) connect the cliques of the Junction tree, and the entire graph must hold the Junction tree property. This property states that for two cliques C_A and C_B, that are connected by a path, the intersection $J = C_A \cap C_B$ is a subset of every clique. Figure 8 shows an example of a junction tree, where cliques are indicated using ovals while separators are indicated using boxes.

Constructing the Junction tree follows the elimination algorithm illustrated in Figure 8. First, an elimination order is selected, because this is also an NP-hard problem; depending on the elimination order, a different Junction tree will be obtained. Similar to the triangulation step, the ordering from ReliefF was used (in Figure 8 portrayed alphabetically).

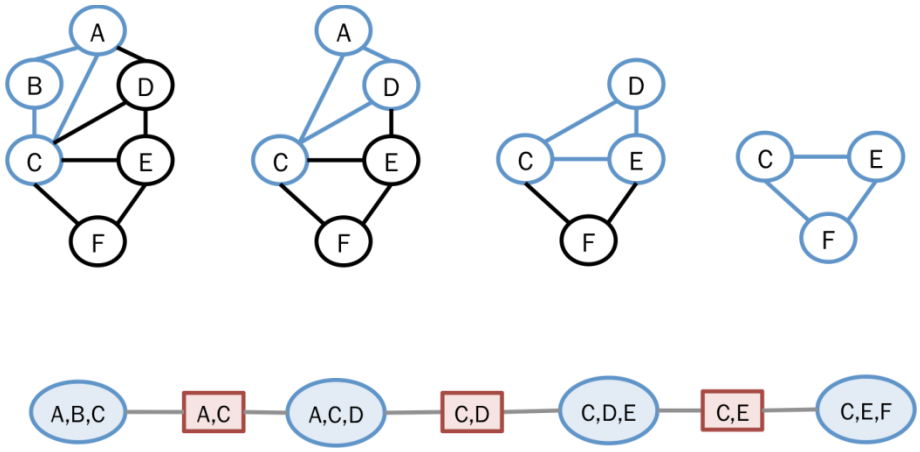

Figure 8. Construction of Junction Tree

The first element in the ordering A is considered. It is part of two possible cliques $C_A = \{ABC, ACD\}$. Because the chosen ordering has node B as its second element, the clique $C = \{ABC\}$ is selected as the first clique (shown in blue in top Figure 8) of JT (bottom Figure 8) and eliminated from G^T. Similarly, the next cliques are sequentially selected and eliminated, until no further nodes remain in G^T, and JT contains all possible cliques.

3.3 THE MODI FRAMEWORK

In this dissertation, a novel framework, called "multi-omic data integration (MODI)", is developed. This framework uses the single-omic parsimonious models created by the J2K framework to create integrated multi-omic models. The MODI framework builds upon the ideas of Wang et al. (W. Wang, Baladandayuthapani, Holmes, et al. 2013) and Wang et al. (W. Wang, Baladandayuthapani, Morris, et al. 2013). In the iNET framework (W. Wang, Baladandayuthapani, Holmes, et al. 2013), multiple models are created (one for each gene), where Bayesian dependencies between gene expression, miRNA and phenotype are explored. In the MODI framework, the relationships from multi-omic elements with the phenotype are used

to create one integrated model that can be used for classification in a whole-genome dataset. In the iBAG framework (W. Wang, Baladandayuthapani, Morris, et al. 2013), a single model is built to consider the interactions between methylation and gene expression. All genes and their corresponding methylation values are connected with two latent variables that represent genes that are methylated and genes that are not. The final classification depends on linear factors that arise between the different layers of the model. In the MODI framework, the known biological regulation that methylation has over gene expression (Phillips 2008) is represented through the use of probabilistic dependencies on the single-omic models. This is an important difference with both iBAG and iNET, because a methylated gene might be part of the methylation-based model, but does not necessarily have to be part of the gene expression-based model. The process of creation of the MODI framework can be seen in Figure 9, which is a model-based integration (Ritchie et al. 2015) because it independently performs analysis on each data type, followed by integration of a resultant model.

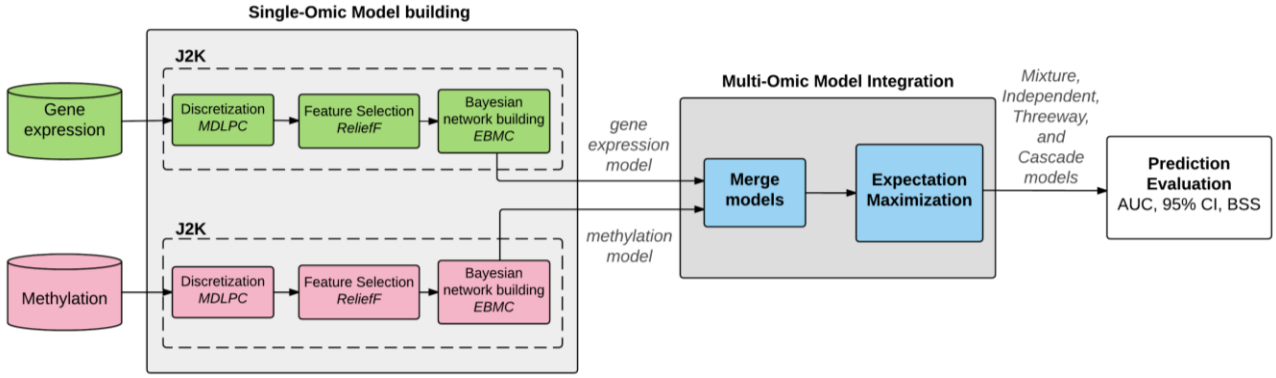

Figure 9. MODI workflow

3.3.1 Integrating Multiple Bayesian Models

The relationships between different data types were modified from Wang et al. (W. Wang, Baladandayuthapani, Holmes, et al. 2013), where models for individual genes were

constructed. In the MODI framework, individual single-omic models are built using the relationships illustrated in Figure 10 and described below.

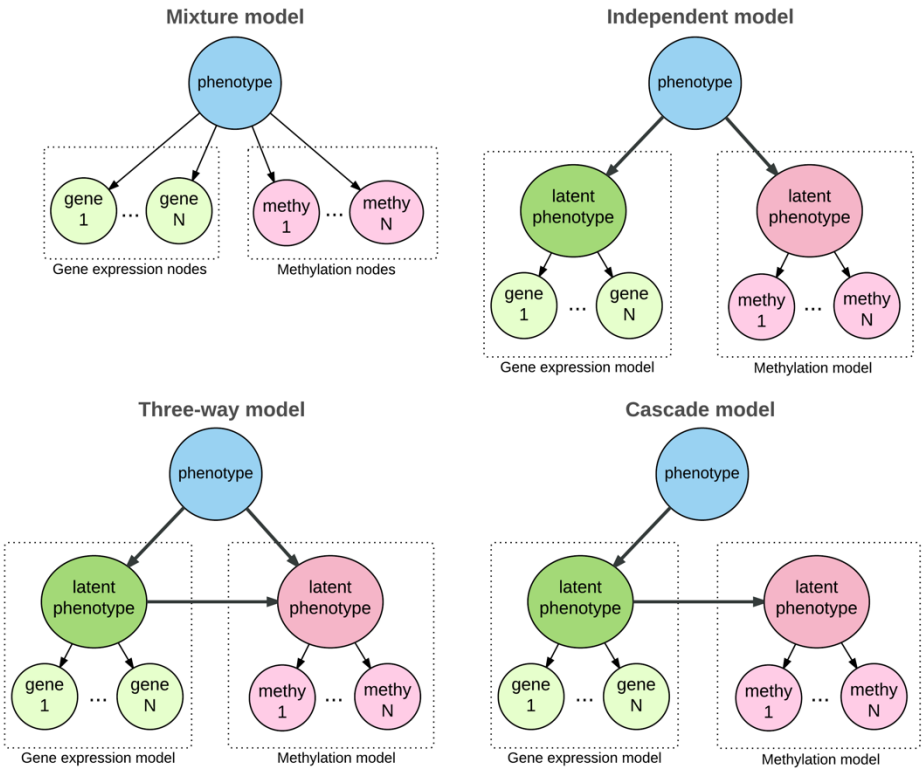

Figure 10. Methods for Bayesian multi-omic data integration
All models are tree-augmented naïve Bayes models (TAN). Gene expression nodes are shaded lighter than methylation nodes. The mixture model might have interactions between gene expression nodes and methylation nodes, while the remaining three models first build single-omic models, and then integrate them using latent variables.

1. *Mixture model*. Gene expression and DNA methylation affect the phenotype, where interactions between gene expression and methylation are possible.
2. *Independent model*. Gene expression and DNA methylation affect the phenotype independently. However, gene expression and DNA methylation are independent, conditioning on the phenotype.
3. *Three-way model*. Both gene expression and DNA methylation affect the phenotype and moreover, gene expression and DNA methylation are dependent, conditioning on the phenotype.
4. *Cascade model*. DNA methylation is correlated with gene expression, which then is correlated with the phenotype. DNA methylation is independent of phenotype, conditioning on gene expression. This relationship is consistent with the underlying biological

mechanisms that DNA methylation has an effect on silencing of the gene expression, which then affects the phenotype.

3.3.2 Latent Variables and the Expectation-Maximization Algorithm

Each single-omic dataset was used to create an augmented naïve Bayesian (ANB) classifiers. All models have a common variable, which is the target node (clinical outcome or phenotype). These single-omic target nodes are considered to be latent (hidden) variables of the multi-omic model, and a new target node is created. Since the creation of latent variables happens after training each single-omic model, the probabilities in both the latent variables and the new target node are not known.

The expectation-maximization (EM (Dempster et al. 1977)) algorithm is used to find the maximum likelihood estimates of the missing parameters in the latent variables. The implementation of the EM algorithm used for this dissertation is in the 'Structural Modeling, Inference, and Learning Engine' (SMILE, (Druzdzel 1999)).

3.4 CLASSIFICATION PERFORMANCE

The evaluation of the models was done using the area under the receiver operator characteristic (AUC) calculated as sensitivity vs. (1 – specificity) , and Brier Skill Score (BSS). The BSS (Wilks 2011) is a measurement of calibration. An ideal BSS is close to 1, while negative numbers indicate models that are less skilled than the weighted dice prediction of 0 (unskilled reference). The Brier Score (BS) in Equation 9 is measured as the average squared

difference between the predicted value y_k and the observed value o_k, with the ideal score being 0 and the worst score being 1. On the other hand the Brier Skill Score (BSS) in Equation 10 is calculated as a scaled representation of the Brier Score relative to the relative frequency of the binary classes or reference Brier Score $\mathrm{BS}_{\mathrm{ref}}$.

Equation 6. Brier Score

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2$$

Equation 7. Brier Skill Score

$$BSS = 1 - \frac{BS}{BS_{ref}}$$

For example, a $\mathrm{BS}_{\mathrm{ref}}$ is equal to 0.098 in a hypothetical test dataset with 9.8% of cases, and let us assume that a hypothetical classification model has a BS of 0.25, then the BSS would be equal to$\mathrm{BSS} = 1 - (0.25/0.098) = -1.55$, which is considered an unskilled prediction. In this sense, it is better to use a BSS because it measures the difference between the score for the prediction and the score for the unskilled reference prediction, normalized by the total possible improvement that can be achieved. The ideal BSS score is1.

4.0 ANNOTATED EXAMPLE

This chapter provides an example of how J2K and MODI create a parsimonious multi-omic model. This example uses the breast cancer dataset with two classes: early stage breast cancer (Stage 0 and I), and advanced stage breast cancer (Stage II to IV). The objective of the modeling task is to capture molecular differences between samples with small tumor size and negative lymph node involvement, and samples with larger tumor size and positive lymph nodes. This task aims to provide better understanding of the genes that are involved in both groups of patients. In the future, it could allow personalized selection of patients that are candidates for sentinel lymph node biopsy and axillary lymph node dissection.

4.1 DATASET

High throughput ‘omic’ technologies are used in the TCGA to process samples from cancer patients across the United States. The output raw data for these technologies is stored in the TCGA data portal. The available breast cancer level 3 data (normalized and aggregated by gene) can be downloaded for research purposes. The gene expression information for each sample, as well as the methylation intensities for each sample, is stored in individual files. Following the process described in Section 3.1.2, these multiple files are converted into a single relational database, where rows correspond to samples, and columns correspond to variables. An

example of such relational databases can be seen in Table 2 (gene expression) and Table 3 (methylation).

Table 2. Raw data example (BRCA gene expression)

Sample	ELMO2	CREB3L1	RPS11	…	class
TCGA-BH-A0AY-01A	0.869	0.878	-0.025		Stage0-I
TCGA-A7-A0DB-01A	0.407	-0.092	0.108		Stage0-I
TCGA-C8-A1HI-01A	0.955	0.529	-0.162		StageII-IV
TCGA-BH-A1F0-01A	-0.290	0.321	0.840		StageII-IV
TCGA-BH-A1EO-01A	0.242	1.060	0.853		StageII-IV

Table 3. Raw data example (BRCA methylation)

Sample	cg00000292	cg00002426	cg00003994	…	class
TCGA-BH-A0AY-01A	0.482	0.159	0.050		Stage0-I
TCGA-A7-A0DB-01A	0.411	0.496	0.066		Stage0-I
TCGA-C8-A1HI-01A	0.863	0.528	0.200		StageII-IV
TCGA-BH-A1F0-01A	0.637	0.272	0.090		StageII-IV
TCGA-BH-A1EO-01A	0.434	0.224	0.077		StageII-IV

In the relational tables, the first column contains the sample ID in TCGA format. This ID can be parsed to extract specific information about the patient (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). All IDs start with 'TCGA', followed by a two-digit alphanumeric code corresponding to the tissue source site (TSS), followed by a four-digit code representing the patient ID, followed by a two-digit numeric code representing the tissue type (e.g., '01' for tumor, '11' for normal), followed by a one-digit letter representing the replicates of that sample (i.e., 'A' means first replica, 'B' means second replica, etc.). The second column and every subsequent column (except the last one) contain variable names and values. For example, in the gene expression platform the variable names correspond to the 17,814 variables in the Agilent platform, e.g., ELMO2, CREB3L1, RPS11, etc. Similarly, in the methylation platform the variables correspond to the 27,578 variables in the Illumina platform, which represent sites of the DNA where there are CpG islands, usually in promoter regions of

genes. The last column represents the class, which in the current example was obtained from the clinical description file for each patient, that is a different file obtained from the TCGA data portal. In the classification tasks of tumor vs. normal the class was obtained from the first column where the sample type is contained. Tables 2 and 3 contain only tumor samples (code '01'), but the selection of samples could have included other sample types depending on the classification task.

Finally, the dataset is transformed into the Attribute-Relation File Format (ARFF, http://www.cs.waikato.ac.nz/ml/weka/arff.html), which is the format used by the WEKA machine learning platform (used in this dissertation).

4.2 DISCRETIZING WITH MDLCP

The expression values and methylation intensities in Tables 2 and 3 respectively are continuous, ranging $\log_{10}$ scale in the case of gene expression, and from 0 to 1 in the case of methylation. However, the J2K framework is based on Bayesian modeling which requires discretization for these values. The WEKA implementation of the MDLPC algorithm is used to select the best discretization strategy in a univariate and supervised approach. Tables 4 and 5 show an example of the discretized datasets.

Table 4. Discretized data example (BRCA gene expression)

PDCL3	MIER1	PIR	…	class
'(0.9795-inf)'	'(-inf-0.81125]'	'(-1.449278-inf)'		Stage0-I
'(0.9795-inf)'	'(-inf-0.81125]'	'(-inf--1.449278]'		Stage0-I
'(-inf-0.9795]'	'(0.81125-inf)'	'(-1.449278-inf)'		StageII-IV
'(-inf-0.9795]'	'(0.81125-inf)'	'(-inf--1.449278]'		StageII-IV
'(-inf-0.9795]'	'(-inf-0.81125]'	'(-1.449278-inf)'		StageII-IV

Table 5. Discretized data example (BRCA methylation)

cg00955451	cg00962459	cg00970325	…	class
'(-inf-0.239753]'	'(0.457448-inf)'	'(-inf-0.340197]'		Stage0-I
'(-inf-0.239753]'	'(0.457448-inf)'	'(0.340197-inf)'		Stage0-I
'(0.239753-inf)'	'(0.457448-inf)'	'(0.340197-inf)'		StageII-IV
'(0.239753-inf)'	'(-inf-0.457448]'	'(0.340197-inf)'		StageII-IV
'(0.239753-inf)'	'(-inf-0.457448]'	'(0.340197-inf)'		StageII-IV

In the discretized form of the datasets, there are a few values that a variable can take. For example, Table 4 shows that the gene PDCL3 was discretized into two groups, where the cutpoint is 0.09795. This creates a group of values smaller or equal to the cutpoint, labeled '(-inf-0.9795]', and another group whose values are greater than the cutpoint, labeled '(0.9795-inf)'. This process is repeated for all variables, with independent discretization strategies for each, depending on the cutpoints found by MDLPC.

All the variables where MDLPC is not able to find cutpoints are removed from the analysis, since they will not be able to contribute to the classification. For example, in this example the gene ELMO2 was deleted because there was no difference between the values in both classes. The resulting database has 254 variables in the gene expression dataset, and 164 in the methylation dataset.

Discretize:

```
weka.filters.supervised.attribute.Discretize -R first-last
```

Remove Useless:

```
weka.filters.unsupervised.attribute.RemoveUseless -M 100.0
```

4.3 FEATURE SELECTION WITH RELIEFF

Although the number of variables is already in the range of hundreds of variables (instead of thousands), there is no guarantee that this behavior will remain across many datasets. Furthermore, parsimony of the models would not be achieved with the number of variables from the discretization step. As mentioned in Section 2.1, a parsimonious model can be achieved by a combination of feature selection and model searching. The ReliefF algorithm is used to select the top 50 variables that have the highest scoring. In other words, the ReliefF score is an assessment of the usefulness of a variable to predict the class variable. A high ReliefF score is preferred. The WEKA implementation of the ReliefF algorithm is used for this purpose. Tables 6 and 7 show the top scoring variables in the gene expression and methylation datasets respectively.

Table 6. Selected features with ReliefF (BRCA gene expression)

#	Variable	ReliefF score	#	Variable	ReliefF score	#	Variable	ReliefF score
1	LCE3B	0.1377	18	CHSY1	0.0772	35	CX3CR1	0.0554
2	HNRPK	0.1345	19	PGM5	0.0749	36	BCL2	0.0546
3	LIN7B	0.1338	20	OR2W3	0.0726	37	SPACA4	0.0536
4	HOXC5	0.131	21	ELL3	0.0706	38	CIDEA	0.0505
5	NFS1	0.1144	22	SORBS1	0.0695	39	SURF4	0.0503
6	LRP6	0.1112	23	C12orf35	0.0687	40	CAMKK1	0.048
7	BRINP1	0.106	24	AQP4	0.0676	41	TLR3	0.0471
8	SLFN12	0.0993	25	SCARB2	0.0669	42	ZFP36L2	0.0459
9	TLR10	0.0993	26	GJA1	0.0666	43	ZC3H12D	0.0451
10	ULBP2	0.0985	27	CYBRD1	0.0663	44	DSN1	0.0449
11	RWDD3	0.0971	28	GTF3C5	0.0661	45	SRXN1	0.0446
12	CNN3	0.0969	29	NTRK2	0.0659	46	ACOT8	0.0444
13	ZNF37A	0.0946	30	SLC13A4	0.0655	47	AGT	0.0435
14	ELAVL4	0.0913	31	RRM2	0.0613	48	MPHOSPH9	0.0434
15	DNASE1L3	0.0855	32	ASPDH	0.0612	49	ATP1A2	0.0428
16	OR51G1	0.0846	33	SLC9A1	0.0591	50	JMJD6	0.0425
17	C9orf140	0.0785	34	LRWD1	0.0565			

Table 7. Selected features with ReliefF (BRCA methylation)

#	Variable	ReliefF score
1	cg07236190	0.0794
2	cg19226099	0.0485
3	cg06224510	0.0429
4	cg21238818	0.0408
5	cg11630242	0.0396
6	cg15028436	0.0388
7	cg10269439	0.0371
8	cg26389232	0.0371
9	cg13482233	0.0369
10	cg02085507	0.0362
11	cg19404979	0.035
12	cg22730004	0.0342
13	cg27341860	0.0339
14	cg11377136	0.0324
15	cg05674036	0.0318
16	cg17503750	0.0305
17	cg21660392	0.0302
18	cg11653864	0.0297
19	cg26538442	0.0292
20	cg25014318	0.0292
21	cg10107671	0.029
22	cg15481539	0.0289
23	cg19664945	0.0283
24	cg19859270	0.0281
25	cg06539804	0.0275
26	cg27488807	0.0275
27	cg10994126	0.0272
28	cg22637834	0.027
29	cg11494699	0.0263
30	cg13131015	0.026
31	cg12732953	0.0249
32	cg04413397	0.0249
33	cg18986165	0.0241
34	cg27071517	0.024
35	cg11830061	0.0238
36	cg17982102	0.0236
37	cg02311163	0.0235
38	cg04740359	0.0227
39	cg15777781	0.0218
40	cg15742700	0.0215
41	cg13204181	0.0212
42	cg06220755	0.0208
43	cg14409083	0.0207
44	cg25384595	0.0205
45	cg26743024	0.0205
46	cg06207804	0.0204
47	cg07911663	0.0201
48	cg03914397	0.0201
49	cg22930187	0.0198
50	cg10129493	0.0194

In this example, the top 50 variables were selected based on previous experiences. This number of variables has been reported to offer a good trade-off between parsimony, relevance, and complexity of model (Dudoit et al. 2002). Another approach could be based on a greedy search of those variables with ReliefF scorings that are not significantly reduced compared to the top scoring feature. One more approach could use a wrapper mechanism to select the optimal number of features. However, these approaches are not in the scope of this dissertation and therefore the simpler approach was selected.

Feature Selection:

```
weka.attributeSelection.ReliefFAttributeEval -W -M -1 -D 1 -K 10 -A 2
weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 30
```

4.4 MODEL BUILDING WITH EBMC

In the Bayesian framework, there are many algorithms that support the construction of a computational classifier that can handle missing data. However, there are few learning algorithms that also reduce the dimensionality of the variables. EBMC is one algorithm that searchers an efficient Bayesian structure that can classify between two states of disease. As mentioned in Section 3.2.3, the implementation used in this dissertation was developed in Java for WEKA, and it is equivalent to the original implementation. EBMC searches for the best available predictors and the conditional dependencies among them. The resulting structure is a tree augmented naïve Bayes, where all variables are conditionally dependent (children) on the target node, and some of them also have other dependencies. After this process, EBMC also implements forward elimination of the arcs that do not contribute much to the classification. In some cases, this will remove an arc between the target node and some of the predictors. However, even in these cases, in EBMC all nodes are part of the Markov blanket of the target node.

Figures 11 and 12 show the EBMC-derived Bayesian model that was constructed for the breast cancer staging example. Both models were constructed using the same parameters, which includes: a) stating the scoring algorithm, in this case K2, b) defining the number of predictors searched, in this case 30, c) defining the number of potential parents for each node and potential children for each node, for both parameters 30. These numbers were selected to allow for a larger search given the number of predictors. Potentially, this parameter description could end up with a completely connected network of 30 variables.

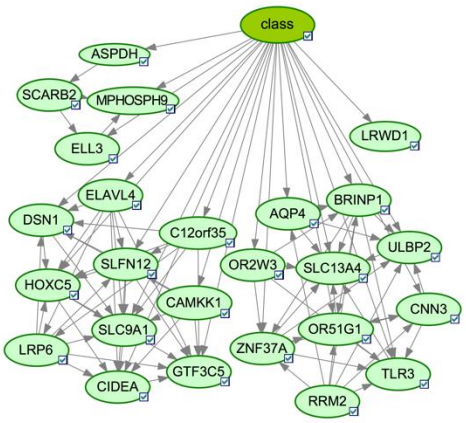

Figure 11. EBMC-derived model (BRCA gene expression)

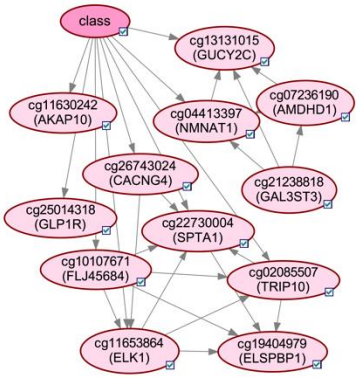

Figure 12. EBMC-derived model (BRCA methylation)

In the model created from gene expression data (Figure 11), there are 25 nodes, from a potential variable pool of 50. In this model, EBMC searched four times for predictors, creating corresponding clusters of up to 10 nodes. Similarly, in the model created from methylation data (Figure 12), there are 12 nodes out of 50. The EBMC search for this model also created clusters that are connected naïvely to the target node. Although the number of variables for both models is small, and significantly reduced from the original number of variables in the high-throughput platforms, the complexity of these models might still pose some cognitive burden for human researchers. Ideally, 7±2 variables would be preferred.

Model Building:

```
weka.classifiers.bayes.BayesNet -D -Q
weka.classifiers.bayes.net.search.local.EBMC -- -T 30 -P 30 -C 30 -S K2 -E
weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
```

4.5 JUNCTION TREE BUILDING

In the J2K framework, a junction tree is built from the Bayesian models created in the previous step. Since EBMC creates an augmented naïve Bayes (ANB) model (or a modified version without some arcs), the junction tree algorithm only needs to moralize the graph, that is to connect nodes with common children node, and remove directionality. The moral graph created from an EBMC-derived Bayesian network is already triangulated, which eliminates this step from the Junction tree algorithm. Finally, all cliques in the network are found to create a junction tree. A clique is a subgraph of the Bayesian network, where all nodes in the subgraph are completely connected. Creating the census of cliques (a list of all possible cliques in the network) is a computationally expensive task in large networks, but the EBMC-derived network is already a small network. The R-package SNA is used to calculate the census of cliques (Makino & Uno 2004).

Figures 13 and 14 show the junction trees created from the EBMC-derived gene expression and methylation datasets, respectively. In Junction tree from the gene expression data (Figure 13), there are 5 cliques (blue circles) and 2 Junctions (red squares). The junctions contain common elements between two cliques. Similarly, the Junction tree from the methylation data (Figure 14) has 8 cliques but only one Junction (with the class).

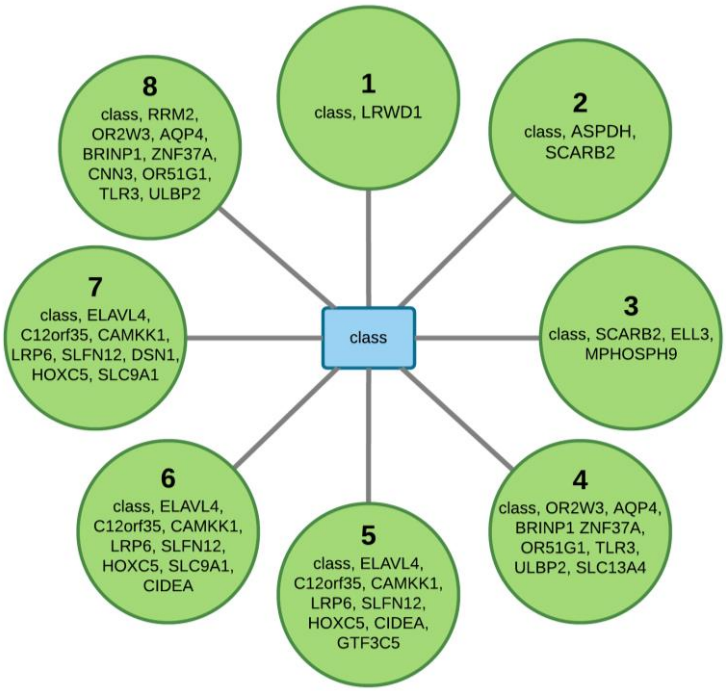

Figure 13. Junction tree (BRCA gene expression)

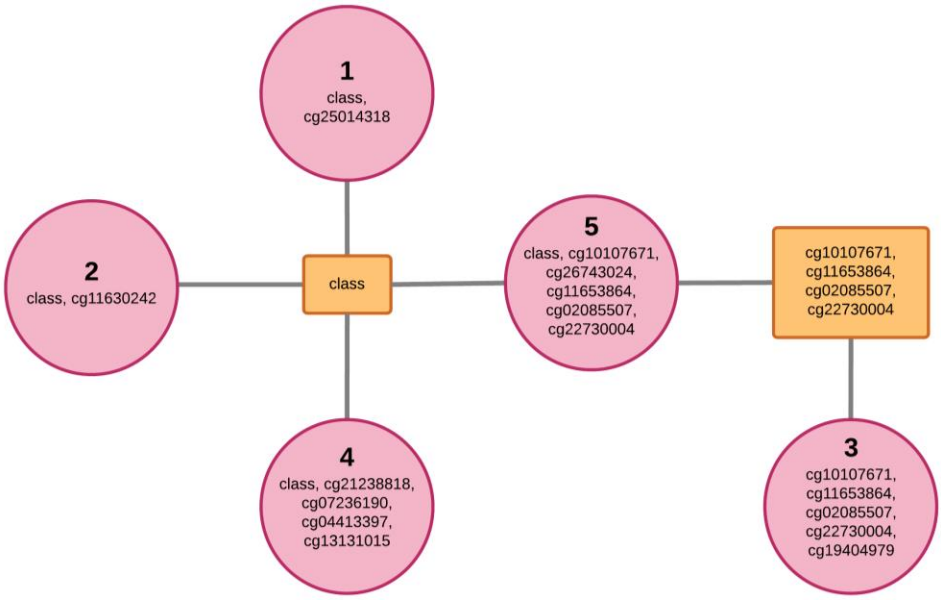

Figure 14. Junction Tree (BRCA methylation)

4.6 CLIQUE EVALUATION

In the junction trees, there is a chance that every clique can become a parsimonious classifier if it satisfies two conditions: 1) the clique contains the class, 2) the classification performance of the clique alone is statistically equivalent to the complete network. In the current example, the Junction tree from the gene expression data has five cliques, where only clique 3 does not satisfy the first condition (does not contain the class). Similarly, the Junction tree from the methylation data has eight cliques, where all contain the class. The second condition can be tested by using the training data from which the model was generated from, evaluating the classification performance (measured by AUC), and comparing this performance to the performance of the complete network. Each clique is evaluated individually, even though there are some nodes that might be part of various cliques. The EBMC model is modified to include only the clique variables, and this model is used to compute classification performance.

Table 8, shows the result of the clique evaluation. In the gene expression data, only clique #5 had a similar classification performance than the complete network. The remaining cliques were significantly worst. In this case, clique #5 can be a parsimonious substitute of the complete network. In the methylation data, only clique #1 had a similar classification performance to the complete network, while the remaining seven cliques had a significantly better classification performance. They would all be considered as potential parsimonious substitutes of the complete network. In this example, clique #8 was selected as the substitute, since it also has the highest AUC from all cliques.

Table 8. Classification performance of cliques (BRCA example)

Omic	Clique	# of variables	AUROC	95% C.I.	*Statistical comparison to complete network* **p-value**
Methylation	Complete	12	0.72	0.68-0.75	1.0
	Clique 1	1	0.61	0.56-0.65	< 0.001
	Clique 2	1	0.61	0.57-0.66	< 0.001
	Clique 3	5	n.a.	n.a.	n.a.
	Clique 4	4	0.64	0.6-0.68	< 0.001
	Clique 5	5	0.71	0.68-0.75	0.93
Gene Expression	Complete	25	0.53	0.51-0.56	1.0
	Clique 1	1	0.63	0.57-0.69	0.004
	Clique 2	2	0.65	0.59-0.7	< 0.001
	Clique 3	3	0.69	0.63-0.74	< 0.001
	Clique 4	8	0.92	0.89-0.95	< 0.001
	Clique 5	8	0.91	0.88-0.93	< 0.001
	Clique 6	8	0.88	0.85-0.92	< 0.001
	Clique 7	8	0.90	0.86-0.92	< 0.001
	Clique 8	9	0.94	0.91-0.96	< 0.001

4.7 MODI FRAMEWORK MODEL

Until this point in the example, each ‘omic’ has been treated independently. Parsimonious single-omic models have been built using the J2K. An extension to the J2K framework is the use of the MODI framework, also covered in this dissertation. MODI integrates single-omic models to create multi-omic models to improve classification performance. The intent of MODI is to provide a comprehensive description of the data given a diverse set of data. In this example only two ‘omics’ are considered (gene expression and methylation), but this idea could be easily extended to incorporate other ‘omics’ and also clinical, and environmental data. As described in Section 3.3, the MODI framework uses a layered approach with latent variables to integrate multiple single-omic models, which are the inputs of the framework. Since every single-omic

model has the same target node, they are all treated as latent variables, and a new target node is created with the same variables. The expectation-maximization (EM) algorithm, is used to compute the probabilities of both the latent variables and the new target node.

Figure 15 shows the MODI created with the two single-omic models, without selecting the best clique, while Figure 16 shows the same MODI created with the best clique scenario (clique #5 in gene expression, and clique #8 in methylation). The interactions from the latent variables (formerly target node in the single-omic models), are modelled to reflect the biology of the 'omics' being considered. Both 'omics' have an effect on the phenotype (target node, or class node), therefore the arcs marked with A and C in Figure 15 are necessary, while arc B reflects the relationship between the 'omics'. In biology, it is known that hyper methylation of the DNA has an effect on the regulation of gene expression.

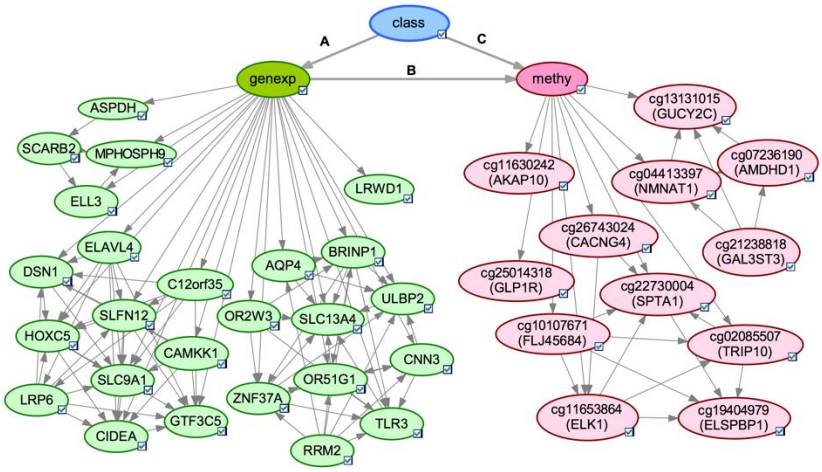

Figure 15. MODI with complete network (BRCA)

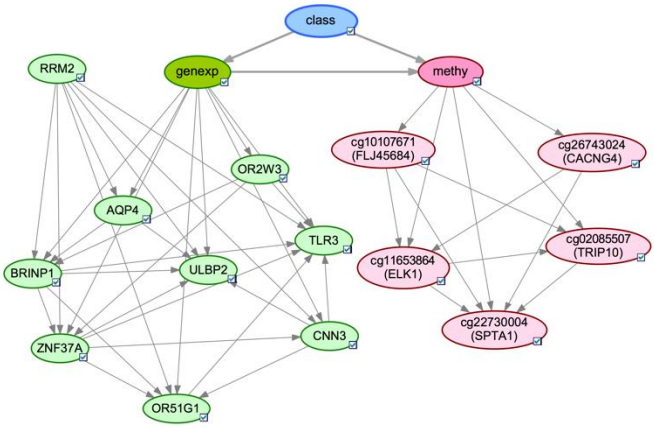

Figure 16. Three-way MODI with best clique selection (BRCA)

Evaluation of a multi-omic network is a difficult task, because it requires a patient to have multi-omic data in the first place. In cancer research, it means that the biopsy should have enough material to be used by the pathology laboratory, and also additional material for each 'omic' machine that will be used. In the future, single-cell technologies could help avoid this inconvenience. In the TCGA breast cancer data, there are enough patients with data in both 'omic' platforms.

Table 9 shows the evaluation for the current example (classification task stage 0-I vs stage II-IV), and another example used as reference (classification task normal vs tumor). The evaluation includes single-omic models (gene model and methylation model), as well as multi-omic models (mixture, independent, three-way, and cascade). All models were created using the J2K framework. The evaluation for the best cliques (parsimonious models) is also provided. These results are based on a stratified random sampling of samples where 70% of them were used for training of the models, and the remaining 30% were used for testing.

Table 9. Classification performance of MODIs (BRCA 70/30)

Classification task	Model	AUC	95% C.I.	Brier skill score
Normal vs. Tumor	Gene model	0.99	(0.99-1.0)	0.64
	Methylation model	0.99	(0.99-1.0)	0.82
	Mixture model	1.0	(1.0-1.0)	0.97
	Independent model	1.0	(1.0-1.0)	0.96
	Three-way model	1.0	(1.0-1.0)	0.97
	Cascade model	1.0	(1.0-1.0)	0.97
Stage 0-I vs. Stage II-IV	Gene model	0.5	(0.37-0.64)	-2.04
	Methylation model	0.54	(0.46-0.62)	-0.47
	Mixture model	0.57	(0.46-0.69)	-2.55
	Independent model	0.8	(0.71-0.89)	0.15
	Three-way model	0.8	(0.71-0.89)	0.15
	Cascade model	0.79	(0.7-0.89)	0.2

The results of Table 9 show that the J2K framework is able to create parsimonious models for molecular classification of cancer. The parsimony of this models is guaranteed by a small number of variables, with at least the same classification performance than the complete models. Furthermore, the MODI models (independent, three-way, and cascade) also improved the classification performance, compared to the single-omic models.

Finally, the MODI shown in Figure 16, created using the complete set of data, shows that the node with the gene RRM2 is spouse of the latent variables (formerly the class node). This node, is a parent of many other nodes, jointly with the gene expression node, which provides an indication that its role in regulating the progression of the disease is important, given the classification task early stage vs late stage. Increased mRNA levels of the gene ribonucleotide reductase M2 (RRM2) have been associated with poor patient outcome in a dose-dependent manner, with prognostic power comparable to that of multiple gene signatures, and superior to TNM stage (H. Zhang et al. 2014), and also with Tamoxifen resistance (Putluri et al. 2014). The findings provided by the J2K framework suggest that there is a molecular change in the

expression of RRM2 between early and late stage patients that regulates other genes in the clique.

The novel knowledge that was derived from the J2K framework was only possible to be observed because of the parsimony of the models. It would be extremely difficult for any human researcher to visually observe that gene from a network with 30 genes, and impossible from models with hundreds or thousands of variables. The strong indication of RRM2 should be validated experimentally in the laboratory, but J2K is already prioritizing a parsimonious list of genes that can be of potential interest.

4.8 PARSIMONY IN J2K

J2K is a framework that facilitates finding parsimonious models with high classification performance and small number of variables. In Figure 17, it is shown the progression these two metrics of parsimony as the J2K successively applies each of its components. The number of variables is reduced significantly from those in a high-throughput dataset into a number that reflects the rule of thumb of the cognitive psychology for human understanding (seven plus minus two). Also, the classification performance is either preserved or improved, when using the post-classification step in J2K. Finally, the new knowledge that is given to the experimental researchers is a list of differentially expressed genes and methylation sites that have a significant role in breast cancer progression.

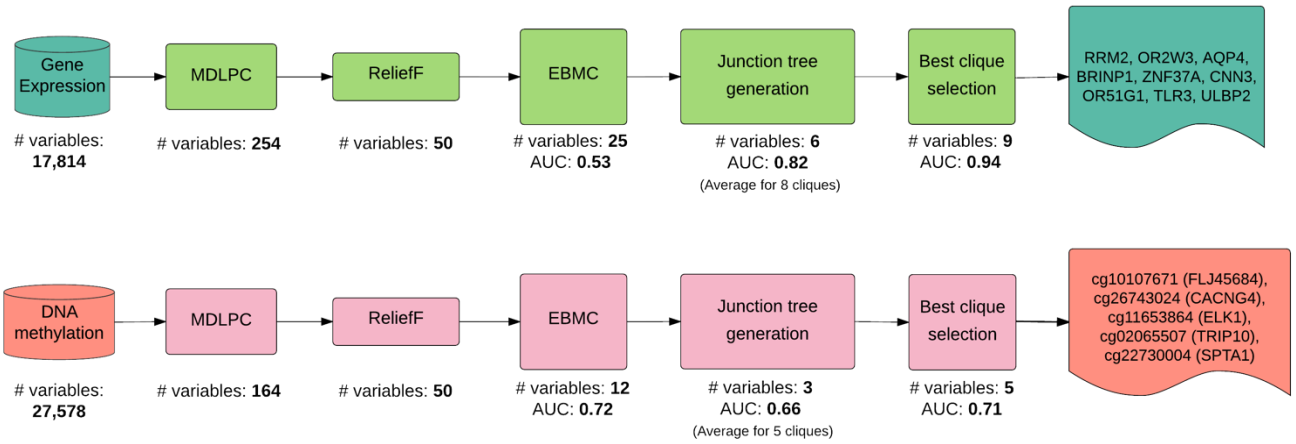

Figure 17. Parsimony in J2K
Top: Gene expression example, Bottom: DNA Methylation example

5.0 EXPERIMENTS AND ANALYSIS

The J2K framework is a novel approach to automate data modeling from genomic datasets. The effectiveness of the framework in finding parsimonious classification models is dependent upon the actual algorithms employed by the framework and the quality of the genomic datasets. For this reason, the empirical effectiveness of the framework under controlled conditions is evaluated here with the data described in Section 3.2. The components of the J2K framework are evaluated following the diagram in Figure 18.

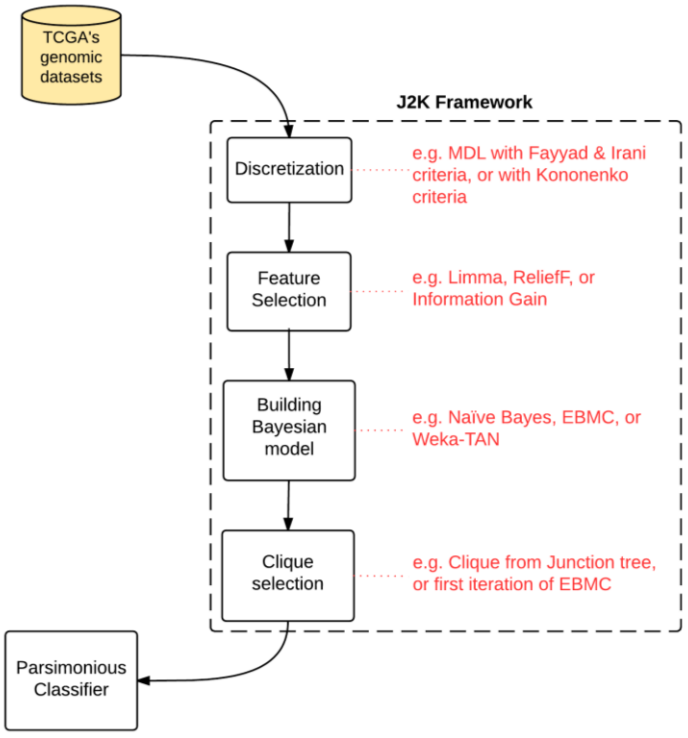

Figure 18. Framework evaluation

The J2K framework uses four components with state-of-the-art algorithms to achieve a parsimonious model: a) Discretization using Minimum Description Length with Fayyad and Irani's Principle Cut criteria (MDLPC) (Fayyad & Irani 1993), or with Kononenko's criteria (Kononenko 1995). b) Feature selection using Kononenko's ReliefF (Robnik-Šikonja & Kononenko 1997), which is compared to Smyth's Limma (Smyth 2004), and Quinlan's Information Gain (Quinlan 1986). c) Building of Bayesian model using Cooper's EBMC (Cooper et al. 2010), and compared to the more traditional naïve Bayesian model (John & Langley 1995), and the TAN algorithm (Friedman et al. 1997). d) Building Junction tree using Lauritzen-Spiegelhalter algorithm (Lauritzen & Spiegelhalter 1988), obtaining the best clique, and comparing with a the first selection of genes done by EBMC. The following sections provide detailed explanations of these comparisons.

5.1 DISCRETIZING CONTINUOUS VALUES

Most 'omic' data, such as gene expression and methylation, are represented with continuous values. However, many machine learning algorithms are designed to handle only discrete (categorical) data, using nominal variables. The reason for that is that discretization, the process of transforming continuous values into discrete ones, has been shown to improve the performance of machine learning classifiers (Garcia et al. 2013).

In the J2K framework, it is necessary to use a discretization algorithm since the selection of a Bayesian framework requires the use of discrete values. The J2K framework uses the minimum description length (MDL) algorithm with Fayyad and Irani's Principle Cut (PC) criterion (Fayyad & Irani 1993). MDLPC is is a well-known algorithm that is frequently used in

machine learning studies with good results (Garcia et al. 2013). The MDL algorithm can also be implemented with the Kononenko criterion (MDLK) (Kononenko 1995), which is being used in this study as a comparison for the PC criterion.

For each of the datasets considered in this dissertation, the experimental design of the discretization experiment uses a 10-fold cross validation to obtain 10 randomly stratified training folds, and their corresponding test folds. Every training fold is independently discretized using both MDLPC and MDLK. It is common that the resulting discretized datasets have a proportion of their original variables to have selected only one bin to which all values in that variables will be assigned, independently of the target variable value. Because this is an irrelevant feature that does not provide new information to the classification, these variables are eliminated, thus contributing to the parsimony of the model.

Table 10 shows for each dataset the number of original variables and the resulting number of variables after discretizing with both methods. The proportion of variables that are eliminated through MDL discretization is the same for both criterions (Fayyad & Irani, and Kononenko) for a given dataset. The differences observed between the numbers of discretized variables in all datasets can be attributed to the difficulty of the classification task, rather than to the discretization method.

Table 10. Experiments: Discretization number of variables.

#	Dataset	# variables (original)	# variables (Fayyad & Irani)	# variables (Kononenko)
A	luad-m-tn	27,578	10,138 (36%)	10,316 (37%)
B	lusc-m-tn	27,578	17,437 (63%)	16,568 (60%)
C	lung-g-adsq	17,814	2,819 (16%)	3,387 (19%)
D	lung-m-adsq	27,578	7,409 (27%)	8,669 (31%)
E	brca-g-tn	17,814	11,672 (66%)	11,363 (64%)
F	brca-m-tn	27,578	14,489 (53%)	13,945 (50%)
G	brca-g-stage	17,814	225 (1%)	260 (1%)
H	brca-m-stage	27,578	138 (<1%)	111 (<1%)
I	ov-g-tn	17,814	1,356 (8%)	2,102 (12%)
J	ov-m-tn	27,578	5,886 (21%)	4,213 (15%)
K	coad-g-tn	17,814	6,165 (35%)	6,132 (34%)
L	coad-m-tn	27,578	15,504 (56%)	15,190 (55%)

Table 11 shows the classification performance for both discretization methods. A naïve Bayes (NB) model was used to classify each of the case-control contrasts, using the corresponding test fold. The results shown in Table 11 do not have a significant statistical difference between the Fayyad & Irani criteria and the Kononenko criteria. The classification performance (measured with AUC) in all the tissue classification (Tumor vs Normal) is above 0.86. The only exception is dataset I, which has the ovarian gene expression dataset. The reason for the poor performance in this dataset has to do with the extreme class imbalance between the cases and controls (590 to 8). Also, datasets G and H had poor performance, which can be attributed to the difficulty of the classification task, in this case determination of breast cancer stage. For these three examples the brier skill score (BSS) was negative.

Table 11. Experiments: Discretization performance

#	Dataset	MDL with Fayyad & Irani criteria + NB			MDL with Kononenko criteria + NB			*Statistical comparison between Fayyad & Irani and Kononenko* p-value
		AUC	95% C.I.	BSS	AUC	95% C.I.	BSS	
A	luad-m-tn	1.0	1.0 - 1.0	0.99	1.0	1.0 - 1.0	0.99	1.0
B	lusc-m-tn	1.0	1.0 - 1.0	0.98	1.0	1.0 - 1.0	0.98	1.0
C	lung-g-adsq	0.86	0.78 - 0.93	0.4	0.86	0.78 - 0.94	0.4	0.96
D	lung-m-adsq	0.89	0.84 - 0.94	0.6	0.9	0.85 - 0.94	0.62	0.83
E	brca-g-tn	0.99	0.97 - 1.0	0.92	0.99	0.97 - 1.0	0.92	1.0
F	brca-m-tn	0.97	0.95 - 0.99	0.79	0.98	0.96 - 0.99	0.81	0.71
G	brca-g-stage	0.63	0.57 - 0.69	-0.76	0.62	0.56 - 0.68	-0.72	0.74
H	brca-m-stage	0.57	0.52 - 0.61	-1.4	0.57	0.52 - 0.61	-1.21	0.95
I	ov-g-tn	0.52	0.5 - 0.54	-0.16	0.52	0.5 - 0.54	-0.16	1.0
J	ov-m-tn	1.0	1.0 - 1.0	0.79	1.0	1.0 - 1.0	0.79	1.0
K	coad-g-tn	0.99	0.97 - 1.0	0.76	0.99	0.98 - 1.0	0.87	0.41
L	coad-m-tn	1.0	1.0 - 1.0	0.98	1.0	1.0 - 1.0	0.98	1.0

There are other potential discretization strategies that could be used instead of MDL. In a preliminary study (Lopez Pineda et al. 2013), Fayyad and Irani's MDLPC and Lustgarten's Efficient Bayesian Discretization (EBD) (Lustgarten et al. 2011) were compared. Both discretizers were independently used with a naïve Bayes classifier (NB), with the Efficient Bayesian Multivariate Classifier (EBMC), and the Bayesian Rule Learner (BRL) (Gopalakrishnan et al. 2010). The results of this study, showed that the classification performance of these classifiers is improved by the use of discretizers, and that EBD performs equivalently to MDLPC.

5.2 SELECTING VARIABLES FOR CLASSIFIERS

DNA methylation microarrays are commonly used in cancer research to identify molecular characteristics of disease. These platforms generate high-dimensional data that can be

very challenging to analyze. Typically, researchers are presented with the task of finding differentially methylated (DM) genes that distinguish between two states of the disease (e.g., tumor and normal). Limma, a tool publicly available from the Bioconductor package in the R-language, is used for the analysis of microarray data (Smyth 2004). Limma uses a t-statistic to rank genes in order of evidence for differential expression. It first fits linear models for each gene (lmFit), and then it uses empirical Bayes (eBayes) moderation to adjust the standard error of the models by borrowing information from the rest of the genes (average variance across all genes). This method is very effective in finding differentially expressed (DE) genes in microarray data, however with methylation datasets it has not been equally successful (Buhule et al. 2014).

A recent study (J. Li et al. 2015) showed that it is possible to use ReliefF (Kononenko et al. 1997) to select features from a multi-omic dataset containing methylation data. ReliefF is an multivariate algorithm for data reduction that has been used to uncover gene-gene interactions (Greene et al. 2009). ReliefF iteratively updates the ranking of variables based on how well they can distinguish the target class. In each iteration, ReliefF sequentially selects a sample and finds its k-nearest hits (same class) and k-nearest misses (different class). The spatial distance of the hits is summed into the ranking, while the distance to the misses is deducted from it.

The J2K framework uses the multivariate feature selection ReliefF, as an alternative to Limma. In addition, Information Gain is also compared to provide an example of a univariate feature selection method. The design of these experiments also uses 10-fold cross-validation. In each fold, the top 50 features are selected in Limma (Bioconductor package in R-language, version 3.24), ReliefF and Information Gain (Waikato Environment for Knowledge Analysis, WEKA (Hall et al. 2009), version 3.7). The training folds were used to learn naïve Bayesian

classifiers and their predictive ability was evaluated using the test folds. Their performance is reported using the average area under the receiver operating characteristic curve (AUC).

Table 12 shows the results of the experimentation of ReliefF, where the concatenation of MDLCP, ReliefF, and NB yield models with 50 variables that have a classification performance higher than AUC≥0.79. These results are at least statistically equivalent or statistically better to the framework when feature selection is not used, thus showing that ReliefF is contributing to the parsimony of the models.

Table 12. Experiments: Feature selection performance of ReliefF

#	Dataset	MDLPC + NB	MDLPC + ReliefF + NB			*Statistical comparison between MDLPC+ReliefF+NB and MDLPC+NB*
		AUC	AUC	95% C.I.	BSS	p-value
A	luad-m-tn	1.0	1.0	1.0 - 1.0	0.99	1.0
B	lusc-m-tn	1.0	1.0	1.0 - 1.0	0.98	1.0
C	lung-g-adsq	0.86	0.98	0.98 - 0.99	0.8	0.001
D	lung-m-adsq	0.89	0.97	0.96 - 0.98	0.79	0.002
E	brca-g-tn	0.99	1.0	1.0 - 1.0	0.95	0.28
F	brca-m-tn	0.97	0.99	0.99 - 0.99	0.83	0.02
G	brca-g-stage	0.63	0.91	0.91 - 0.92	0.08	< 0.001
H	brca-m-stage	0.57	0.79	0.78 - 0.8	-0.69	< 0.001
I	ov-g-tn	0.52	0.8	0.75 - 0.86	-0.96	< 0.001
J	ov-m-tn	1.0	1.0	1.0 - 1.0	0.68	0.27
K	coad-g-tn	0.99	1.0	1.0 - 1.0	0.97	0.04
L	coad-m-tn	1.0	1.0	1.0 - 1.0	0.98	1.0

Table 13 shows the results with Information Gain, with similar characteristics as those observed by ReliefF. In fact, there are only a couple of datasets (G, H, I) where Information Gain had significantly lower performance than ReliefF. In the previous section these three datasets were already been identified to be hard classification performance tasks. The selection of algorithm in this case was for ReliefF because it is still better than Information Gain.

Table 13. Experiments: Feature selection performance of Information Gain

#	Dataset	MDLPC + NB	MDLPC + ReliefF + NB	MDLPC + Information Gain + NB			*Statistical comparison between MDLPC+ReliefF+NB and MDLPC+IG+NB* **p-value**	*Statistical comparison between MDLPC+IG+NB and MDLPC+NB* **p-value**
		AUC	**AUC**	**AUC**	**95% C.I.**	**BSS**		
A	luad-m-tn	1.0	1.0	1.0	1.0 - 1.0	0.99	1.0	1.0
B	lusc-m-tn	1.0	1.0	1.0	1.0 - 1.0	0.98	1.0	1.0
C	lung-g-adsq	0.86	0.98	0.98	0.97 - 0.99	0.76	0.2	0.002
D	lung-m-adsq	0.89	0.97	0.96	0.96 - 0.97	0.78	0.3	0.003
E	brca-g-tn	0.99	1.0	0.99	0.99 - 1.0	0.94	0.003	0.92
F	brca-m-tn	0.97	0.99	0.99	0.99 - 0.99	0.83	0.4	0.025
G	brca-g-stage	0.63	0.91	0.92	0.91 - 0.93	0.11	0.6	< 0.001
H	brca-m-stage	0.57	0.79	0.82	0.81 - 0.83	-0.55	< 0.001	< 0.001
I	ov-g-tn	0.52	0.8	0.8	0.75 - 0.86	-0.56	0.94	< 0.001
J	ov-m-tn	1.0	1.0	1.0	1.0 - 1.0	0.72	0.75	0.37
K	coad-g-tn	0.99	1.0	1.0	1.0 - 1.0	0.97	1.0	0.04
L	coad-m-tn	1.0	1.0	1.0	1.0 - 1.0	0.98	1.0	1.0

So far, a comparison has been made considering that the order of the framework should be to discretize first and select features as a second step. However, the opposite order is also possible, and in fact, it is necessary for the comparison of ReliefF and Limma (since Limma requires the use of continuous values).

In Table 14 the results of ReliefF are shown, considering it the first step in the framework. There is a possibility that when the number of variables in the classifier model will be further reduced by the combined effect of feature selection as first step and discretization as second. In the datasets tested with ReliefF and MDLPC, they all kept the original top 50 variable selection. There is no statistically significant difference between this order and those in previous tables, with the exception of those datasets with hard classification tasks (G, H, I). Therefore, using discretization as a first step is preferred.

Table 14. Experiments: Feature selection performance of ReliefF being performed before discretization

#	Dataset	MDLPC + ReliefF + NB	ReliefF + MDLPC + NB				*Statistical comparison between MDLPC+ReliefF+NB and ReliefF+MDLPC+NB* p-value
		AUC	# variables	AUC	95% C.I.	BSS	
A	luad-m-tn	1.0	50	1.0	1.0 - 1.0	0.99	1.0
B	lusc-m-tn	1.0	50	1.0	1.0 - 1.0	0.95	1.0
C	lung-g-adsq	0.98	50	0.95	0.92 - 0.99	0.55	0.09
D	lung-m-adsq	0.97	50	0.95	0.91 - 0.98	0.7	0.14
E	brca-g-tn	1.0	50	0.98	0.96 - 1.0	0.91	0.14
F	brca-m-tn	0.99	50	0.99	0.98 - 1.0	0.77	0.18
G	brca-g-stage	0.91	50	0.64	0.57 - 0.7	-0.62	< 0.001
H	brca-m-stage	0.79	50	0.57	0.53 - 0.61	-1.11	< 0.001
I	ov-g-tn	0.8	50	0.52	0.5 - 0.54	-0.16	< 0.001
J	ov-m-tn	1.0	50	1.0	1.0 - 1.0	0.72	0.88
K	coad-g-tn	1.0	50	1.0	1.0 - 1.0	0.97	1.0
L	coad-m-tn	1.0	50	1.0	1.0 - 1.0	0.98	1.0

Table 15 shows the results of the use of Limma to find DE or DM genes for the construction of a NB model. The top-50 variables selected by Limma, are further reduced when MDLPC is not able to select at least one cut-point to create two discretization bins. The statistical significance between the frameworks using ReliefF and Limma in two cases: A) when MDLPC is used after both feature selection algorithms, yielding a mix of results that are either equivalent or significantly worst, and B) when MDLPC is used before ReliefF, but after Limma, yielding mostly results that are significantly worst in the case of Limma.

Table 15. Experiments: Feature selection performance of Limma

#	Dataset	MDLPC+ ReliefF+ NB	ReliefF+ MDLP +NB	Limma + MDLPC + NB				*Statistical comparison between ReliefF+MDLPC +NB and Limma + MDLPC+NB*	*Statistical comparison between MDLPC+ReliefF +NB and Limma +MDLPC+NB*
		AUC	**AUC**	**# variables**	**AUC**	**95% C.I.**	**BSS**	**p-value**	**p-value**
A	luad-m-tn	1.0	1.0	7	0.88	0.8 - 0.97	0.43	0.008	0.008
B	lusc-m-tn	1.0	1.0	19	0.97	0.95 - 0.99	0.43	0.02	0.02
C	lung-g-adsq	0.98	0.95	39	0.89	0.82 - 0.95	0.14	0.08	0.003
D	lung-m-adsq	0.97	0.95	18	0.83	0.76 - 0.89	0.23	0.001	< 0.001
E	brca-g-tn	1.0	0.98	18	0.92	0.88 - 0.95	0.14	0.003	< 0.001
F	brca-m-tn	0.99	0.99	39	0.98	0.97 - 1.0	0.69	0.7	0.1
G	brca-g-stage	0.91	0.64	1	0.49	0.44 - 0.54	-0.01	< 0.001	< 0.001
H	brca-m-stage	0.79	0.57	1	0.49	0.45 - 0.54	-0.01	0.02	< 0.001
I	ov-g-tn	0.8	0.52	35	0.52	0.5 - 0.54	-0.15	0.9	< 0.001
J	ov-m-tn	1.0	1.0	14	0.84	0.7 - 0.99	-0.81	0.04	0.04
K	coad-g-tn	1.0	1.0	51	0.95	0.88 - 1.0	0.87	0.1	0.1
L	coad-m-tn	1.0	1.0	25	0.8	0.71 - 0.89	0.25	< 0.001	< 0.001

There is, however, still an open question about the importance of using a multivariate method like ReliefF, given that the default bioinformatics analysis prefers Limma. A potential explanation is that ReliefF is indeed better than Limma in those cases where methylation data is being used, with mixed results when using gene expression. A potential explanation for this behavior is that normalized beta values in methylation microarray data have a probability density function that follows a bimodal distribution, where two distinct peaks are observed for hyper-methylated and hypo-methylated samples. Bi-modality in the data is a hard problem to solve for univariate linear models, such as the ones used by Limma. In contrast, the multivariate approach of ReliefF facilitates the exploration of both peaks in the bimodal distribution. This is the main reason why ReliefF finds differentially methylated genes that perform better than Limma. The selection of genes is, in the worst-case scenario, as good as those found by Limma in both platforms.

5.3 BUILDING BAYESIAN NETWORK CLASSIFIERS

Bayesian network classifiers are commonly used in biomedical problems, yielding excellent results for classification. Recent comparisons between several machine learning classifiers, it was observed the advantages of using a Bayesian framework (Lopez Pineda et al. 2015; Jiang et al. 2014). Among these machine learning models, the logistic regression seems to be the most used in bioinformatics laboratories. However, it has been suggested that the predictions obtained from a logistic regression model are the same as those predictions originated from a naïve Bayes model (Sebastiani et al. 2012). Logistic regression is consistent with the conditional independence assumption used in naïve Bayes. Nevertheless, there are important differences in each algorithm. For example, logistic regression will adjust its parameters to maximize the conditional likelihood of the data, even if the resulting parameters are inconsistent with the naïve Bayes parameter estimates (Mitchell 2015). Furthermore, the possibility of making predictions in the presence of missing data is a characteristic that is better modeled in a Bayesian approach. The Bayesian model has prior parameter estimates that are obtained during the training step, and the prediction of a new case can be done without imputing any missing data.

The J2K framework uses EBMC, a novel algorithm for the construction of Bayesian classifiers. EBMC facilitates the discovery of Bayesian network structures that have a good classification performance. The search strategy has two main phases: a) the greedy search for predictors of the target class node, and b) the search for conditional dependencies between those predictors, given the target class node. EBMC continuously switches between phases, until no improvement can be achieved, measured by the K2 score. At this point, the EBMC has obtained

a tree augmented naïve Bayes structure, but a last step is to remove any arc between nodes that improve the overall score of the network. EBMC potentially generates classifiers with less variables when fewer predictors are searched than those available, but this parameter is only an indication on when to stop the search, because the algorithm could stop by itself it no improvement is obtained.

The design of these experiments uses 10-fold cross-validation. Each fold is discretized with MDLPC, and the top 50 ReliefF features are selected. Three algorithms are used to build a Bayeisan network: a) naïve Bayes (NB), which was shown in Table 12, b) EBMC, which was implemented in Java for WEKA, and c) Bouckaert's TAN algorithm (Bouckaert 2008). The TAN algorithm first creates a tree-like structure from all the possible variables, and then it connects all nodes naïvely to the target node, obtaining a tree-augmented naïve Bayes structure (TAN) similar to that in EBMC. The main difference between EBMC structure and Bouckaerts's TAN structure is the number of variables used, since Bouckaert's TAN does not try to reduce the number of variables used in the final Bayesian structure, while EBMC starts with an empty structure and iteratively builds it up until no improvement can be obtained.

Table 16 shows the classification performance of the framework with EBMC, where most classifiers achieved AUC > 0.89. The exceptions are the same as in previous experiments (G, H, I). In all cases the number of variables was reduced to 30 or less. Results for datasets A, B, D, F, K, and L where statistically equivalent or better to the results obtained when using NB with 50 variables, which confirms the contribution of EBMC to the parsimony of these models. The AUC for datasets C, E, and J were lower than when using NB, but there is no statistical difference. Finally, results for datasets G, H, and I are statistically worst than when using NB. However, for these datasets the Brier Skill Score is negative when using EBMC and also when

using NB, which talks about the poor calibration of the models, and the difficulty of these classification problems. Overall, EBMC contributed to the parsimony of the models by reducing the number of variables from 50 to 30. The optimal number of reduction that EBMC could obtain without compromising the AUC performance is still an open question, which would require the use of a wrapper approach to keep exploring the best alternative.

Table 16. Experiments: building models with EBMC

#	Dataset	MDLPC + ReliefF+ NB	MDLPC + ReliefF + EBMC				*Statistical comparison between MDLPC+ReliefF+NB and MDLPC+ReliefF+EBMC*
		AUC	# variables	AUC	95% C.I.	BSS	p-value
A	luad-m-tn	1.0	30	1.0	1.0 - 1.0	0.94	1.0
B	lusc-m-tn	1.0	30	1.0	0.99 - 1.0	0.94	0.32
C	lung-g-adsq	0.98	29	0.89	0.82 - 0.96	0.46	0.008
D	lung-m-adsq	0.97	30	0.97	0.95 - 0.99	0.73	0.82
E	brca-g-tn	1.0	30	0.99	0.97 - 1.0	0.92	0.32
F	brca-m-tn	0.99	20	1.0	0.99 - 1.0	0.82	0.06
G	brca-g-stage	0.91	26	0.62	0.56 - 0.68	-0.32	< 0.001
H	brca-m-stage	0.79	15	0.6	0.56 - 0.65	-0.19	< 0.001
I	ov-g-tn	0.8	3	0.55	0.49 - 0.62	-0.14	< 0.001
J	ov-m-tn	1.0	30	0.96	0.87 - 1.0	0.49	0.31
K	coad-g-tn	1.0	30	1.0	1.0 - 1.0	0.97	1.0
L	coad-m-tn	1.0	30	1.0	1.0 - 1.0	0.98	1.0

Table 17 shows the results of using Bouckaert's TAN algorithm, where almost all models achieved a classification performance AUC > 0.93, expect for datasets G, H, I. The use of this algorithm does not reduce the number of variables, which means that all models were constructed using 50 variables. The AUC variation with the models built using EMBC has mixed results (some are better, some are worst), but none of them has a statistically significant difference. Therefore, the use of EBMC is encouraged.

Table 17. Experiments: building models with Bouckaert's TAN

#	Dataset	MDLPC + ReliefF + EBMC	MDLPC + ReliefF + TAN			*Statistical comparison between MDLPC+ReliefF+EBMC and MDLPC+ReliefF+TAN*
		AUC	AUC	95% C.I.	BSS	p-value
A	luad-m-tn	1.0	0.98	0.96 - 1.0	0.88	0.16
B	lusc-m-tn	1.0	0.98	0.94 - 1.0	0.9	0.44
C	lung-g-adsq	0.89	0.93	0.88 - 0.99	0.49	0.34
D	lung-m-adsq	0.97	0.96	0.93 - 0.99	0.67	0.59
E	brca-g-tn	0.99	0.99	0.97 - 1.0	0.92	0.94
F	brca-m-tn	1.0	0.98	0.97 - 1.0	0.84	0.07
G	brca-g-stage	0.62	0.63	0.57 - 0.7	-0.29	0.80
H	brca-m-stage	0.6	0.6	0.56 - 0.64	-0.55	0.91
I	ov-g-tn	0.55	0.56	0.53 - 0.59	-0.04	0.81
J	ov-m-tn	0.96	1	0.99 - 1.0	0.56	0.32
K	coad-g-tn	1.0	1	0.99 - 1.0	0.92	0.32
L	coad-m-tn	1.0	1	1.0 - 1.0	0.98	1.0

5.4 SELECTING PARSIMONIOUS MODELS

Until this point, a traditional machine learning classification process has been followed were parsimony is gradually achieved by a specific sequence of algorithms. The J2K framework, proposes the use of a post-classification approach were inspection of the Bayesian model, which created in the classification step, can lead to the identification of an even more parsimonious model. The selection of this model with fewer nodes can be done by reducing the number of steps that the EBMC search algorithm is taking, or by inspection of the graphical structure of the network. The second approach has been described in this dissertation with the use of Junction trees, and the selection of one of its cliques.

The experimental design of this section applies 10-fold cross-validation. Each fold is discretized with MDLPC, and the top 50 ReliefF features are selected. A Baysian classifier is

constructed using EBMC, where 30 predictors are searched and a completely connected network is allowed (maximum number of parents and children is also 30). Then, all cliques are extracted from the network, and evaluated individually using the training fold. The clique that obtains the best AUC with training data is selected as the best clique and evaluated with the test fold.

Table 18 shows the results of the evaluation of the best clique. Datasets A, B, J, and L obtained classification performances of AUC > 0.87, but only B and J are statistically equivalent to those from the complete network. These are all parsimonious models that can be reported. In contrast, dataset K had an AUC of 0.76, which is significantly worst than the complete network. However, for datasets C, D, E, F, G, H, and I, the training data was not sufficient to build a parsimonious model that could be evaluated with the test data. When removing nodes from the original network to keep only those in the clique, the training data does not consider some conditional possibilities that are present in the test folds, therefore the predictions of the network are assigned to only one class in most cases.

Table 18. Experiments: Selection of best clique from Junction tree

#	Dataset	**MDLPC + ReliefF + EBMC**	**MDLPC + ReliefF + EBMC + Junction tree**					*Statistical comparison between MDLPC + ReliefF + EBMC and MDLPC + ReliefF + EBMC + BestClique* **p-value**
		AUC	**# cliques in network**	**# nodes in cliques**	**AUC (best clique)**	**95% C.I. (best clique)**	**BSS (best clique)**	
A	luad-m-tn	1.0	8	7	0.87	0.78 - 0.95	0.58	0.004
B	lusc-m-tn	1.0	11	5	0.93	0.87 - 0.99	0.74	0.05
C	lung-g-adsq	0.89	8	8	0.53	0.42 - 0.64	0	< 0.001
D	lung-m-adsq	0.97	9	7	0.53	0.44 - 0.61	0	< 0.001
E	brca-g-tn	0.99	14	6	0.51	0.44 - 0.58	-0.02	< 0.001
F	brca-m-tn	1.0	8	7	0.51	0.46 - 0.56	-0.02	< 0.001
G	brca-g-stage	0.62	7	8	0.51	0.45 - 0.58	-0.01	0.004
H	brca-m-stage	0.6	7	4	0.51	0.46 - 0.55	0	< 0.001
I	ov-g-tn	0.55	2	4	0.61	0.45 - 0.76	-0.13	0.06
J	ov-m-tn	0.96	15	5	1	1.0 - 1.0	0.92	0.3
K	coad-g-tn	1.0	4	11	0.76	0.67 - 0.84	0.07	< 0.001
L	coad-m-tn	1.0	8	7	0.91	0.84 - 0.97	0.62	0.003

Table 19 shows the classification results when the first iteration of EBMC is used. Almost all models included 10 variables, and achieved classification performances of AUC > 0.9. Datasets G, H, and I are cases that were not able to classify in previous steps of the framework. For all results, there is no statistical significance between the models created with the first iteration of EBMC and those models that undertake every iteration. Parsimony of these models is guaranteed by the first iteration only, which reduces the uncertainty of parameterization of EBMC algorithm, given that fewer predictors are searched, and also because the parents and children can be exhaustively investigated. EBMC-first is a better alternative than J2K to always provide a parsimonious model, given the available training data.

Table 19. Experiments: Selection of the first iteration of EBMC

#	Dataset	MDLPC + ReliefF + EBMC-first				*Statistical comparison between MDLPC + ReliefF + EBMC-first and MDLPC + ReliefF + EBMC* **p-value**	*Statistical comparison between MDLPC + ReliefF + EBMC-first and MDLPC + ReliefF + EBMC + BestClique* **p-value**
		# variables	**AUC**	**95% C.I.**	**BSS**		
A	luad-m-tn	10	1.0	0.99 - 1.0	0.87	0.30	0.005
B	lusc-m-tn	10	1.0	1.0 - 1.0	0.98	0.32	0.04
C	lung-g-adsq	10	0.9	0.84 - 0.97	0.33	0.81	< 0.001
D	lung-m-adsq	10	0.95	0.91 - 0.98	0.63	0.18	< 0.001
E	brca-g-tn	10	1.0	1.0 - 1.0	0.89	0.28	< 0.001
F	brca-m-tn	10	1.0	0.99 - 1.0	0.82	0.83	< 0.001
G	brca-g-stage	10	0.6	0.54 - 0.66	-0.28	0.59	0.02
H	brca-m-stage	6	0.59	0.55 - 0.64	-0.08	0.74	0.002
I	ov-g-tn	3	0.56	0.5 - 0.62	-0.14	0.83	0.05
J	ov-m-tn	10	1.0	0.99 - 1.0	0.47	0.33	0.07
K	coad-g-tn	10	1.0	1.0 - 1.0	0.95	1.0	< 0.001
L	coad-m-tn	10	1.0	1.0 - 1.0	0.93	0.23	0.003

5.5 HYPOTHESIS TESTING

The J2K framework has four components with specific algorithms in each one. Alternative algorithms are tested in each component, as shown in Tables 11 to 19, using the TCGA datasets described in Table 1. Although there is a statistical comparison (p-value with DeLong's method) associated with each comparison for each dataset in Tables 11 to 19, a statistical inference testing could reveal more concrete results across all datasets. These tests are enumerated in Table 20, and aggregated by J2K component.

Table 20. Hypothesis testing

J2K component	Algorithm testing
Discretization	1) Fayyad & Irani vs. Kononenko
Feature Selection	2) ReliefF vs. Information Gain vs. No Feature Selection 3) Discretization before Feature Selection vs. Discretization after Feature Selection 4) Limma vs. ReliefF
Model Building	5) Naïve Bayes vs. EBMC vs. TAN
Post-classification	6) Best clique from Junction vs. First iteration of EBMC vs. No post classification

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true (Weisstein 2004a). Two statistical hypothesis testing methods were used for comparing classification performance across all datasets: a) Paired 2-tailed t-test (Weisstein 2004b), and b) Wilcoxon signed rank test (Lowry 2014). In the paired t–test, given two paired sets X_i and Y_i of n measured values, the paired t-test determines whether they differ from each other in a significant way under the assumptions that the paired differences are independent and identically normally distributed. Therefore, it can be calculated as seen in Equation 8.

Equation 8. T-test

$$\hat{X} = (X_i - \bar{X})$$

$$\hat{Y} = (Y_i - \bar{Y})$$

$$t = (\bar{X} - \bar{Y}) \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(X_i - Y_i)^2}}$$

When the data within two correlated samples fail to meet one or another of the assumptions of the t-test, an appropriate non-parametric alternative can often be found in the Wilcoxon Signed-Rank Test (W statistic). The Wilcoxon test begins by transforming each instance of $X_i - Y_i$ into its absolute value, which is accomplished simply by removing all the positive and negative signs. Those cases in which there is a zero difference are eliminated from consideration, since they provide no useful information. The remaining absolute differences are then ranked from lowest to highest, with tied ranks included where appropriate. The positive or negative sign that was removed from the $X_i - Y_i$ difference is re-attached. Lastly, the W statistic is calculated by obtaining an average of the available ranks.

For both test the null hypothesis is that the classification performance between method A and method D are identical, while the alternative is that they are different. The alpha parameter in all tests was established ad 0.05 given that the number of experiments is $n = 12$.

H_0: classification performance between Method A and Method B are identical.

H_1: classification performance between Method A and Method B are different.

Discretization. In Table 21, it is shown that MDL Fayyad & Irani criteria for discretization is statistically indistinguishable to MDL Kononenko criteria. Both algorithms have an average

performance of AUC=0.87, with standard error of the mean (SEM) of 0.05. The number of variables is statistically significantly reduced from an average of 23,510 in the original datasets to 7,770 with Fayyard & Irani, and 7,688 with Kononenko. The number of variables is statistically equivalent in both algorithms.

Table 21. Hypothesis testing: Kononenko vs Fayyad&Irani

	P-value
Paired 2-tailed t-test	0.59
Wilcoxon signed rank test	1.0

P-values for two hypothesis test methods. T: p-value in paired 2-tailed t-test, W: p-value in Wilcoxon signed rank test. Only top triangle is shown. Alpha is 0.05

Feature Selection. In Table 22, it is shown that feature selection significantly improves classification. The AUC improves from 0.87, when no feature selection algorithm is used, to 0.95 with a SEM of 0.02. The classification performance of ReliefF is statistically indistinguishable to that from Information Gain (IG). The number of variables is statistically significantly reduced from 7,770 to 50 variables.

Table 22. Hypothesis testing: ReliefF vs Information Gain

	MDLPC + ReliefF	**MDLPC + IG**
MDLPC	0.01 (T) 0.02 (W)	0.02 (T) 0.03 (W)
MDLPC + ReliefF	–	1.0 (T) 0.6 (W)

P-values for two hypothesis test methods. T: p-value in paired 2-tailed t-test, W: p-value in Wilcoxon signed rank test. Only top triangle is shown. Alpha is 0.05

In Table 23, it is shown that it is preferable to discretize before using a feature selection method, and not after. This is a soft claim, given that the non-parametrical W-statistic is not rejecting the null hypothesis, but the t-statistic is (with an alpha of 0.05). However, upon inspection of Table 14, discretization before feature selection is recommended since in the worst case scenario it

would be the same as discretizing after feature selection. Table 23 also shows that using ReliefF statistically significantly improves classification when compared to Limma.

Table 23. Hypothesis testing: Feature Selection and Discretization

	ReliefF + MDLPC	**Limma + MDLPC**
MDLPC + ReliefF	0.03 (T) 0.06 (W)	< 0.01 (T) < 0.01 (W)
ReliefF + MDLPC	–	< 0.01 (T) < 0.01 (W)

P-values for two hypothesis test methods. T: p-value in paired 2-tailed t-test, W: p-value in Wilcoxon signed rank test. Only top triangle is shown. Alpha is 0.05

Model building. In Table 24 it is shown that NB has statistically significantly improved classification when compared to both EBMC and TAN. The AUC of both EBMC and TAN is 0.88, with a SEM of 0.05. EBMC is statistically indistinguishable to TAN. However, the number of variables that NB and TAN use (50, selected from ReliefF) cannot be said to be a parsimonious model, and therefore an acceptable error might be to use EBMC in favor of a simplistic model. The average BSS in NB is 0.53, while in EBMC is 0.53. The search of EBMC is truncated with the expectation of 30 predictors (features). However, NB assumes a strong independence between variables, which creates a structure that does not allow for any discrimination or prioritization from the variables. Therefore, EBMC is still a valid choice to use.

Table 24. Hypothesis testing: EBMC vs TAN vs NB

	EBMC	**TAN**
NB	0.04 (T) 0.04 (W)	< 0.01 (T) < 0.01 (W)
EBMC	–	0.95 (T) 0.68 (W)

P-values for two hypothesis test methods. T: p-value in paired 2-tailed t-test, W: p-value in Wilcoxon signed rank test. Only top triangle is shown. Alpha is 0.05

Post classification. In Table 25, it is shown that the first iteration of EBMC is statistically indistinguishable to EBMC (AUC=0.88, SEM=0.05). The First iteration of EBMC improves classification when compared to the best clique from the Junction tree (AUC=0.68, SEM=0.05). There is a clear evidence that the use of a post-classification strategy can find a model that is more parsimonious, given a graphical structure. A NB structure would require more search in order to obtain a smaller subset of variables.

Table 25. Hypothesis testing: Use of post classification

	EBMC-first	**EBMC + Junction**
EBMC	1.0 (T) 0.7 (W)	< 0.01 (T) < 0.01 (W)
EBMC-first	–	< 0.01 (T) < 0.01 (W)

P-values for two hypothesis test methods. T: p-value in paired 2-tailed t-test, W: p-value in Wilcoxon signed rank test. Only top triangle is shown. Alpha is 0.05

5.6 SUMMARY

The main claim in this dissertation is to investigate whether the J2K framework provides a mechanism for the identification of parsimonious Bayesian models, by using specific algorithms in sequence. The arguments to support this claim can be seen by the following Sections.

a. MDLPC improves parsimony of models: From Sections 4, 5.1.1, and 5.1.5, it was shown that this claim was strongly supported.

b. ReliefF improves parsimony of models: From Sections 4, 5.1.2, and 5.1.5, it was shown that this claim was strongly supported.

c. EBMC improves parsimony of models: From Sections 4, 5.1.3, and 5.1.5, it was shown that this is supported.

d. Post classification improves parsimony of models: From Sections 4, 5.1.4, and 5.1.5, it was shown that this claim was inconclusive for Junctions, but supported for the first iteration of EBMC.

A secondary claim is that the J2K framework facilitates the creation of parsimonious multi-omic data integration models (MODI). From the annotated example in Section 4 it was shown that this claim is supported.

6.0 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

6.1 CONCLUSIONS

Computational models can accelerate translational research, facilitating improved diagnosis and personalized treatment options for patients. Using a Bayesian framework, this dissertation has shown the feasibility of building parsimonious models for classification in cancer. It demonstrated how microarray data can be transformed into a Bayesian network with high classification performance with few variables.

In machine learning, the parsimony principle states that if two models can adequately model a given set of data, the one that is described by a fewer number of parameters will have better predictive ability given new data (Seasholtz & Kowalski 1993). The experiments presented in this dissertation using real life cancer datasets from The Cancer Genome Atlas, have shown that a parsimonious model is possible using state of the art algorithms.

Each component of the J2K framework is contributing individually to the parsimony of the model. Discretization with MDLPC provides a mechanism for the selection of genes or methylation sites that can take discrete values. The use of ReliefF facilitates finding groups of genes or methylation sites that jointly classify between disease states. The use of EBMC to build a Bayesian network efficiently selects an accurate classifier. EBMC facilitates the search of a Bayesian network structure that can become a parsimonious classifier. Although the algorithms

selected for the J2K framework could be replaced by others that might become available in the future, the selection of those algorithms should be based on the principle of parsimony.

The J2K framework enables the creation of parsimonious multi-omic data integration models. Handling the large number of variables that cancer molecular studies have (and especially multi-omic studies) is a complex task. The MODI framework builds upon the J2K framework by taking these parsimonious models and creating hierarchical models. There is a potential to explore other algorithms for the novel genomic problems, such as pan-cancer analysis.

Finally, in this dissertation many known algorithms were used in a novel framework to create classification models for cancer. The informatics novelty of this approach is the selection of algorithms, the post-classification of cliques and clusters of nodes, and the application to a multi-omic problem. The impact of this dissertation to cancer problems lie in the future use that these novel frameworks can have in the downstream analysis of biomarker discovery.

6.2 LIMITATIONS

The results presented in this dissertation suggest that the J2K framework is a valid approach to create parsimonious models. The results have to be interpreted in light of the following limitations:

a. For each component of the J2K framework, only one algorithm was used. The J2K framework is only one approach to create a parsimonious model from data, but it does not address the selection of the most parsimonious model.

b. The discretization component of J2K is necessary for the Bayesian-EBMC component, but using a different machine learning classifier might not require the discretization step.

c. The results presented only consider classification tasks using microarray data from the TCGA, in four cancer types. The external validity of the J2K needs to be tested when using other data sources.

d. Only one instance of the algorithm for Junction tree generation was used.

6.3 FUTURE WORK

The work described in this dissertation has shown the potential to create parsimonious models for cancer research. This work directly leads to the following future directions:

6.3.1 Investigate the Junction Structure.

There is a potential to find interesting biological meaning in the structure generated from the Junction tree algorithm. For example, one application could be in the determination of gene-gene-phenotype regulatory networks. Epistasis describes how gene interactions can affect phenotypes, and the data-driven approach of the J2K could be optimized to find such relationships. The impact of this method would be in the identification of relevant functional modules that could lead to new translational applications.

Typically, the result of a microarray case-control experiment is a list of differentially expressed genes (Smyth 2004). The list of genes are then processed with gene set analysis tools, which are statistical methodologies to 1) rank the top scoring genes given a condition, 2) produce

small p-values for those genes (Tarca et al. 2013). However, a significant result from these methods does not necessarily mean that the gene set of interest contains genes that are associated with the phenotype (Maciejewski 2014). To solve this problem, heuristic methods are used to compare the gene sets with known biological pathways such as Gene Ontology GO (Ashburner et al. 2000) and the Kyoto Encyclopedia of Genes and Genomes KEGG (Ogata et al. 1999).

From a systems biology perspective, functional modules are a group of molecular components (i.e., genes, gene products, or metabolites) that coordinately participate in accomplishing a specific biological function in the cell (Resendis-Antonio et al. 2012). Functional modules are subnetworks of the interactome composed of elements with physical interactions (i.e., protein-protein) or genomic interactions (i.e., gene-gene, gene-disease) (Mitra et al. 2013). Since the J2K provides a quick way of extracting groups of genes that are highly correlated between each other, and also to the disease, there is an initial thought that this would be a tool to find functional modules (in addition to parsimonious classifiers).

6.3.2 Explore Novel Search Strategies for Bayesian Model Building.

The EBMC search strategy has been shown in this dissertation to have an important role in finding Bayesian classifiers that accurately capture the complexity of data, while at the same time reducing the number of variables. Other search strategies could lead to the identification of causal relationships that can be explored in the experimental laboratory setting.

The use in J2K of a univariate approach for discretization (MDLPC), multivariate approach for feature selection (ReliefF), and multivariate approach for model building (EBMC), creates a pipeline that facilitates the search. However, there is still a need to identify models that can provide feedback to these components. A wrapper approach is a potential solution to this

problem, but increases the time needed for training the models. Therefore, some heuristic methods should be considered when building a J2K-wrapper.

Other alternatives could be to apply post-processing to state-of-the-art algorithms such as Model Averaging Naïve Bayes (MANB) algorithm (W. Wei et al. 2011), or the Bayesian Rule Learning (BRL) algorithm (Gopalakrishnan et al. 2010), which already expand on the idea of model searching, and are also well calibrated.

6.3.3 Expand the MODI Framework to Integrate more 'Omics'.

The study of the molecular differences in cancer samples still has open questions of interest to personalized medicine experts. The use of a single-omic approach has limitations that are partially addressed by the MODI framework. Multi-omic data integration aims to enable personalized medicine by using information from various molecular elements, however, in MODI only two 'omics' are considered (gene expression and DNA methylation). However, the MODI framework could be easily extended to include more 'omic' platforms, as well as clinical data and imaging studies from the Electronic Health Record (EHR). The use of latent variables is an important contribution that could lead to an automated search of other latent variables.

6.3.4 Modeling from Liquid Biopsy Samples

In spite of the great advances in machine learning, the main source of information is case-control studies that were generated with data from sequencing biopsy samples. Recently, it has been suggested that the use of blood-based liquid biopsies can be used to develop a noninvasive method to detect and monitor tumors (Bettegowda et al. 2014). The current tumor staging

procedure cannot detect early tumor cell dissemination as a key event in tumor progression, but there use of circulating tumor cells (CTC) (Alix-Panabières et al. 2012), and tumor-educated platelets (Best et al. 2015), might be able to help accomplish this task. Cell-free fragments of DNA can circulate cell-free in the blood stream. Trying to correlate these fragments with tumor staging and prognosis is a promising area of research (Diaz & Bardelli 2014). The sensitivity when using this approach in metastatic tumors is very high (Diehl et al. 2008), given the high content of these circulating fragments. However, in early stages this has not been the case, given the smaller number of circulating fragments. There is a lack of methods that look at liquid samples to find a group of biomarkers that can serve as a screening, diagnostic, or monitoring mechanism.

Future work of this dissertation would be to explore the potential of liquid biopsies to become a screening mechanism for breast cancer. The goal of this potential study would be to obtain a parsimonious classifier, composed by small number of biomarkers from blood, that can detect breast cancer staging with a high sensitivity and specificity. The output of this classifier, could be tested and implemented in clinical care. The publicly available data in The Cancer Genome Atlas (TCGA) contains datasets that were generated using high-throughput technologies (microarrays, and next-generation sequencing). An exploration of the most informative type of data for correlation with liquid samples (e.g., SNPs, expression, methylation, etc) would be needed. The proposed method would also explore a selection of various types of samples, including: a) solid tumor biopsies from early stage breast cancer patients; b) solid tumor biopsies from late stage breast cancer patients; c) blood from early stage patients; and d) blood from late stage patients. Matched normal tissue samples are needed to exclude germ-line mutations from the tumors.

A machine learning analysis is suitable for identifying early stage cancer biomarkers from liquid biopsies. A challenge of this task is the amount of features available from where the machine learning process can take action. The process of creating a computational model, requires a training and test datasets, in the format of a data matrix where columns are features, and rows are samples. The target class for each sample would be assigned using a retrospective study where the true value can be obtained from an invasive procedure (solid tumor biopsy, and IHC analysis). The features for training the machine learning model would be from both tumor-derived and blood-derived genotyping, while only the blood-derived genotyping features would be used for testing. The computational models created with this approach could be tested using either an independent dataset, or a prospective analysis. This study can be repeated over time, given that obtaining blood from patients is a far less invasive procedure than obtaining a tumor biopsy.

A liquid biopsy screening that requires testing a small number of biomarkers has the potential to be implemented in clinical care. The methods presented in this dissertation could be adapted for searching blood-derived biomarkers. Data from the TCGA could be used to find an initial selection of biomarkers, and other available datasets with blood derived genotyping would have to be investigated, e.g., the Norwegian Women and Cancer study (NOWAC) (Dumeaux et al. 2008).

APPENDIX A

J2K APPLIED TO LUNG CANCER SUBTYPING

BACKGROUND

Lung cancer is the leading cause of human cancer death in the United States. Adenocarcinoma (ADC) and squamous cell carcinoma (SCC) are the most common histological subtypes among all lung cancers. Both of them are a form of cancer that develops in the epithelial cells (carcinoma), and belong to the category of non-small cell lung cancer. Several studies have shown that molecular profiling of lung carcinoma is a viable tool for disease diagnosis (Cai et al. 2014), and prognosis (Subramanian & Simon 2010). What is more, distinguishing between ADC and SCC has significant clinical implications – both can have different treatment regimens. Furthermore, ADC and SCC have distinct progression rate and progression free survival, which determines the selection of treatment (Chiu et al. 2014). The standard molecular testing for lung cancer is to check for mutations of two molecules: epidermal growth factor receptor (EGFR) and rearrangement of anaplastic lymphoma kinase (ALK). Each protein has mutations that lead to the development of lung cancer. However, EGFR is found to be mutated only in around 10% of tumors (Dacic et al. 2010). Similarly, ALK mutation occurs only in 6% of tumors (Soda et al. 2007). Although some drugs target EGFR and ALK positive

tumors with therapeutic benefits for the patient, 75% of lung tumors do not possess these molecular alterations (Richer et al. 2015). The high sensitivity and low specificity of these diagnostic molecules is a motivation to research into new diagnostic models.

Typically, a biopsy tissue represents a very small portion of the lung. In spite of ultrasound guidance, it is easy to miss a small focal malignancy, and end up retrieving tumor-adjacent histologically-normal tissue (TAHN) along with Tumor tissue. In those cases, the biopsy is discarded if it cannot retrieve more than 50% of tumor tissue (Dooms et al. 2014). The patient would have to undergo a new procedure to obtain another biopsy. Thus, it is worth exploring computational alternatives for classifying lung cancer subtypes given a small biopsy sample and a mix of TAHN and tumor tissue.

The goal of the work presented in this Appendix was to test whether computational modeling can be a viable approach to accurately differentiate between lung cancer subtypes, given molecular profiles of tumor tissue and than using DNA methylation data. Specifically, the hypothesis that was tested was that "Bayesian modeling is sufficient to classify lung cancer subtypes, regardless of the tissue sample being tumor or tumor-adjacent". Also, it was evaluated the ability of a Bayesian classifier to accurately differentiate lung cancer subtypes using molecular profiles of real lung cancer data sets that are also publicly available.

EXPERIMENTAL DESIGN

The lung cancer datasets from Table 1 were used. In particular, the methylation and gene expression datasets where ADC and SCC are used as a classification task (datasets C and D). A supervised classification process was followed on 10-fold cross-validation. That is, for each fold

the dataset was partitioned into training and test, where the former contains 90% of the samples, while the latter contains the remaining 10%. Each partition maintains the same class distribution as the whole dataset (stratified). In each fold, the datasets were analyzed using the experimental design as illustrated in Figure 19. According to the design, there are four main components, namely, a) Feature Selection, b) Discretization, c) Model Building, and d) Evaluation. Additionally, it was performed a Gene Functional Analysis, and Clustering methods were applied to better understand the characteristics of the features chosen by this framework.

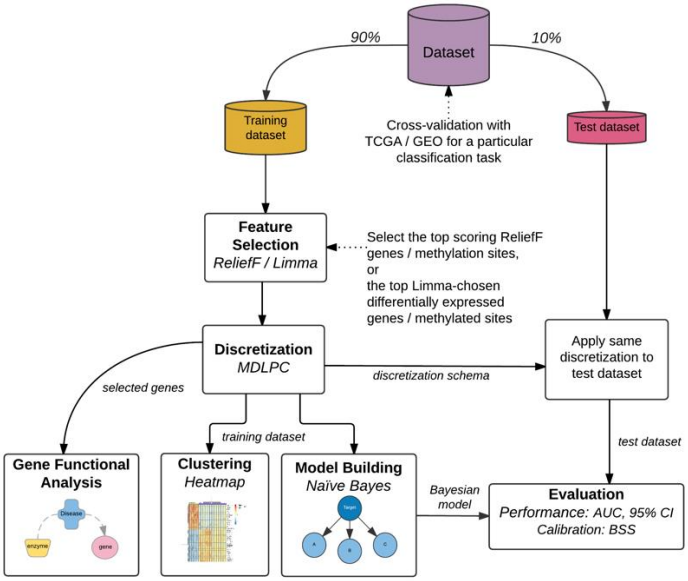

Figure 19. Appendix. Experimental Design

Cross-validation (10-folds) experimental design for a particular classification task, using feature selection, discretization. There are three outcomes: a simple naïve Bayesian model with its test evaluation; clustering of samples based on selected genes; and gene enrichment analysis. Algorithms: ReliefF, Limma, minimum description length principle cut (MDLPC). Evaluation: area under the receiver operating characteristic (AUC), 95% confidence interval (CI), and Brier Skill Score (BSS).

RESULTS

Four classification tasks were investigated depending on the tissue type. These tasks test the hypothesis that the TAHN tissue has distinct genomic signatures that can differentiate among non-small cell lung cancer subtypes. The classification tasks can be described as follows:

1. $TAHN_{ADC}$ *vs.* $Tumor_{ADC}$, and $TAHN_{SCC}$ vs $Tumor_{SCC}$, searches for molecular differences between tumor tissue and TAHN tissue. These tasks are only applied to one lung cancer subtype at a time, either adenocarcinoma or squamous cell carcinoma patients;
2. $Tumor_{ADC}$ *vs.* $Tumor_{SCC}$, which searches for molecular differences between subtypes using only Tumor tissue;
3. $TAHN_{ADC}$ *vs.* $TAHN_{SCC}$, which searches for molecular differences between subtypes using only TAHN tissue; and
4. $TAHN\text{-}Tumor_{ADC}$ *vs.* $TAHN\text{-}Tumor_{SCC}$, which searches for molecular differences between subtypes using both TAHN and Tumor tissue.

The classification performance for every naïve Bayes classifier was calculated by averaging the AUCs over all folds from the experimental design illustrated in Figure 1. Table 26 shows results for the classification tasks, including 95% confidence interval (C.I.) and Brier Skill Score (BSS) as a calibration measurement. Figure 20 shows heatmaps and clusters for each classification task with the methylation probe sites selected using ReliefF.

Table 26. Appendix I. Results
AUC classification performance for different classification tasks.

Classification Task	Omic	Feature selection with ReliefF			Feature selection with Limma		
		AUC	95% C.I.	BSS	AUC	95% C.I.	BSS
$TAHN_{ADC}$ vs. $Tumor_{ADC}$	G	0.99	0.97-1.0	0.89	0.94	0.82-1.0	0.73
	M	1.0	1.0-1.0	0.99	0.81	0.58-0.97	0.17
$TAHN_{SCC}$ vs. $Tumor_{SCC}$	M	1.0	0.99-1.0	0.94	0.99	0.96-1.0	0.66
$Tumor_{ADC}$ vs. $Tumor_{SCC}$	G	0.89	0.83-0.96	0.29	0.90	0.89-0.9	0.81
	M	0.97	0.94-0.99	0.71	0.89	0.74-1.0	0.38
$TAHN_{ADC}$ vs. $TAHN_{SCC}$	M	1.0	1.0-1.0	0.92	1.0	1.0-1.0	0.99
$TAHN\text{-}Tumor_{ADC}$ vs. $TAHN\text{-}Tumor_{SCC}$	M	0.92	0.89-0.95	0.42	0.94	0.87-1.0	0.56

G: gene expression, M: DNA methylation. The Brier Skill Score is a measurement of calibration of the classifier. A positive value on the BSS means that the classifier is well calibrated. A baseline classification is the work by Chang and Ramoni (Chang & Ramoni 2009) which obtained an accuracy of 0.95 in the classification task $Tumor_{ADC}$ vs. $Tumor_{SCC}$.

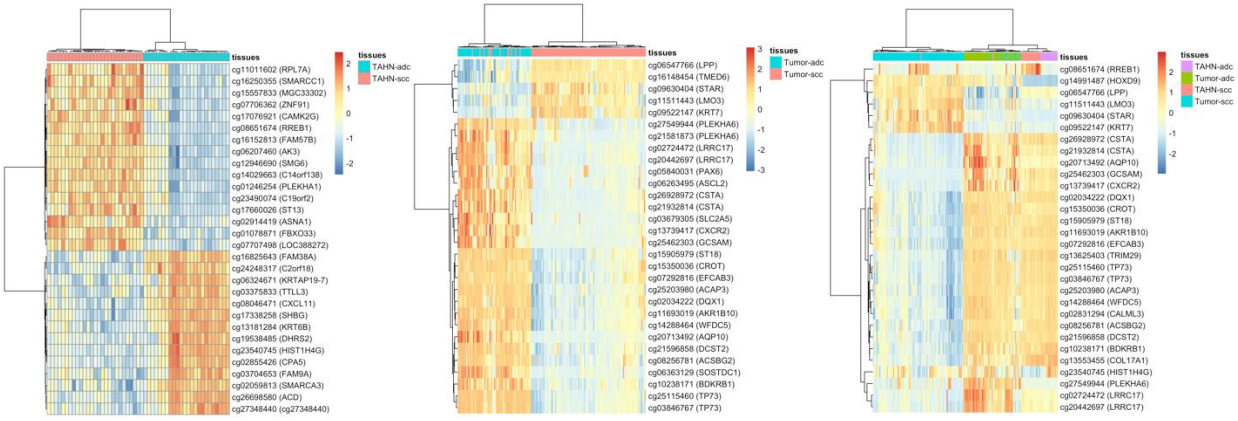

Figure 20. Appendix I. Heatmaps

Heatmaps for classification task (A) $TAHN_{ADC}$ vs. $TAHN_{SCC}$, (B) $Tumor_{ADC}$ vs. $Tumor_{SCC}$, and (**C**) TAHN-$Tumor_{ADC}$ vs. TAHN-$Tumor_{SCC}$ using the ReliefF feature selection algorithm. In the vertical axis the corresponding methylation site and gene symbol (in parenthesis) are shown. Some methylation sites do not lie in a particular gene, therefore, no symbol is provided. When multiple methylation sites are selected for the same gene, these sites should have similar methylation intensity, for it to be included. In the horizontal axis, a color-coded representation of the tissue samples is provided. Two distinct groups are observed in all three heatmaps. Cluster purity (accuracy by classification using clustering) for each task is calculated to be 1.0, 0.94, and 0.85 respectively.

The genes found by ReliefF in the classification task of TAHN-$Tumor_{ADC}$ vs TAHN-$Tumor_{SCC}$ using IPA® were analyzed. The results of the IPA® core analysis show a significant association between ReliefF-selected genes and the following diseases: cancer (25 out of 27) connective tissue disorder (13 out of 27), dermatological diseases and conditions (13 out of 27). Interestingly, the ReliefF-selected genes (19 out of 27) are associated with either adenocarcinoma (16 genes), squamous-cell carcinoma (4 genes), or carcinoma of the lung (4 genes). The list of genes and their associations can be seen in Table 27. Using these interesting 19 genes, we generated a gene interaction network to graphically visualize the relationships between genes and the disease class (adenocarcinoma, squamous-cell carcinoma, and carcinoma of the lung). The network is illustrated in Figure 21.

Table 27. Appendix I. Gene Enrichment

Genes selected for the classification task of TAHN-Tumor$_{ADC}$ Vs. TAHN-Tumor$_{SCC}$.

Gene Symbol	Gene Name	Known Literature Evidence to Cancer
ST18	suppression of tumorigenicity 18, zinc finger	Yes (Forbes et al. 2015)
CSTA	cystatin A (stefin A)	Yes (Forbes et al. 2015; Costea et al. 2013)
LPP	LIM domain containing preferred translocation partner in lipoma	Yes (Forbes et al. 2015)
CROT	carnitine O-octanoyltransferase	Yes (Forbes et al. 2015)
BDKRB1	bradykinin receptor B1	Yes (Dlamini & Bhoola 2005)
AKR1B10	aldo-keto reductase family 1, member B10 (aldose reductase)	Yes (B. Kim et al. 2007)
TP73	tumor protein p73	Yes (Flores et al. 2005; Lu et al. 2011; Tomasini et al. 2008)
EFCAB3	EF-hand calcium binding domain 3	Yes
RREB1	ras responsive element binding protein 1	Yes (Forbes et al. 2015)
HIST1H4G	histone cluster 1, H4g	No
STAR	steroidogenic acute regulatory protein	Yes
ACSBG2	acyl-CoA synthetase bubblegum family member 2	Yes (Forbes et al. 2015)
DQX1	DEAQ box RNA-dependent ATPase 1	Yes (Forbes et al. 2015)
AQP10	aquaporin 10	Yes (Forbes et al. 2015)
PLEKHA6	pleckstrin homology domain containing, family A member 6	Yes (The Cancer Genome Atlas Research Network, Getz, Saksena, Zhang, et al. 2012; Seshagiri et al. 2012)
GCSAM	germinal center-associated, signaling and motility	No
WFDC5	WAP four-disulfide core domain 5	Yes
KRT7	keratin 7, type II	Yes (Laurell et al. 2006)
DCST2	DC-STAMP domain containing 2	Yes (Forbes et al. 2015)
CALML3	calmodulin-like 3	Yes
ACAP3	ArfGAP with coiled-coil, ankyrin repeat and PH domains 3	Yes
LRRC17	leucine rich repeat containing 17	Yes (Forbes et al. 2015)
TRIM29	tripartite motif containing 29	Yes (L. Wang et al. 2015)
CXCR2	chemokine (C-X-C motif) receptor 2	Yes (Forbes et al. 2015; Raghuwanshi et al. 2008; Raghuwanshi et al. 2013)
HOXD9	homeobox D9	Yes (Pickering et al. 2013)
COL17A1	collagen, type XVII, alpha 1	Yes (Forbes et al. 2015)
LMO3	LIM domain only 3 (rhombotin-like 2)	Yes

The list of genes is ordered by their ranks, as selected by ReliefF for the classification task of TAHN-Tumor$_{ADC}$ Vs. TAHN-Tumor$_{SCC}$. The Entrez gene symbol, and the gene name are listed in the first two columns respectively. The 'Known Literature Evidence to Cancer' indicates if links to cancer were detected by the IPA® software. Citations are provided to literature indicating links to— adenocarcinoma, squamous-cell carcinoma, and carcinoma in lung.

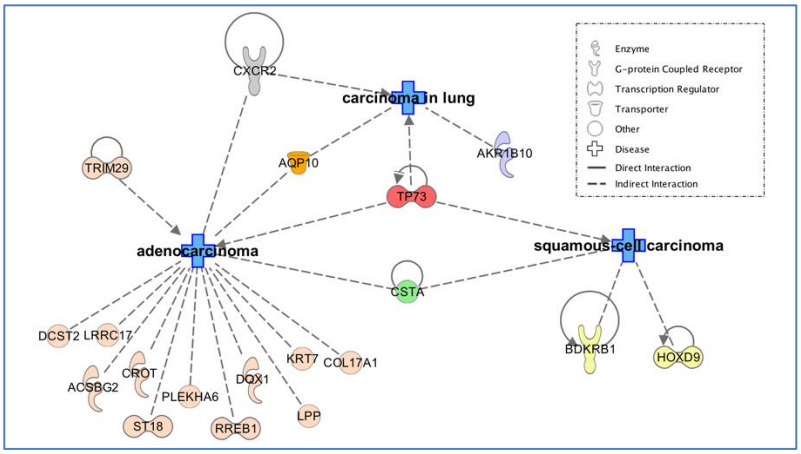

Figure 21. Appendix I. Pathway analysis

Gene interaction network generated by the IPA® software. It shows an analysis of the genes found by ReliefF in the classification task TAHN-Tumor$_{ADC}$ vs TAHN-Tumor$_{SCC}$. Three diseases are being shown (carcinoma of the lung, adenocarcinoma, and squamous cell carcinoma), and the selected genes were connected to these diseases via literature evidence that indicates: direct interactions (straight line), or indirect interactions (dashed line). Some of those interactions have arrow-heads indicating causation (e.g. BDKRB1). An arrow-head with a bar (i.e. TP73) indicates inhibition.

DISCUSSION

Evaluation of Classifiers. The classification performance for all models is high (AUC $\geq$ 0.81), with positive calibration (BSS $>$ 0). This positive calibration is a good indication that the models will perform well for other cases, and that they were not biased by the distribution of the data.

The value of using TAHN tissue for classification. Lung cancer patients could benefit with a potentially novel approach for subtyping. The diagnosis of adenocarcinoma vs. squamous cell carcinoma is routinely accomplished using histology supplemented by immunohistochemistry (TTF-1 and p63/p40). It is therefore not likely that the approach presented here would change this practice, which is well established, quick and inexpensive. Rather, this approach suggests that the use of epigenomic changes could help in the small

number of tumors which remain difficult to classify. However, the primary importance of this work may be in providing additional understanding of the origins of squamous cell and adenocarcinomas, which suggest that these phenotypes are associated with, or perhaps even derived from, different epigenomic phenotypes. Epigenomic alterations, in the form of DNA methylation, prevent the binding of transcription machinery, resulting in gene silencing (Brzeziańska et al. 2013). Moreover, DNA methylation signatures are different between tissue types and between tumors and normal surrounding tissue (Szyf 2012). In this study, tumor-adjacent histologically normal tissue samples were used to classify lung cancer subtypes with excellent results. This classification performance was achieved when no tumor samples were involved ($TAHN_{ADC}$ vs. $TAHN_{SCC}$), and when a mix of tissue was used ($TAHN\text{-}Tumor_{ADC}$ vs. $TAHN\text{-}Tumor_{SCC}$). The high AUC results are an indication of the diagnostic potential of this technology.

APPENDIX B

J2K APPLIED TO BREAST CANCER STAGING

This Appendix follows the example in Section 4. The classification performance, under a 10-fold cross validation, of the single-omic and multi-omic data integration models is shown in Table 28 for the classification tasks of distinguishing between samples from early stage and advanced stage patients. Each model is evaluated for area under the receiver operating characteristic (AUC), 95% confidence interval (C.I.), and Brier Skill Score (BSS) for calibration of the model. The multi-omic models are shown in Section 4 (Figure 15), except for the mixture model, which is shown here in Figure 22.

Table 28. Appendix II. Classification Performance

Model Name	AUC	95% C.I.	BSS
Single-Omic			
Gene Expression model	0.64	0.45-0.83	-0.09
DNA Methylation model	0.6	0.47-0.74	-0.19
Multi-Omic			
Mixture model	0.65	0.47-0.83	0.01
Independent model	0.89	0.79-0.99	-0.18
Three-way model	0.88	0.78-0.99	-0.02
Cascade model	0.88	0.78-0.99	-0.01

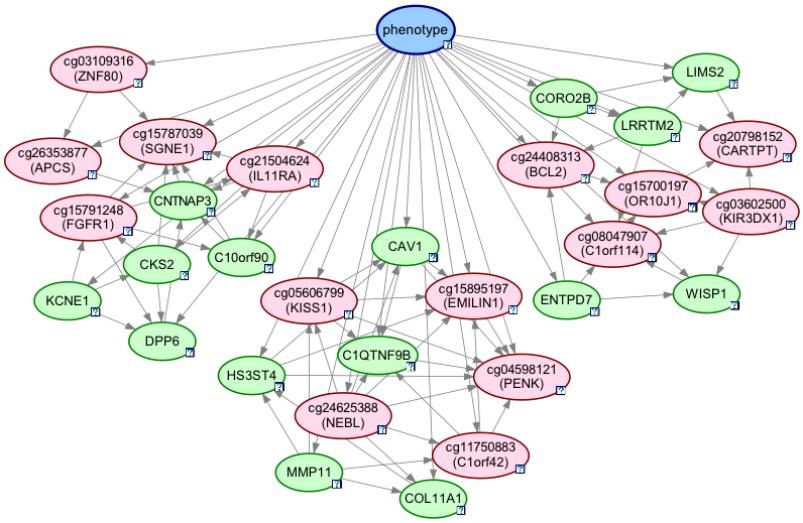

Figure 22. Appendix II. Mixture Model

The selected set of differentially expressed genes and methylation probes is different between the mixture model and the models with latent variables. To build a ANB structure, the mixture model searches over all possible variables in both 'omics', while the other three multi-omic models first build single-omic ANB structures, and then integrate them using latent variables. There is no intersection in the genes selected using each method. Table 29 presents a list of genes selected from the multi-omic models that have been associated with breast cancer and cancer progression in the literature

Table 29. Appendix II. List of genes from multi-omic models involved in breast cancer

M/G	Entrez Gene Symbol	Gene Name
Mixture model		
G	CAV1	Caveolin 1, caveolae protein, 22kDa
M	FGFR1	Fibroblast growth factor receptor 1
G	MMP11	Matrix metallopeptidase 11
M	BCL2	B-cell CLL/lymphoma 2
G	COL11A1	Collagen, type XI, alpha 1
M	KISS1	KiSS-1 metastasis-suppressor
MODI models		
G	RRM2	Ribonucleotide reductase M2
M	AKAP10	A kinase (PRKA) anchor protein 10
G	AQP4	Aquaporin 4
G	BRINP1	Bone morphogenetic protein/retinoic acid inducible neural-specific 1
G	TLR3	Toll-like receptor 3
G	HOXC5	Homeobox C5

Genes found by the mixture model. The caveolin 1, caveolae protein, 22kDa gene (CAV1) is a tumor suppressor gene candidate and has been shown to be differentially methylated among breast cancer subtypes (Z. Li et al. 2015). Expression of the fibroblast growth factor receptor 1 gene (FGFR1) in triple–negative breast cancers is independently prognostic of overall survival (Cheng et al. 2015). The matrix metallopeptidase 11 gene (MMP11) encodes a member of the proteins of the matrix metalloproteinase (MMP) family, which participate in normal physiological processes, as well as in disease processes, such as metastasis (Entrez). Expression of MMP11 by intratumoral mononuclear inflammatory cells has been associated with distant metastasis development and worse prognosis in breast cancer (Eiró et al. 2012). The B-cell CLL/lymphoma 2 gene (BCL2) is thought to be the cause of follicular lymphoma. It is overexpressed in ~75% of breast cancer (Merino et al. 2015), where it is also predictive of lymph node metastasis (H. Kim et al. 2015) and its hypermethylation has been associated with favorable response to endocrine treatment (Stone et al. 2013). The collagen, type XI, alpha 1 gene (COL11A1) is highly expressed by activated stromal cells of breast tumors, and correlates with tumor progression and lymph node metastasis (Vázquez-Villa et al. 2015), is therefore a marker of invasiveness in breast tumor lesions (Freire et al. 2014). The KiSS-1 metastasis-suppressor gene (KISS1) is known to suppress metastases of melanomas and breast carcinomas (Entrez Gene) via inhibition of breast cancer cell invasiveness by its protein product, kisspeptin (Tan et al. 2014). Kisspeptin-10 (KP-10) is a shorter fragment of KISS1 (Song & Zhao 2015), which suppresses breast cancer and human umbilical vein endothelial cell (HUVEC) growth both in vivo and in vitro. KP-10 is a novel regulator of EMT in breast cancer cells.

Genes found by the MODI models. Increased mRNA levels of the gene ribonucleotide reductase M2 (RRM2) have been associated with poor patient outcome in a dose-dependent

manner, with prognostic power comparable to that of multiple gene signatures, and superior to TNM stage (H. Zhang et al. 2014). The A kinase anchor protein 10 gene (AKAP10) regulates protein kinase A (PKA). Overexpression of PKA is a hallmark of the great majority of human cancers including breast cancer (Wirtenberger et al. 2006) and expression of AKAP10 has been correlated with deeper tumor invasion, lymph nodes metastasis and advanced tumor stage in colorectal cancer (M. Wang et al. 2013). The gene aquaporin 4 (AQP4) is markedly underexpressed in various cancers, including breast cancer (Shi et al. 2012). The gene bone morphogenetic protein/retinoic acid inducible neural-specific 1 (BRINP1, also known as DBC1) is reported to be hypermethylated in breast cancer and is high performing in cancer prediction (Z. Li et al. 2015). Stimulation of the toll-like receptor 3, encoded by the toll-like receptor 3 gene (TLR3) has been found to promote breast cancer cells toward a cancer stem cell phenotype in vitro and in vivo (Jia et al. 2015). The expression level of the homeobox C5 gene (HOXC5) was lower in breast cancer tissues with mutated-type p53 than in normal and cancerous tissues with wild-type p53, suggesting that aberrant expression of this gene is related to the development of breast cancer (Makiyama et al. 2005).

BIBLIOGRAPHY

Alberg, A.J. et al., 2007. Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, 132(3 Suppl), pp.29S–55S.

Aliferis, C.F., Hardin, D. & Massion, P.P., 2002. Machine learning models for lung cancer classification using array comparative genomic hybridization. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp.7–11.

Alix-Panabières, C., Schwarzenbach, H. & Pantel, K., 2012. Circulating Tumor Cells and Circulating Tumor DNA. *dx.doi.org*.

Allison, D.B. et al., 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews: Genetics*, 7(1), pp.55–65.

Arnett, D.B., Laverie, D.A. & Meiers, A., 2003. Developing parsimonious retailer equity indexes using partial least squares analysis: a method and applications. *Journal of Retailing*, 79(3), pp.161–170.

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), pp.25–29.

Bagaria, S.P. et al., 2014. Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. *JAMA Surgery*, 149(2), pp.125–129.

Balbin, O.A. et al., 2013. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nature Communications*, 4, pp.2617–2617.

Bast, R.C., Hennessy, B. & Mills, G.B., 2009. The biology of ovarian cancer: new opportunities for translation. *Nature Reviews Cancer*, 9(6), pp.415–428.

Bastien, R.R.L. et al., 2012. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC medical genomics*, 5(1), p.44.

Baylin, S.B. et al., 2001. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics*, 10(7), pp.687–692.

Best, M.G. et al., 2015. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell*.

Bettegowda, C. et al., 2014. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Science Translational Medicine*, 6(224), pp.224ra24–224ra24.

Bonneau, R. et al., 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, 7(5), p.R36.

Bouckaert, R.R., 2008. *Bayesian networks in Weka*, Computer Science Department. University of Waikato.

Braun, R., 2014. Systems Analysis of High-Throughput Data. In S. J. Corey, M. Kimmel, & J. N. Leonard, eds. *A Systems Biology Approach to Blood*. Advances in Experimental Medicine and Biology. New York, NY: Springer New York, pp. 153–187.

Brzeziańska, E., Dutkowska, A. & Antczak, A., 2013. The significance of epigenetic alterations in lung carcinogenesis. *Molecular biology reports*, 40(1), pp.309–325.

Buhule, O.D. et al., 2014. Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Frontiers in genetics*, 5, p.354.

Bullard, J.H. et al., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1), p.94.

Cai, Z. et al., 2014. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular bioSystems*.

Capra, J.A. & Kostka, D., 2014. Modeling DNA methylation dynamics with approaches from phylogenetics. *Bioinformatics (Oxford, England)*, 30(17), pp.i408–i414.

Carpenter, J.M., 1988. Choosing among multiple equally parsimonious cladograms. *Cladistics*.

Chang, H.-H. & Ramoni, M.F., 2009. Transcriptional network classifiers. *BMC Bioinformatics*, 10 Suppl 9, pp.S1–S1.

Cheng, C.L. et al., 2015. Expression of FGFR1 is an independent prognostic factor in triple-negative breast cancer. *Breast cancer research and treatment*, 151(1), pp.99–111.

Chiu, C.-H. et al., 2014. Should EGFR mutations be tested in advanced lung squamous cell carcinomas to guide frontline treatment? *Cancer chemotherapy and pharmacology*, 74(4), pp.661–665.

College of American Pathologists, 2011a. Lung Adenocarcinoma. pp.1–2.

College of American Pathologists, 2011b. Lung Squamous Cell Carcinoma. pp.1–2.

Compton, C.C. et al., 2012. *AJCC cancer staging atlas: a companion to the seventh editions of the AJCC cancer staging manual and handbook*,

Cooper, G.F. et al., 2010. An efficient bayesian method for predicting clinical outcomes from genome-wide data. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA*

Symposium, 2010, pp.127–131.

Correia, C. et al., 2014. Hope for GWAS: relevant risk genes uncovered from GWAS statistical noise. *International journal of molecular sciences*, 15(10), pp.17601–17621.

Cossalter, M., Mengshoel, O. & Selker, T., 2011. Visualizing and Understanding Large-Scale Bayesian Networks. pp.1–10.

Costea, D.E. et al., 2013. Identification of two distinct carcinoma-associated fibroblast subtypes with differential tumor-promoting abilities in oral squamous cell carcinoma. *Cancer Research*, 73(13), pp.3888–3901.

Coughlin, S.S., 2014. Toward a Road Map for Global -Omics: A Primer on -Omic Technologies. *American journal of epidemiology*, p.kwu262.

Cyr, A., 2015. Toward less-invasive management of early-stage breast cancer. *Journal of the National Comprehensive Cancer Network : JNCCN*, 13(5 Suppl), pp.646–648.

Dacic, S. et al., 2010. Clinicopathological predictors of EGFR/KRAS mutational status in primary lung adenocarcinomas. *Modern Pathology*, 23(2), pp.159–168.

Daly, R., Shen, Q. & Aitken, S., 2011. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26(02), pp.99–157.

de Kok, J.B. et al., 2005. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Laboratory investigation; a journal of technical methods and pathology*, 85(1), pp.154–159.

De Souto, M.C.P., Jaskowiak, P.A. & Costa, I.G., 2015. Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics*, 16(1), p.64.

Demir, L. et al., 2014. Hormone receptor, HER2/NEU and EGFR expression in ovarian carcinoma--is here a prognostic phenotype? *Asian Pacific journal of cancer prevention : APJCP*, 15(22), pp.9739–9745.

Dempster, A.P., Laird, N.M. & Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*, 39(1), pp.1–38.

Diaz Prado, J.A., Lopez Pineda, A. & Cruz Ramos, M.P., 2010. Corporate Technology Intelligence Research System through Recycling Public Patent Databases. *Communications of the IBIMA*, 10, pp.1–10.

Diaz, L.A., Jr & Bardelli, A., 2014. Liquid Biopsies: Genotyping Circulating Tumor DNA. *Journal of Clinical Oncology*, 32(6), pp.579–586.

Diehl, F. et al., 2008. Circulating mutant DNA to assess tumor dynamics. *Nature Medicine*, 14(9), pp.985–990.

Dlamini, Z. & Bhoola, K.D., 2005. Upregulation of tissue kallikrein, kinin B1 receptor, and kinin B2 receptor in mast and giant cells infiltrating oesophageal squamous cell carcinoma. *Journal of Clinical Pathology*, 58(9), pp.915–922.

Dooms, C. et al., 2014. Suitability of small bronchoscopic tumour specimens for lung cancer genotyping. *Respiration*, 88(5), pp.371–377.

Dougherty, J., Kohavi, R. & Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. *... learning: proceedings of ...*.

Druzdzel, M.J., 1999. SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: a development environment for graphical decision-theoretic models. *AAAI/IAAI*.

Dudoit, S., Fridlyand, J. & Speed, T.P., 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97, pp.77–87.

Dumeaux, V. et al., 2008. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. *Breast Cancer Research (Online Edition)*, 10(1), p.R13.

Eiró, N. et al., 2012. Relationship between the inflammatory molecular profile of breast carcinomas and distant metastasis development. *PLoS ONE*, 7(11), p.e49047.

Ezkurdia, I. et al., 2013. The shrinking human protein coding complement: are there now fewer than 20,000 genes? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.–.

Fayyad, U. & Irani, K., 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning.

Flores, E.R. et al., 2005. Tumor predisposition in mice mutant for p63 and p73: Evidence for broader tumor suppressor functions for the p53 family. *Cancer Cell*, 7(4), pp.363–373.

Forbes, S.A. et al., 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(Database issue), pp.D805–D811.

Freire, J. et al., 2014. Collagen, type XI, alpha 1: an accurate marker for differential diagnosis of breast carcinoma invasiveness in core needle biopsies. *Pathology, research and practice*, 210(12), pp.879–884.

Friedman, N., Geiger, D. & Goldszmidt, M., 1997. Bayesian Network Classifiers. *Machine Learning*, 29(2-3), pp.131–163.

Fu, C. et al., 2012. *Identification of oncogenic genes for colon adenocarcinoma from genomics data*, IEEE.

Garcia, S. et al., 2013. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *Knowledge and Data Engineering, IEEE Transactions on*,

25(4), pp.734–750.

Goemans, M.X. & Bertsimas, D.J., 1993. Survivable networks, linear programming relaxations and the parsimonious property. *Mathematical Programming*, 60(1-3), pp.145–166.

Gopalakrishnan, V. et al., 2010. Bayesian rule learning for biomedical data mining. *Bioinformatics (Oxford, England)*, 26(5), pp.668–675.

Greene, C.S. et al., 2009. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining*, 2(1), pp.1–9.

Guvenen, F., 2009. A Parsimonious Macroeconomic Model for Asset Pricing. *Econometrica*, 77(6), pp.1711–1750.

Guyon, I. et al., 2010. Model Selection: Beyond the Bayesian/Frequentist Divide. *The Journal of Machine Learning Research*, 11.

Hall, M. et al., 2009. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1), p.10.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. Unsupervised Learning. In *The Elements of Statistical Learning*. Springer New York, pp. 485–585.

Ho, T.-H. & Chong, J.-K., 2003. A Parsimonious Model of Stockkeeping-Unit Choice. *Journal of Marketing Research*, 40(3), pp.351–365.

Huang, D.W. et al., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35(Web Server issue), pp.W169–75.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), pp.1–13.

Jayawardana, K. et al., 2015. Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. *International journal of cancer. Journal international du cancer*, 136(4), pp.863–874.

Jemal, A.A. et al., 2011. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2), pp.69–90.

Jia, D. et al., 2015. β-Catenin and NF-κB co-activation triggered by TLR3 stimulation facilitates stem cell-like phenotypes in breast cancer. *Cell death and differentiation*, 22(2), pp.298–310.

Jiang, X. et al., 2014. A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. *Journal of the American Medical Informatics Association*, 21(e2), pp.e312–e319.

John, G.H. & Langley, P., 1995. *Estimating continuous distributions in Bayesian classifiers*, Morgan Kaufmann Publishers Inc.

Kairov, U. et al., 2012. Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures. *Bioinformation*, 8(16), pp.773–776.

Karimpour-Fard, A., Epperson, L.E. & Hunter, L.E., 2015. A survey of computational tools for downstream analysis of proteomic and other omic datasets. *Human genomics*, 9(1), p.28.

Karr, J.R. et al., 2012. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2), pp.389–401.

Kerr, M.K., 2003. Design considerations for efficient and effective microarray studies. *Biometrics*, 59(4), pp.822–828.

Kim, B. et al., 2007. Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Research*, 67(15), pp.7431–7438.

Kim, H. et al., 2015. Expression of SIRT1 and apoptosis-related proteins is predictive for lymph node metastasis and disease-free survival in luminal A breast cancer. *Virchows Archiv : an international journal of pathology*, pp.1–8.

Kira, K. & Rendell, L.A., 1992. A practical approach to feature selection. In Proceedings of the ninth international workshop on …. pp. 249–256.

Kononenko, I., 1995. On Biases in Estimating Multi-Valued Attributes. *IJCAI*, pp.1034–1040.

Kononenko, I., Šimec, E. & Robnik-Šikonja, M., 1997. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence*, 7(1), pp.39–55.

Laurell, H. et al., 2006. Identification of biomarkers of human pancreatic adenocarcinomas by expression profiling and validation with gene expression analysis in endoscopic ultrasound-guided fine needle aspiration samples. *World Journal of Gastroenterology*, 12(21), pp.3344–3351.

Lauritzen, S.L. & Spiegelhalter, D.J., 1988. *Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems*,

Li, J. et al., 2015. Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics*, 16 Suppl 5(Suppl 5), p.S10.

Li, M. & Vitanyi, P., 2013. *An Introduction to Kolmogorov Complexity and Its Applications*,

Li, Z. et al., 2015. Methylation profiling of 48 candidate genes in tumor and matched normal tissues from breast cancer patients. *Breast cancer research and treatment*, 149(3), pp.767–779.

Lin, D. et al., 2014. Integrative analysis of multiple diverse omics datasets by sparse group

multitask regression. *Frontiers in cell and developmental biology*, 2, pp.62–62.

Lin, J.H. & Haug, P.J., 2008. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, 41(1), pp.1–14.

Liu, J., Yang, X.Y. & Shi, W.J., 2014. Identifying differentially expressed genes and pathways in two types of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. *Genetics and Molecular Research*, 13(1), pp.95–102.

Liu, J., Zhao, D. & Fan, R., 2015. Shared and unique mutational gene co-occurrences in cancers. *Biochemical and biophysical research communications*, 465(4), pp.777–783.

Liu, Y. et al., 2013. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC Systems Biology*, 7(1), pp.14–14.

Lopez Pineda, A. et al., 2015. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *Journal of Biomedical Informatics*.

Lopez Pineda, A. et al., 2013. Machine Learning Classification of Non-Small Cell Lung Cancer Subtypes from Gene Methylation Data. In Great Lakes Bioinformatics Conference. Pittsburgh, PA, pp. 1–1.

Lowry, R.R., 2014. *Concepts and Applications of Inferential Statistics*,

Lu, H. et al., 2011. TNF-alpha Promotes c-REL/Delta Np63 alpha Interaction and TAp73 Dissociation from Key Genes That Mediate Growth Arrest and Apoptosis in Head and Neck Cancer. *Cancer Research*, 71(21), pp.6867–6877.

Lustgarten, J.L. et al., 2011. Application of an efficient Bayesian discretization method to biomedical data. *BMC Bioinformatics*, 12(1), p.309.

Lyman, G.H. et al., 2007. Impact of a 21-gene RT-PCR assay on treatment decisions in early-stage breast cancer: an economic analysis based on prognostic and predictive validation studies. *Cancer*, 109(6), pp.1011–1018.

Maciejewski, H., 2014. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, 15(4), pp.504–518.

Madsen, A.L., 2004. An empirical evaluation of possible variations of lazy propagation. In Proceedings of the 20th conference on Uncertainty in ….

Makino, K. & Uno, T., 2004. New Algorithms for Enumerating All Maximal Cliques. In *Algorithm Theory - SWAT 2004*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 260–272.

Makiyama, K. et al., 2005. Aberrant expression of HOX genes in human invasive breast carcinoma. *Oncology Reports*, 13(4), pp.673–679.

Maksimovic, J. et al., 2015. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Research*, 43(16), pp.e106–e106.

Martini, P. et al., 2013. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*, 41(1), pp.e19–e19.

Mason, C.E., Porter, S.G. & Smith, T.M., 2014. Characterizing Multi-omic Data in Systems Biology. *Systems Analysis of Human Multigene …*, pp.15–38.

McAuley, J.J., Caetano, T.S. & Barbosa, M.S., 2008. Graph rigidity, Cyclic Belief Propagation and Point Pattern Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, pp.2047–2054.

McCall, M.N., Jaffee, H.A. & Irizarry, R.A., 2012. fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. *Bioinformatics (Oxford, England)*, 28(23), pp.3153–3154.

Merino, D. et al., 2015. Targeting BCL-2 to enhance vulnerability to therapy in estrogen receptor-positive breast cancer. *Oncogene*.

Miller, D.S. et al., 2009. Phase II evaluation of pemetrexed in the treatment of recurrent or persistent platinum-resistant ovarian or primary peritoneal carcinoma: a study of the Gynecologic Oncology Group. *Journal of Clinical Oncology. Official journal of the American Society of Clinical Oncology*, 27(16), pp.2686–2691.

Miller, E. et al., 2014. Current treatment of early breast cancer: adjuvant and neoadjuvant therapy. *F1000Research*, 3, p.198.

Miller, G.A., 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2), pp.81–97.

Mitchell, T.M., 2015. Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression. In *Machine Learning*. pp. 1–17.

Mitra, K. et al., 2013. Integrative approaches for finding modular structure in biological networks. *Nature Reviews: Genetics*, 14(10), pp.719–732.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), pp.621–628.

Moulos, P. et al., 2011. Unifying the integration, analysis and interpretation of multi-omic datasets: exploration of the disease networks of Obstructive Nephropathy in children. *International Conference of the IEEE Engineering in Medicine and Biology Society. Proceedings*, 2011, pp.3716–3719.

Myers, E.R. et al., 2015. Benefits and Harms of Breast Cancer Screening: A Systematic Review. *JAMA*, 314(15), pp.1615–1634.

Neapolitan, R.E., 2012. *Probabilistic Reasoning in Expert Systems*,

Nelson, C.R. & Siegel, A.F., 1987. Parsimonious Modeling of Yield Curves on JSTOR. *Journal of business*.

Ogata, H. et al., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), pp.29–34.

Olesen, K.G. & Madsen, A.L., 2002. Maximal prime subgraph decomposition of Bayesian networks. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 32(1), pp.21–31.

Orucevic, A. et al., 2015. Is the TNM Staging System for Breast Cancer Still Relevant in the Era of Biomarkers and Emerging Personalized Medicine for Breast Cancer - An Institution's 10-year Experience. *The breast journal*, 21(2), pp.147–154.

Pakzad, P. & Anantharam, V., 2005. Estimation and marginalization using the Kikuchi approximation methods. *Neural computation*, 17(8), pp.1836–1873.

Palsson, B. & Zengler, K., 2010. The challenges of integrating multi-omic data sets. *Nature Chemical Biology*, 6(11), pp.787–789.

Parker, J.S. et al., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology. Official journal of the American Society of Clinical Oncology*, 27(8), pp.1160–1167.

Payne, S.H., 2015. The utility of protein and mRNA correlation. *Trends in biochemical sciences*, 40(1), pp.1–3.

Peixoto, L. et al., 2015. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Research*, 43(16), pp.7664–7674.

Pfeifer, G.P. & Rauch, T.A., 2009. DNA methylation patterns in lung carcinomas. *Seminars in Cancer Biology*, 19(3), pp.181–187.

Phillips, T., 2008. The Role of Methylation in Gene Expression | Learn Science at Scitable. *Nature Education*.

Pickering, C.R. et al., 2013. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discovery*, 3(7), pp.770–781.

Piorkowski, M., Sarafijanovic-Djukic, N. & Grossglauser, M., 2009. A parsimonious model of mobile partitioned networks with clustering. In 2009 First International Communication Systems and Networks and Workshops (COMSNETS). IEEE, pp. 1–10.

Pruteanu-Malinici, I., Majoros, W.H. & Ohler, U., 2013. Automated annotation of gene expression image sequences via non-parametric factor analysis and conditional random fields. *Bioinformatics (Oxford, England)*, 29(13), pp.i27–35.

Pufulete, M. et al., 2005. Effect of folic acid supplementation on genomic DNA methylation in patients with colorectal adenoma. *Gut*, 54(5), pp.648–653.

Putluri, N. et al., 2014. Pathway-centric integrative analysis identifies RRM2 as a prognostic marker in breast cancer associated with poor survival and tamoxifen resistance. *Neoplasia (New York, N.Y.)*, 16(5), pp.390–402.

Qabaja, A. et al., 2014. Prediction of novel drug indications using network driven biological data prioritization and integration. *Journal of cheminformatics*, 6(1), p.1.

Quinlan, J.R., 1986. Induction of Decision Trees. *Machine Learning*, 1(1), pp.81–106.

Raghuwanshi, S.K. et al., 2008. Depletion of beta-arrestin-2 promotes tumor growth and angiogenesis in a murine model of lung cancer. *J Immunol*, 180(8), pp.5699–5706.

Raghuwanshi, S.K. et al., 2013. G protein-coupled receptor kinase 6 deficiency promotes angiogenesis, tumor progression, and metastasis. *J Immunol*, 190(10), pp.5329–5336.

Rauch, T.A. et al., 2012. DNA methylation biomarkers for lung cancer. *Tumor Biology*, 33(2), pp.287–296.

Resendis-Antonio, O. et al., 2012. Functional modules, structural topology, and optimal activity in metabolic networks. *PLoS computational biology*, 8(10), p.e1002720.

Rex, D.K. et al., 2011. The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps. *Gastrointestinal Endoscopy*, 73(3), pp.419–422.

Richer, A.L. et al., 2015. Genomic profiling toward precision medicine in non-small cell lung cancer: getting beyond EGFR. *Pharmacogenomics and personalized medicine*, 8, pp.63–79.

Risso, D. et al., 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9), pp.896–902.

Ritchie, M.D. et al., 2015. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews: Genetics*, 16(2), pp.85–97.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), p.R25.

Robnik-Šikonja, M. & Kononenko, I., 1997. An adaptation of Relief for attribute estimation in regression. *Machine Learning: Proceedings of …*.

Robson, M.E. et al., 2010. American Society of Clinical Oncology Policy Statement Update: Genetic and Genomic Testing for Cancer Susceptibility. *Journal of Clinical Oncology*, 28(5), pp.893–901.

Saeys, Y., Inza, I. & Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Journal of Gerontology*, 23(19), pp.2507–2517.

Saleem, M. et al., 2013. *Linked cancer genome atlas database*, ACM.

Sanchez-Palencia, A. et al., 2011. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International journal of cancer. Journal international du cancer*, 129(2), pp.355–364.

Sawhney, M.S. & Eliashberg, J., 1996. A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, 15(2), pp.113–131.

Seasholtz, M.B. & Kowalski, B., 1993. The parsimony principle applied to multivariate calibration. *Analytica Chimica Acta*, 277(2), pp.165–177.

Sebastiani, P., Solovieff, N. & Sun, J.X., 2012. Naïve Bayesian Classifier and Genetic Risk Score for Genetic Risk Prediction of a Categorical Trait: Not so Different after all! *Frontiers in genetics*, 3, pp.26–26.

Seong, S.J. et al., 2015. Controversies in borderline ovarian tumors. *Journal of gynecologic oncology*.

Serang, O., 2014. The probabilistic convolution tree: efficient exact Bayesian inference for faster LC-MS/MS protein inference. *PLoS ONE*, 9(3), pp.e91507–e91507.

Serang, O. & Noble, W.S., 2012. Faster Mass Spectrometry-Based Protein Inference: Junction Trees Are More Efficient than Sampling and Marginalization by Enumeration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(3), pp.809–817.

Seshagiri, S. et al., 2012. Recurrent R-spondin fusions in colon cancer. *Nature*, 488(7413), pp.660–664.

Shi, Z. et al., 2012. Aquaporins in human breast cancer: identification and involvement in carcinogenesis of breast cancer. *Journal of surgical oncology*, 106(3), pp.267–272.

Siegel, R. et al., 2014. Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, 64(1), pp.9–29.

Slooten, K., 2011. Validation of DNA-based identification software by computation of pedigree likelihood ratios. *Forensic science international. Genetics*, 5(4), pp.308–315.

Smyth, G.K., 2004. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Soda, M.M. et al., 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153), pp.561–566.

Song, G.-Q. & Zhao, Y., 2015. Kisspeptin-10 inhibits the migration of breast cancer cells by regulating epithelial-mesenchymal transition. *Oncology Reports*, 33(2), pp.669–674.

Statnikov, A. et al., 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Machine Learning: Proceedings of …*.

Stetson, L.C. et al., 2014. Computational identification of multi-omic correlates of anticancer therapeutic response. *BMC Genomics*, 15(Suppl 7), p.S2.

Stewart, B.W. & Wild, C.P., 2014. *World Cancer Report 2014*,

Stone, A. et al., 2013. BCL-2 hypermethylation is a potential biomarker of sensitivity to antimitotic chemotherapy in endocrine-resistant breast cancer. *Molecular cancer therapeutics*, 12(9), pp.1874–1885.

Strogatz, S.H., 2001. Exploring complex networks. *Nature*, 410(6825), pp.268–276.

Subramanian, J. & Simon, R., 2010. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *Journal of the National Cancer Institute*, 102(7), pp.464–474.

Szyf, M., 2012. DNA methylation signatures for breast cancer classification and prognosis. *Genome medicine*, 4(3), pp.26–26.

Takeuchi, Y. et al., 2015. An alternative option for "resect and discard" strategy, using magnifying narrow-band imaging: a prospective “proof-of-principle” study. *Journal of gastroenterology*, 50(10), pp.1017–1026.

Tan, K. et al., 2014. KiSS1-induced GPR54 signaling inhibits breast cancer cell migration and epithelial-mesenchymal transition via protein kinase D1. *Current molecular medicine*, 14(5), pp.652–662.

Tarca, A.L., Bhatti, G. & Romero, R., 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE*, 8(11), p.e79217.

The Cancer Genome Atlas Research Network, 2012a. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417), pp.519–525.

The Cancer Genome Atlas Research Network, 2012b. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–337.

The Cancer Genome Atlas Research Network, 2014. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511), pp.543–550.

The Cancer Genome Atlas Research Network, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353), pp.609–615.

The Cancer Genome Atlas Research Network, Getz, G., Saksena, G., Park, P.J., et al., 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70.

The Cancer Genome Atlas Research Network, Getz, G., Saksena, G., Zhang, J., et al., 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–337.

Tieri, P. et al., 2014. Multi-omic landscape of rheumatoid arthritis: re-evaluation of drug adverse effects. *Frontiers in cell and developmental biology*, 2, pp.59–59.

Tomasini, R. et al., 2008. TAp73 knockout shows genomic instability with infertility and tumor suppressor functions. *Genes Dev*, 22(19), pp.2677–2691.

Totir, L.R., Fernando, R.L. & Abraham, J., 2009. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics, selection, evolution : GSE*, 41(1), p.52.

Vázquez-Villa, F. et al., 2015. COL11A1/(pro)collagen 11A1 expression is a remarkable biomarker of human invasive carcinoma-associated stromal cells and carcinoma progression. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 36(4), pp.2213–2222.

Wang, L. et al., 2015. ATDC induces an invasive switch in KRAS-induced pancreatic tumorigenesis. *Genes Dev*, 29(2), pp.171–183.

Wang, M. et al., 2013. A-Kinase Anchoring Proteins 10 Expression in Relation to 2073A/G Polymorphism and Tumor Progression in Patients with Colorectal Cancer. *Pathology & Oncology Research*, 19(3), pp.521–527.

Wang, W., Baladandayuthapani, V., Holmes, C.C., et al., 2013. Integrative network-based Bayesian analysis of diverse genomics data. *BMC Bioinformatics*, 14(Suppl 13), p.S8.

Wang, W., Baladandayuthapani, V., Morris, J.S., et al., 2013. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics (Oxford, England)*, 29(2), pp.149–159.

Wang, X. et al., 2003. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics (Oxford, England)*, 19(11), pp.1341–1347.

Wei, C., Li, J. & Bumgarner, R.E., 2004. Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, 5, pp.87–87.

Wei, W., Visweswaran, S. & Cooper, G.F., 2011. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *Journal of the American Medical Informatics Association*, 18(4), pp.370–375.

Weisstein, E.W., 2004a. Hypothesis Testing.

Weisstein, E.W., 2004b. *Paired t-Test -- from Wolfram MathWorld*, Available at: http://mathworld.wolfram.com/Pairedt-Test.html.

Wilks, D.S., 2011. *Statistical Methods in the Atmospheric Sciences*, Academic Press.

Wirtenberger, M. et al., 2006. The functional genetic variant Ile646Val located in the kinase binding domain of the A-kinase anchoring protein 10 is associated with familial breast cancer. *Carcinogenesis*, 28(2), pp.423–426.

World Health Organization, 2012. GLOBOCAN Cancer Fact Sheets: All Cancers (excluding nonmelanoma skin cancer). Estimated Incidence, Mortality and Prevalence Worldwide in 2012. (http://globocan.iarc.fr/), pp.1–6.

Zakharkin, S.O. et al., 2006. Optimal allocation of replicates for measurement evaluation studies. *Genomics, proteomics & bioinformatics*, 4(3), pp.196–202.

Zarocostas, J.J., 2010. Global cancer cases and deaths are set to rise by 70% in next 20 years. *British Medical Journal (Abstracts)*, 340, pp.c3041–c3041.

Zhang, H. et al., 2014. Prognostic and therapeutic significance of ribonucleotide reductase small subunit M2 in estrogen-negative breast cancers. *BMC cancer*, 14(1), p.664.

Zhang, Q., Burdette, J.E. & Wang, J.-P., 2014. Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*, 8(1), p.1338.

Zhang, W. et al., 2007. Pooling mRNA in microarray experiments and its effect on power. *Bioinformatics (Oxford, England)*, 23(10), pp.1217–1224.