SURVIVAL PREDICTION VIA PARTIAL ORDERING IN FEATURE SPACE AND SAMPLE SPACE

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

By
Mustafa Büyüközkan
March 2016

Survival Prediction via Partial Ordering in Feature Space and Sample Space
By Mustafa Büyüközkan
March 2016

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

Öznur Taştan(Advisor)

Tolga Can

Ercüment Çiçek

Approved for the Graduate School of Engineering and Science:

Levent Onural
Director of the Graduate School

ABSTRACT

SURVIVAL PREDICTION VIA PARTIAL ORDERING IN FEATURE SPACE AND SAMPLE SPACE

Mustafa Büyüközkan
M.S. in Computer Engineering
Advisor: Öznur Taştan
March 2016

Predicting the survival of a cancer patient is critical for choosing patient specific treatment strategies and is traditionally based on clinical or pathological factors such as patient age and tumor stage. In this thesis, we present two methodologies to build effective and interpretable survival models that utilize high-dimensional molecular profiles made available through next-gen sequencing technologies.

Firstly, we present a method that focuses on partial ordering in the feature space. Existing models rely on the individual molecular quantities recorded in tumors; however, cancer is a complex disease where molecular mechanisms are dysregulated in various ways. This study, based on a system level perspective, incorporates the partial ordering of molecules (POF) in lieu of individual quantities. This strategy not only unveils predictive features with direct relevance to the biological mechanism and but also yields better performance in survival prediction compared to multivariate ℓ_1 penalized Cox proportional hazard and Random Survival Forest models. Testing the partial order representation of features in the subgroup identification task, we find that these features yield groups of patients, which are more quantifiably distinct in terms of survival distributions.

Secondly, we develop a survival prediction method based on ranking and support vector machines – Ranking Survival Vector Machines (RsurVM). RsurVM obtains a pairwise ranking of the patient survival times by learning to rank. It focuses on optimizing the most commonly used metric concordance index and can handle the censored data without making any assumptions. Our extensive tests on the ovarian adenocarcinoma patient molecular data demonstrate that RsurVM achieves better survival predictions regardless of the input molecular data (mRNA,

protein, miRNA, Copy number variation and DNA methylation) than the two most commonly used methods: Cox-proportional hazards model and Random Survival Forest.

Keywords: survival estimation, pairwise ranking, partial ordering, biologically interpretible features.

ÖZET

ÖZNİTELİK YA DA ÖRNEKLEM UZAYINDA KISMİ SIRALAMA YOLUYLA SAĞKALIM TAHMİNLEME

Mustafa Büyüközkan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Öznur Taştan
Mart 2016

Bir hastanın sağkalımını tahminleme, hastaya uygun tedavi planını belirleyebilmek için kritik öneme sahiptir ve geleneksel olarak hastanın yaşı, tümör evresi gibi klinik ve patolojik değişkenlere dayandırılarak yapılmaktadır.

Bu tezde, yeni nesil dizilime yöntemleri ile elde edilen moleküler verileri kullanarak başarılı ve yorumlanabilir sağkalım tahminleme modellerine ulaştıran iki yöntem sunuyoruz. Mevcut sağkalım modellerinin çoğu, moleküllerin tekil ifadeleme değerlerinin tekil temsili üzerine kurulmaktadır. Fakat kanser, moleküler mekanizmaların çok farklı yollarla bozulması ile meydana çıkan karmaşık bir hastalıktır. Bu çalışmada sistemsel biyoloji yaklaşım ile, moleküler değerlerin birbirlerine göre sıralamasını göz önünde bulunduran bir öznitelik temsili öneriyoruz. Önerdiğimiz model moleküler özniteliklerin doğrudan bağlantılı oldukları biyolojik mekanizmaları açığa çıkmarmayı mümkün kılmanın yanında, genel kabul görmüş, ℓ_1 regülasyonlu çok değişkenli Cox nisbi risk ve Rastgele Sağkalım Ormanı modelleri ile over kanserinde sağ kalımı tahmin için kullanıldığında, tekil değerler ile kurulan modellerden daha iyi sonuçlar vermektedir. Önerdiğimiz öznitelik temsilini over kanseri hasta alt gruplarını bulmakta kullandığımızda da alt grupların sağkalım dağılımları arasında istatiksel olarak anlamlı ölçüde bir ayrım sağladığını görmekteyiz.

İkinci olarak ise, sıralama tahminini karar destek vektörleri ile öğrenen yeni bir model (RsurVM) sunuyoruz. RsurVM, sağkalım modellerinin başarımlarını ölçmek

için en yaygınca kullanılan metrik olan konkordans indisini optimize etmeye odaklanmakta ve sansürlü örneklemleri de hiç bir ön kabul yapmadan kullanabilmektedir. Over kanseri hastalarının moleküler verileri üzerinde yaptığımız geniş kapsamlı testler göstermektedir ki RsurVM, kullanılan moleküler veriden bağımsız olarak (mRNA, protein, miRNA, kopya sayısı farklılıkları ve metilasyon), net bir şekilde, Cox nisbi risk ve Rastgele Sağkalım Ormanı modellerini geri bırakarak en iyi sağkalım tahminlemesini yapmaktadır.

Anahtar sözcükler: sağkalım tahminleme, ikili sıralama, yarım sıralama, biyolojik olarak yorumlanabilir öznitelik.

Acknowledgement

Research is like walking at night around a city, which has no streetlight. As if you do not know where you are and leastwise where you are going to go and how. During this journey, there are some persons, who help you, like your supervisor who has the map of this city and gives you a torch, or your family, whose support keeps you alive emotionally or your friends whose brotherhood you feel around.

First of all, I must express my highest gratitude to my supervisor Oznur Tastan. She broadened my horizon intellectually and scientifically, I am very grateful for her supervision and very fruitful brainstormings that she triggered. Even though I sometimes run off the rails during my journey, she always motivated me and never drove me to desperation.

Secondly, I owe my deepest gratitude to my parents, Rabia and Bilal, and my sister, Kubra. Therewithal, I would not be able to go through this journey without the perseverant support of the love of my life, Esra. Without their love, support, sacrifice and encouragement, I would not find my way in this journey.

Thirdly, I would like to thank my dearest friends who have always supported me; Aykut, Burak, Muammer, Selahaddin and Serkan.

Finally, I would like to thank to Ercument Cicek and Tolga Can, who read my thesis and provided very constructive feedback. I also thank to Bilkent University faculty members, who contributed to my education and work throughout my undergraduate and graduate studies.

This thesis is a product of such a journey. I hope you enjoy reading.

With my kindest regards.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is one of the most prevalent diseases in the world. Nearly one out of every four deaths in the US occur due to cancer and it is the second biggest cause of death[1]. Cancer stems from the malignant growth of some cells, which are unable to regulate their reproduction rate. The uncontrolled multiplication of cells triggers the formation of cancer; a process called carcinogenesis. Cancer is termed as a disease of the genome because the division of cells is directly controlled by the genome. Apart from this fact the underlying phenotypic and genomic mechanisms are not completely understood yet[2].

The past decade has showcased fascinating advances in genome level characterization of a wide range of cancers thanks to the next-generation sequencing technologies. Data obtained via high throughput sequencing technologies have been used in cancer research to catalogue the genome scale alterations in the cancer cells. The next generation sequencing technologies have been used to produce a wide variety of genome-scale data while the cost of sequencing has been dropping steadily.

Nowadays the complete genome sequences pertaining to a large number of cancer types are being obtained and they are providing us with a comprehensive view of cancer development. This trend accelerates the genome scale data driven cancer research while paving the way for personalized medical treatment strategies based on genomic profiling. Although the profiling efforts are fledgling they hold a genuine potential for intensive use in the near future[3].

Various molecules such as messenger-RNA's (mRNA), micro-RNA's (miRNA), proteins collaboratively carry out the functions for a cell whereas changes like DNA-methylation (DNAmethy) aberrations and copy number variations (CNV) influence the working of these molecules. To gain a better understanding of cancer, large scale projects devoted to molecular profiling of tumors such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGA) have been initiated[4, 5]. These projects systematically review and produce different kinds of molecular profiles of tumors culled from a large cohort of cancer studies. Such large scale sequencing projects has made a large-scale trascriptomic, epigenomic (methylomic), metabolomic and proteomic available. This thesis focuses on developing computational models that utilize this rich set of problems.

Complexity of cancer has been originated from not only being driven by combinations of genes and patient specific heterogeneity but also interactions among genome scale multi dimensions[6].This large-scale multi-dimensional data have high potential the way patients are treated and also reveal molecular underpinnings of the cancer. One key use is estimating survival rate of the patients, which is critical for patients, families and clinicians. Accurate estimates allow clinicians to choose targeted treatment and surveillance strategies. Traditionally these estimates are based on clinical variables such as patient age or tumor stage[7, 8, 9]. On one hand, having only few clinical variables to consider facilitates the formulation of a survival prediction model. On the other hand incorporating molecular data enriches the predictive models. Recently, molecular information is being incorporated in these estimates. However, due to lack of multi-dimensional data these computational models are based on single molecular data type or multiple

data types but compiled from different sources[10]. Incorporating these datasets is critical, however, the fact that molecular data are inherently high dimensional and heterogeneous renders such models that rather challenging to build as well. Consequently there is a pressing need to have powerful models that are capable of handling complex and voluminous molecular data[11].

One other use of this rich set of information is to identify biomarkers that are signatures for a subtype or a survival state. Biomarkers can provide valuable insights into the biological mechanisms behind cancer development[12]. They can also be brought into use in clinics after being validated. For the latter a systems biology approach is necessary to integrate heterogeneous data and to reach conclusions that are biologically relevant and readily interpretable[13].

Toward these two goals, in this thesis we provide two contributions. In the first one, we present **Partial Ordering Features(POF)** a new way of representing molecular data suited for producing biologically interpretable features. Secondly, we propose a method-**RSurVM** a ranking based survival model, which modifies the support vector machine for ranking (RSVM)[14]. Our method is non-parametric and designed to directly capture the characteristics of censored survival.

We test our approaches on Ovarian adenocarcinoma (OV) which is one of the most difficult cancer types. OV is the first leading cause of death among gynecologic malignancies due to the lack of early symptoms and difficulties in detection and a high metastasis rate[15]. We conduct intensive experiments on mRNA, miRNA, CNV, RPPA, DNA-methy and show that RSurVM outperforms current state of approaches such as random survival forest and cox proportional hazard models.

This thesis comprises four chapters. This chapter introduces the topic of the thesis and summarizes the related works in the literature. The second chapter discusses partial ordering features. I provide also an outline of basic concepts

and methods of survival analysis and briefly introduce and explain different kinds of data used in this study during second section. The third chapter introduces RsurVM and intensive experiments made on mRNA, miRNA, CNV, RPPA and DNA-methy. The last chapter presents our conclusions and future work that can be followed.

Chapter 2

Background and Related Work

2.1 Data

In this study we use right censored survival times as the clinical observables as well as five different molecular data; CNV, RPPA, mRNA, miRNA and DNAmethy are used in this study. In the following sections we will describe these datasets.

2.1.1 Survival Time

Survival analysis deals with the time elapsed until a specific failure event takes place. The duration between the start of an observation and the failure event is termed the survival time. In this study survival time refers to the time interval between the first visit to the clinic and the date of death. Studies that resort to this methodology necessitate non-standard methods because the resulting distributions are too skewed to be handled under normality assumptions and the survival data are often censored. Censoring refers to the cases for which the failure event does not take place within the observation window. For patients in a clinical study, this

also includes instances where a patient withdraws from a study or the observer loses his/her track. Therefore, the survival information is incomplete. Survival time can be right censored start of the observation is known but the time the event occurs is unknown) or left censored (only the occurrence time of an event is known). In this study, the survival time of the patients is right censored.

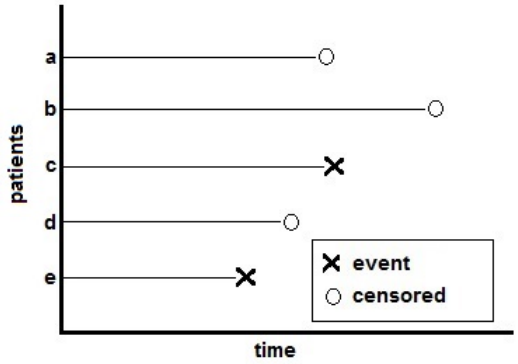

Figure 2.1: Censored survival times

Occurrence of death is not known for some patients so that some patients can be excluded from the observation before an event occurs. δ_i represents whether patient i, p_i censored $\delta_i = 1$ or not $\delta_i = 0$. Censored data is inherently incomplete because the value of a measurement (death in this case) is only partially known. Let us denote, C_i^* as censoring time and T_i^* as the actual survival time:

$$\delta_i = 0 \quad \text{if} \quad T_i^* \leq C_i^* \tag{2.1}$$

$$\delta_i = 1 \quad \text{if} \quad T_i^* > C_i^* \tag{2.2}$$

observed survival time S_i in our data is $S_i = min(T_i^*, C_i^*)$. Our survival data for N patients can be summarized as $D = \{x_i, S_i, \delta_i\}_{i=1}^N$ where x_i is explanatory variables[16, 17]. For example,p_a, p_b, p_d have censored survival time indicating $T_i^* > C_i^*$ for $i = \{a, b, d\}$. Patient a, b and d left from observation before an event occurs. See figure 2.1.

2.1.2 Molecular Data

In this subsection, I provide an overview of the description of the molecular data. Since the detailed discussions of the molecules and sequencing technologies are beyond the scope of this study, we list the sources without offering comprehensive explanations. Below we briefly overview the functions of the molecules used in this study in relevance to cell biology. [1]

- **mRNA:** Messenger RNA molecules convey genetic information from DNA to the ribosome. mRNAs are transcript from gene coding parts of the DNA and the mRNA transcripts are then translated to protein sequence. In this study, term "mRNA" indicates the expression levels of mRNA transcripts obtained from the platform Agilent 244K Custom Gene Expression G4502A[18].

- **miRNA:** A micro RNA is a typically 22-nucletotide long non-coding RNA molecule. MicroRNAs functions in RNA slicing and regulate gene expression post-transcriptionally. These are fine-tuning parameters of regulation of gene expression and their transcript abundance can signal different cell states. In this study, the term "miRNA" indicates the expression levels of microRNA transcripts obtained from the Agilent 8 15K Human miRNA-specific microarray platform[18].

- **RPPA:** Reverse phase protein array (RPPA) is a high-throughput antibody-based technique that quantifies protein expression. The protein expression levels obtained from MD Anderson Reverse Phase Protein Array (RPPA) Core platform[18].

- **CNV:** Copy number alterations indicate the structural variations in the genome such as deletions or duplications. The term "CNV" denotes which

[1] Preprocessed data used in this study have been taken from the Pan-cancer project (syn300013) on Synapse (`http://www.synapse.org`) Overall survival data and CNV data from Firehouse (`https://confluence.broadinstitute.org/display/GDAC/Home`). All clinical and genomic/proteomic data used in this study are available at the Synapse homepage of Pan-cancer project with the accession number syn1710282 (`doi:10.7303/syn1710282`).

level alterations occur for some common regions among patients. Finally the platform used herein is Affymetrix Genome-Wide Human SNP Array 6.0[18].

- **DNAmethy:** Methylation groups attach and modify the function of the DNA according to location where they bind. The data obtained from Illumina Infinium Human DNA Methylation 27K reveals to what extent a specified region(gene) is methylated[18].

2.2 Survival Estimation

Survival estimation is a challenging task because the survival distribution is skewed and non-normal and the information is missing due to censoring [17]. Dichotomized, unsupervised models and survival analysis via hazard rates have been proposed in the literature. Dichotomized survival models reduce survival modeling to a classification problem and dichotomize the censored continuous survival data through a designated cut off time (survival milestone). According to these cut off values, survival times are grouped into two: high survival, and low survival. Dichotomized survival gives an idea associated to survival (e.g. low, high) but does not produce estimates on the risk precisely. It is problematic due to being dependent too much on the dichotomization process (cut off point, inclusion, exclusion criteria). It is possible that dichotomized survival classes do not capture the data very well if dichotomization criteria do not align well with the data; however, it expedites the feature selection and variant analysis since it reduces the problem to a binary classification problem. Unsupervised methods are generally used for sub group detection. Patients are clustered into groups and survival distributions of these groups are compared with statistical tests. Most widely used test is the log-rank test[19]. It is a non-parametric test that compares two or more groups. Clustering is helpful for sub group identification but does not make a prediction for each sample separately. It is unsupervised but hard to validate.

The most realistic scenario is to analyze the survival time of patients as is. At that point, operating on censored data is the most challenging aspect of the survival modeling. Because it involves a learning task with the risk of being misled from the data in which patient survivals are censored. Parametric models such as *exponential,weibull or gompertz*[20] semi-parametric models such as *Cox proportional hazard rates*[21] and finally non-parametric models such as *random survival forests*[22] methods have been proposed to address this problem.

2.2.1 Kaplan-Meier Estimator and Plot

The KaplanMeier (KM) estimator is a non-parametric statistic used to estimate the survival function. KM plot divides the survival function into steps for each censored points. Next, for each step, it produces the maximum likelihood estimates of survival function $S(t)$ non-parametrical by accounting for the proportion of the patients who survives at each step[23]:

$$for \quad t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_n, \quad \hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}, \tag{2.3}$$

where d_i is the number of deaths before t_i and n_i is the number of patients at risk at t_i.

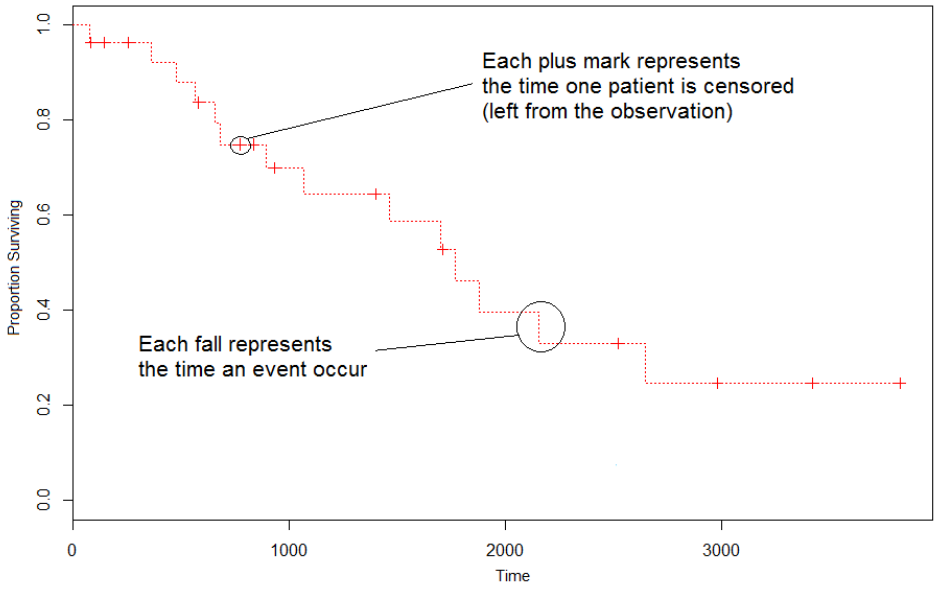

Figure 2.2: Sample Kaplan-Meier plot

2.2.2 Log-rank Test

Survivals of two or more groups of patients can be compared with a log-rank test. The log-rank test[19] is the most widely used method to compare survival curves. It is a non-parametric test. For each group at each event time, log-rank test calculates the expected number of events since the previous event. These values are then summed over all event times. This summation gives the total expected number of events in each group (E_i for group i). The log-rank test compares the observed number of events(O_i for group i) to the expected number of events (E_i) via the test statistic:

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \tag{2.4}$$

X^2 can be compared according to test statistic with a χ^2 distribution $g-1$ degrees of freedom, where g is the number of groups. In this manner, p-value can be calculated to indicate the statistical significance of the differences between the survival distributions of groups.

2.2.3 Hazard Functions and Estimated Hazard Rates

Hazard rates are estimated from the hazards ratios of deaths via a hazard function conventionally denoted as λ. Hazard function $\lambda(t)$ measures the instantaneous rate of failure[16]. Hazard function estimates the failure or death rate at time t conditional on survival until time t or later ($T \geq t$). Let $S(t)$ be the survival function that characterizes the distribution of survival times and p be the density function of T. $S(t) = Pr[T > t]$ for $t > 0$ is the probability that an individual is still alive at time t. $\lambda(t) = \lim_{\triangle t \to 0} Pr[t < T \leq t + \triangle t | T > t]/\triangle t$ is the hazard function and $\lambda(t) = p(t)/S(t)$.

2.2.4 Concordance Index- Harrell's c-index

Concordance index (c-index) is the modification of Kendall-Goodman-Kruskal-Somers type rank correlation index[24, 25]. It measures in which level estimated hazards ($\hat{h_i}$) are concordant with the actual survivals (s_i, δ_i) for comparable pairs(ε) among patients(p_i). The bigger the estimated hazard is, the more likely an early occurrence of an event is .

$$\varepsilon = \{\forall(p_i, p_j) \quad | s_i < s_j \quad and \quad \delta_i = 0\} \tag{2.5}$$

A comparable pair means valid comparison between survivals of two patients. As it was mentioned in the censored data section, some of survival times are incomplete (censored, $\delta_i = 1$). The comparison of two patients would not be valid for the pairs in which survival time pertaining to a patient whose data is censored, is less than the other patient being compared. On the other hand, all the pairs that includes a patient whose survival data is not censored and lower than the patients he/she is being compared to, would be valid. Comparable pairs constitute the set all of complete pairwise ranking of survivals. Concordance index (c-index) refers to concordance between survivals that are referred by the estimated hazards of comparable pairs.

$$c = \frac{1}{|\varepsilon|} \sum_{\forall(i,j)\in\varepsilon, s_i<s_j} \mathbf{1}_{\hat{h_i}>\hat{h_j}} \tag{2.6}$$

$$\mathbf{1}_{\hat{h_i}>\hat{h_j}} = \begin{cases} 1 & if \ \hat{h_i} > \hat{h_j} \\ 0 & otherwise \end{cases} \tag{2.7}$$

For example, let us consider p_b, p_c, p_d and p_e in figure 2.1. If we describe valid comparison of a pair (p_i, p_j) as $S_i < S_j$. Comparable pairs in this set will be $\epsilon = \{(p_e, p_c), (p_e, p_d), (p_e, p_b), (p_c, p_b)\}$ but $\{(p_d, p_b), (p_d, p_c)\} \notin \epsilon$.2

2.2.5 ℓ_1 Regularized Cox Proportional Hazards

Cox Proportional Hazards Model: Hazard function $\lambda(t)$ measures the instantaneous rate of failure as described in the previous sections. $\Lambda(t) = \int_0^t \lambda(u)du$ represents the cumulative hazard function and Cox porportional hazard model(Cox-ph) assumes that $S(t) = e^{-\Lambda(t)}$[26].

Proportional Hazard (PH) model has been considered the standard method to analyze the effects of explanatory variables for censored survival data[27]. PH models are based on the assumption that effects of the covariates to risk (hazard function) are multiplicative, which can be formulated as follows:

$$\lambda(t|x) = \lambda_0(t)e^{\mathrm{w}^T x} \tag{2.8}$$

Here $e^{\mathrm{w}^T x}$ is the proportional hazard term in which the effects of covariates x are multiplicative with the regression parameters $\vec{\mathrm{w}}$ and $\lambda_0(t)$ is the baseline hazard function (i.e. when $x = 0$). In this model, hazard function of a patient $\lambda(t|x)$ is proportionally dependent to x with coefficients $\vec{\mathrm{w}}$ on the baseline hazard $\lambda_0(t)$. Cumulative hazard function with baseline cumulative hazard $\Lambda_0(t)$ is:

$$\Lambda(t|x) = \Lambda_0(t)e^{\mathrm{w}^T x}, \quad \text{and} \quad S(t|x) = e^{-\Lambda_0(t)e^{\mathrm{w}^T x}} \Rightarrow S(t|x) = e^{-[e^{\mathrm{w}^T x}\int \lambda_0(t)dt]} \tag{2.9}$$

If the proportional hazard is assumed to hold, estimating the effect of the parameters without any consideration of the baseline hazard function is possible[28]. Cox proportional hazard model is a semi-parametric model that baseline hazard function is completely unspecified and only the effects of covariate parameters are estimated. Cox proposed to maximize *the partial likelihood*[21] function $L(\hat{\mathrm{w}})$ in order to estimate the $\hat{\mathrm{w}}$:

$$L(\hat{\mathrm{w}}) = \prod_{T_i uncensored} \frac{e^{\mathrm{w}^T x_i}}{\sum_{T_i \leq T_j} e^{\mathrm{w}^T x_j}} \tag{2.10}$$

ℓ_1 penalized log partial likelihood: Lasso[29] is a regularization approach to estimate the regression coefficients constrained by the ℓ_1 norm. ℓ_1 norm forces some

of the coefficients to be 0 and enable model to make automatic feature selection. The optimization function is:

$$\hat{\mathrm{w}} = arg\,max\, L(\mathrm{w}), \quad \text{s.t.} \quad \|\mathrm{w}\|_1 \leq s, \tag{2.11}$$

where $L(\mathrm{w})$ is partial log likelihood of Cox ph model and $\vec{\mathrm{w}}$, coefficients of the model in equation 2.10.

2.2.6 Random Survival Forest

Random Survival Forest(RSF)[22] is the extension of Random Forest(RF)[30] to right censored survival data. RF is an ensemble model, where decision trees are bagged. Introducing diversity in the ensemble models in general provides better learning mechanism. In RF, variation is introduced in two forms. First, each tree grows from randomly drawn bootstrap samples of the data. Second, at each node of the tree, a randomly selected subset of variables (covariates) is used for splitting. Averaging over trees, in combination with the randomization used in growing a tree, enables RF to approximate rich classes of functions while maintaining low generalization error[22]. Most importantly, RF can capture the non-linear effects of covariates and it is a non-parametric method that requires no parameters to be estimated.

RSF adopts RF methodology to censored data by changing splitting criteria and uses hazard estimation at the leaves. Each node in RSF is split as maximizing survival distributions of daughter nodes. Trees grow until no more than d_0 unique deaths remain in a terminal node. For each terminal node u, cumulative hazard function is estimated via a Nelson-Aalen estimator:

$$\hat{\Lambda}_u(t) = \sum_{t_{j,u} \leq t} \frac{d_{j,u}}{Y_{j,u}} \tag{2.12}$$

$$\Lambda(t|x_i) = \hat{\Lambda}_u(t), \quad x_i \in u \tag{2.13}$$

where x_i is the d-dimensional covariates of case i, $d_{j,u}$ is the number of deaths at time t and $Y_{j,u}$ represents the number of individuals at risk at time $t_{j,u}$ for

each terminal node u. If we consider that there are B bootstrapping steps to grow different random trees in RSF, the boostrap ensemble cumulative hazards for patient i, $\Lambda_e^*(t|x_i)$ would be:

$$\Lambda_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B}\Lambda_b^*(t|x_i) \tag{2.14}$$

where $\Lambda_b^*(t|x_i)$ is the cumulative hazard function for a tree grown from b^{th} bootstrap sample.

2.3 Related Work

Most commonly used model for survival prediction is the Cox Proportional Hazard Model(Cox-ph)[21]. It is a semi parametric model and standard for survial analysis. Cox-ph models are based on the assumption that effects of the covariates to risk (hazard function) are multiplicative. If the proportional hazard is assumed to hold, estimating the effect of the parameters without any consideration of the baseline hazard function is possible[28]. Other parametric models also have been proposed such as, Weibull or Gompertz[20], which assume survival distributions follows a predefined distributions. These model estimates the survival of patients based on these distributions.

Conversely, Random Survival Forest[22] non-parametrically estimates the survivals. It is motivated from Random Forest[30], which is an ensemble model consisting of many decision trees. In Random Forests (RF), randomization is applied in both sample space and feature space so that RF maintains low generalization errors as well as capturing variety of functions. Random Survival Forest(RSF) adopts Random Forest methodology to censored data by changing splitting criteria and adds the hazard estimation. Each node in RSF is split as maximizing survival distributions of daughter nodes and hazard of patients are estimated for each terminal nodes.

Besides these widely accepted models, SVM based model[31], which modifies regression optimization term in regular SVM as considering censored cases, is proposed as well as artificial neural network based models[32], which simply omits censored cases or uses estimated survival rates of censored cases. Bair and Tibshirani[33] proposed semi-supervised method based on clustering to predict patient survivals. In this model, patients are clustered and survival of patients are estimated differently for each cluster using Expectation-Maximization (EM) algorithm. Semi-supervised clustering, which iteratively selects features based on log-rank test that is calculated on clusters obtained at each iteration [34]. Ritchie

et al[35] proposed a model based on genetic algorithm implemented on neural networks. This model offers an integrative framework designed to predict survivals while integrating meta dimensional omics data through identifying interactions among them.

Steck et al[16] show that classical survival analysis involving censored data can be cast as a ranking problem. They showed that log sigmoid and exponential lower bound on concordance index, which is standart performance measure in the survival prediction and quantifies pairwise ranking quality. Log sigmoid bound emerges directly from the description of Cox-ph model based on Cox partial likelihood. Empirical experiments shows that these two bounds and Cox-ph perform equally well. They also explain the relation between Cox-ph and concordance index that refers that survival analysis can be cast to a ranking problem.

In addition to the survival prediction models, dichotomized survival models are used[11, 36, 37, 38, 39, 40]. These models cast the survival estimation to classification task via predefined survival milestones. Survial times are transformed to discreate survial classes such as low survial and high survival with these milestones and classification problem is solved with such classes. This approach suggests a general framework that makes any classification model possible to be applied for survival modelling. However, this approach solves an easier problem and is dependent to dichotomization criteria.

Clustering is another methodology that leverages survial times to capture biological insights. In this approach, patients are grouped first with a clustering method then clusters are compared in terms of survial time disributions. Various methods have been proposed for clustering samples to find subtypes and validate survival separation. Hierarchical clustering(HC) has been successfully utilized to analyze expression patterns among patients[41] such as predicting outcome of lymphoma patients[42], or providing molecular patterns found in breast cancer[43]. Self-organizing maps (SOM) was also used to find molecular patterns[44] such as recognizing subtypes of leukemia[40]. SOM and HC, because of being unstable

and imposing stringent structure respectively, have lost their popularities, instead consensus clustering[45] has been proposed as an newer approach followed by NMF consensus[46], which is very powerful and robust.

Yuan et al[47] assessed the utility of clinical or molecular data according to survival estimation. Similarly, TCGA data have been compared with alternatively collaborative influence of different molecular data to survival[48, 49]. Some frameworks proposed to predict survival of patients with the integration of different molecular data[50, 13, 35, 51].

Chapter 3

Partial Order Features

Better representation of the molecular data has potential to increase the predictive performance of survival models. In the context of survival prediction, models are built not only to predict the outcome or survival but also to identify predictive biomarkers that can be used clinically or to shed light molecular alterations that lead to the phenotype.

So far the efforts to identify novel prognostic factors have focused on molecular markers via high throughput molecular profiling of large tumor sets. In most of these approaches, models rely on the individual expression quantities of the molecules in the tumors. However, cancer is a complex disease where molecular mechanisms are dysregulated in various ways. In this chapter, we propose partial order features (POF) methodology based on a system level perspective. We incorporate interactions within different molecules (mRNA, protein, miRNA). We use partial ordering of the molecules in lieu of individual expression values. Our strategy unveils predictive features with direct relevance to the biological mechanism of cancer and it efficiently accounts for order dsyregulations among the individual features.

3.1 Partial Ordering

Let $X = \{x_1, x_2, \ldots, x_n\}$ is N dimensional molecular data and $Y = \{S, \delta\}$ is the survival data, where S is survival time and δ denotes censoring. The problem is finding model $f\{\ldots\}$ so that $f(X) \sim Y$.

We propose a new way of feature representation: pof, ϕ:

$$f(\phi(X)) = f(\tilde{X}) \sim Y \quad where \tag{3.1}$$

$$\phi(X) = \tilde{X} = \{\tilde{x}_{12}, \tilde{x}_{13} \ldots \tilde{x}_{23}, \tilde{x}_{24}, \ldots \tilde{x}_{(n-1)(n)}\} \tag{3.2}$$

$$\tilde{x}_{ij} = \begin{cases} 1 & if\ x_i > x_j \\ -1 & otherwise \end{cases} \tag{3.3}$$

$\tilde{x}_{ij} = 1$ indicates that the molecule i is more abundant with respect to molecule j whereas $\tilde{x}_{ij} = -1$ indicates that the molecule i is less abundant with respect to molecule j. We compare the performance of partial order features with two well established methods: ℓ_1 regularized(Lasso) Cox Proportional Hazard Model[52] and Random Survival Forest[22].

3.2 Pipeline for Continuous Survival

Overall pipeline, similar to pipeline in Yuan at al[47], followed in continuous survival models consists of five phases (see figure 3.1). In the first phase (figure 3.1-A), data are randomly split to training and test sets. Partial order features (pof), are created in the second phase as a new form of feature representation (figure 3.1-B). Feature preselection with univariate Cox screen takes place in the third phase (figure 3.1-C), followed by performance tests of newly created features with molecular data on two survival models, Cox-ph and RSF (figure 3.1-D). The fifth phase is the last step wherein the proposed methods are evaluated statistically with c-index (figure 3.1-E).

3.2.1 Bootstrapping

The data are split into test and training sets with 100 bootstrapping steps. Training sets comprise 300 patients and the test sets comprise 75 patients selected out of 379 patients. Phases from **B** to **E** are implemented on training sets to come up with selected features and estimated parameters for multivariate Cox model and RSF. C-indices are calculated on test sets.

3.2.2 POF Representation

POF offers a discrete binary representation for continuous molecular data that is biologically interpretable. In this methodology the features are produced from all possible pair combinations of variables at hand. Therefore, the number of features will grow quadratically [please note that $c(n, 2) = n(n-1)/2$], a trend too expensive to tackle computationally for most cases. For example size of POF for mRNA, miRNA and mRNA are $13530, 318003$ and 158642578 respectively. To handle feature explosion, we put a threshold (t) for the number of variables used

to produce POF. We determined t with respect to the size of RPPA, which has the minimum number of features. At this point, it is required to choose t features from N. In order to select t features from N variables, all the variables are ranked with respect to median absolute deviation (MAD) and the top t of them are selected. A ranking based on MAD is expected to reveal the features whose influence or collaborative effect to survival is possibly manifested more strongly. Partial order features are then created based on these t variables.

3.2.3 Univariate Cox-Screening

Most of the time, putting all the features to survival prediction model is inefficient in terms of runtime and space requirements both for POF and raw molecular data. One plausible way is to conduct a pre-selection step. Towards this aim, we eliminate some of the features according to the p-value of partial likelihood ratio test, which is calculated on the univariate Cox model for each feature set separately[47]. We exclude the features that are deemed statistically insignificant by conducting a partial likelihood test. Statistical significance criterion is decided by setting the threshold p-value to 0.05. p-values are calculated by Wald's test[53]. Features with p values smaller than 0.5 are excluded. Partial likelihood ratio test checks whether the hazard ratio is less than 0.05 to determine the association of each feature to survival based on univariate Cox-ph. Therein, hazard ratios greater than 1 indicate a weak association.

3.2.4 Creating Survival Models-RSF and Cox-ph

Following univariate Cox screening, we obtain a manageably large set of features relevant to survival. For each bootstrapping training and test pair, we run ell_1 regularized Cox-ph and Random Survival Forest. In conclusion, we attain hazard ratios, $\hat{w}_i$ in Cox-ph and a forest model $\hat{\Upsilon}_i$ in RSF for each bootstrap i, B_i.

3.2.5 Comparing Performances

Performance of POF with respect to individual expression data is compared with C-indices. The concordance index decides quality in survival predictions by evaluating estimated hazards of the patients. For each bootstrapping sample, we calculate C-indices for POF and individual expression data on Cox-ph and RSF. Ultimately, we obtain c_i^{rsf} and c_i^{cox} for each B_i with $\hat{w}_i$ and $\hat{\Upsilon}_i$ accordingly. We use Wilcoxon signed-rank test to be able to see whether C_{pof}^{cox} and C_{pof}^{rsf} show an improvement comparing with C_{raw}^{cox} and C_{raw}^{rsf} where $C_{data}^{method} = \{c_i^{method} \quad \forall \quad B_i^{data}, i \in [1, 100]\}$.

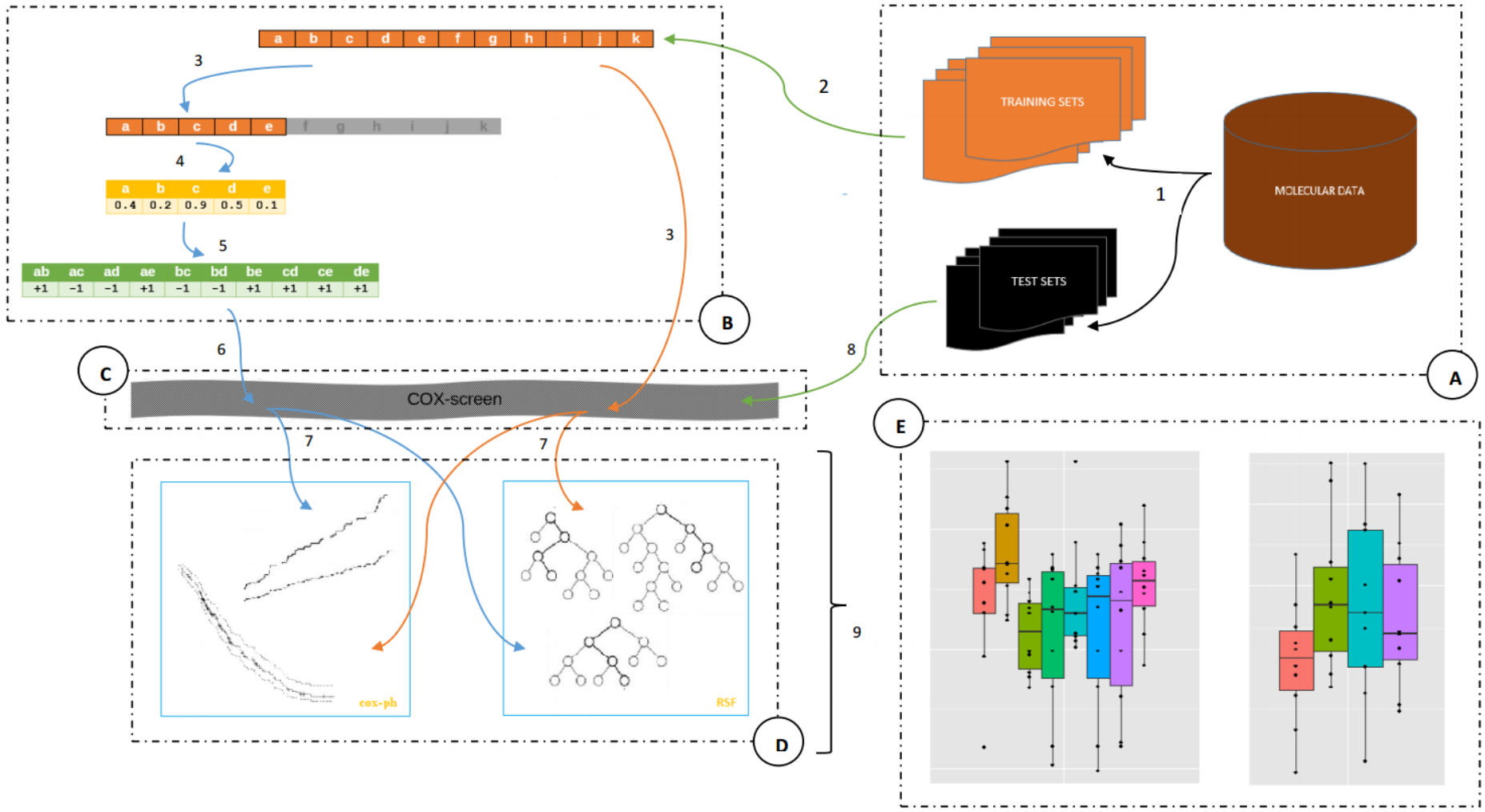

Figure 3.1: Overall pipeline. Data are randomly split 100 times (1) with bootstrapping. POF are created (5) after MAD filtering (2-3-4). Feature pre-selection is done with univariate Cox for POF (6) and raw data (3). Models are created with RSF and Cox-ph (7). Performances are compared with c-idex (8-9).

3.3 POF Results on Cox-PH and RSF

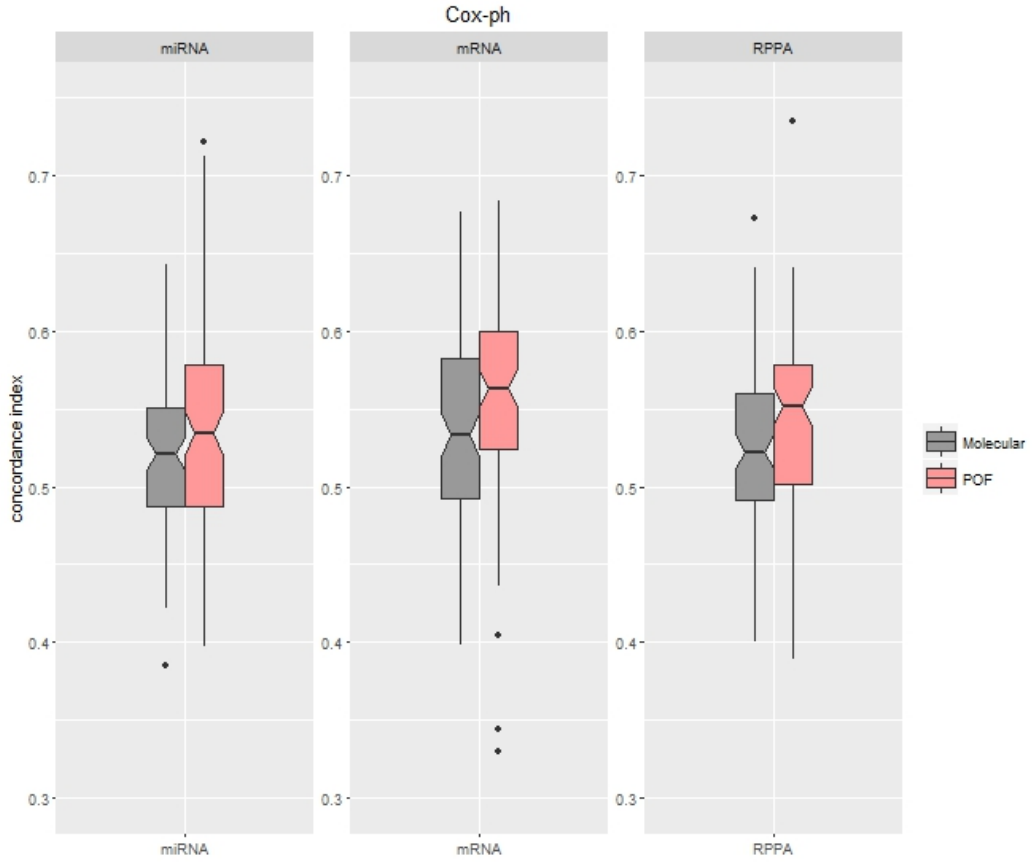

Figure 3.2: Concordance indices of molecular data and POF with Cox Proportional Hazards Model

Partial ordering features represent comparative or relative expression levels of two molecules and they offer biological insight. For this reason, we use mRNA, miRNA and RPPA in our experiments and exclude CNV and DNAmethy since their comparative effects would not further clarify the underlying mechanisms. For each bootstrapping training set, we obtain C-index for Cox-ph model and RSF model. We test our approach on each bootstrapping test set and obtain C-indices accordingly. Although, the level of improvement differs for each molecular data, partial

ordering of all data yield statistically significantly better prognosis capability according to Wilcoxin-signed rank test (see table 3.1 and 3.2) .

Table 3.1: Comparison between molecular data and POF with Cox Proportional Hazards Model

Cox-ph							
	p-value	c-index					
		1^{st}	2^{nd}	median	3^{rd}	4^{th}	std
mRNA	**0.005**	.40	.49	.53	.58	.68	.06
mRNA-pof		.33	.53	**.57**	.60	.76	.07
miRNA	**0.017**	.39	.49	.52	.55	.64	.05
miRNA-pof		.40	.59	**.53**	.58	.72	.07
RPPA	**0.002**	.40	.49	.52	.56	.67	.05
RPPA-pof		.39	.50	**.55**	.58	.73	.05

$mRNA$ yields slightly better results and $mRNA_{pof}$ is the most predictive data. For l1-regularized Cox-ph model(see figure 3.2), improvement obtained by $mRNA$ surpasses other data types. Additionally, it yields more robust estimation of hazards as can be seen on the standard deviations (ref table). $RPPA_{pof}$ shows better prognosis capability than $RPPA$ and reaches the level of C-index=0.55. Though overall concordance indices of $miRNA_{pof}$ is better than $miRNA$, improvement is very slight. Lower quartile of $miRNA_{pof}$ is almost same with $miRNA$. This along with standard deviations indicate that $miRNA_{pof}$ is more sample specific.

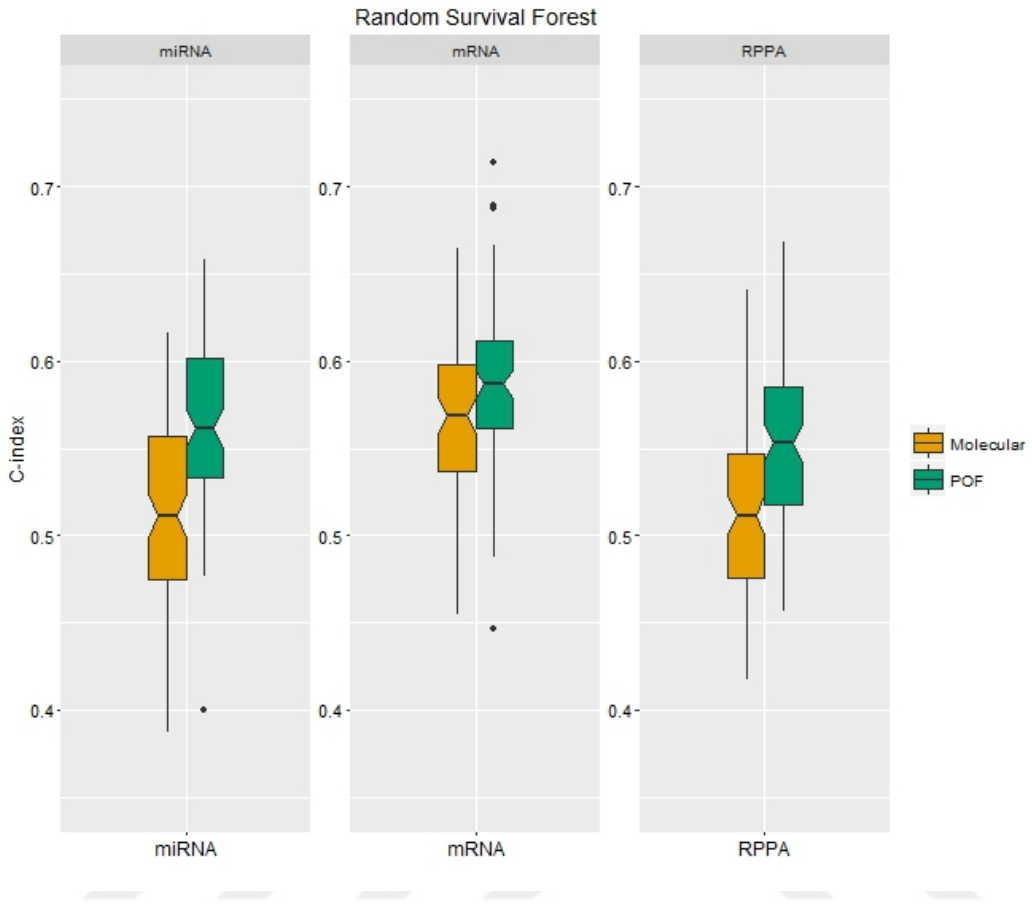

Figure 3.3: Concordance indices of molecular data and POF with Random Survival Forest

The concordance indices obtained through Random Survival Forest together with POF are statistically more significant (see table 3.2). POF in $miRNA$ and $RPPA$ improves more considerable than $mRNA$ with lower quartiles. For all cases, RSF improves better than Cox-ph indicating that RSF benefits from the partial order representation very well. All quartiles for each POF data are higher than corresponding molecular data quartiles hinting that RSF with POF is more robust than the alternatives. For all data types, figure 3.3 clearly shows that the non-linear interaction among pair of molecules are more contingent in the scope of survival prediction.

Table 3.2: Comparion between molecular data and POF with Random Survival Forest

RSF							
	p-value	c-index					
		1st	2nd	median	3rd	4th	std
mRNA	**0.001**	.45	.54	.57	.60	.66	.05
mRNA-pof		.45	.56	**.59**	.61	.71	.04
miRNA	**0.000**	.39	.47	.51	.56	.62	.05
miRNA-pof		.40	.53	**.56**	.60	.66	.05
RPPA	**0.000**	.42	.48	.51	.55	.64	.05
RPPA-pof		.46	.52	**.55**	.58	.67	.05

Consequently, partial ordering offers a promising way of feature representation as it can infer by three points. Firstly, RSF captures a variety of functions and non-linear interactions. Besides, Cox-ph is the standard for survival modeling especially for those considering linear relations. The finding that POF yields better survival estimation with 100 bootstrap samples over two widely used methods in the literature, most likely originates from its direct biological relevance. Secondly, some molecular data like miRNA, which are less competent in estimating survivals, works as good as mRNA, which gives the best survival estimation, in the POF case. Thirdly, the differences between improvements obtained by POF on two models hints at the presence of nonlinear interactions among the molecules.

3.3.1 Obtained POF

To be able to arrive testable biological hypothesis it is important to understand which features have predictive power. As demonstrated in the previous sections, POF representation outperforms individual molecular representation of features when used as input to both Cox-Ph model and RSF model. Among the two, RSF performance is better (see figure 3.3). Therefore, we analyzed the feature importance in RSF models.

In RSF, variable importance of a feature is quantified as the difference between the performances of the original ensemble and the ensemble where this features assignments are randomized. Thus, a variable importance measures the increase or decrease in misclassification error solely due to that feature[54]. Large importance values implies the feature contributes to the model.

Variable importance of features are calculated by `randomForestSRC` package in R[55]. The average importance of each feature over 100 bootstrapping samples is calculated and ranked by the average variable importance. This are shown in figures 3.4,3.6,3.8. In this figures the most predictive features are highlighted. For each molecular data type, we also focused on the top performing pof (see figures 3.5,3.7,3.9).

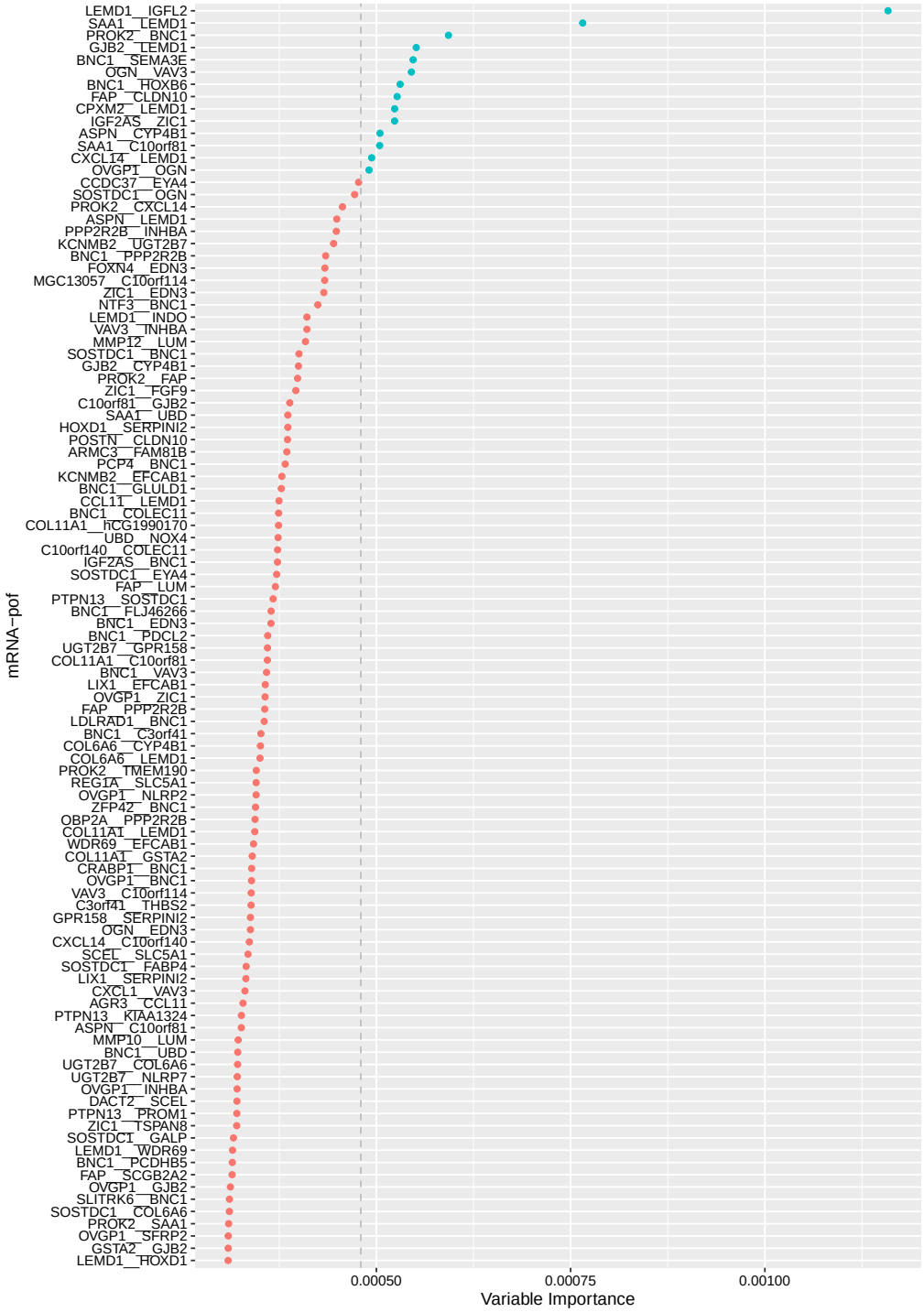

Figure 3.4: Variable importance of significant mRNA-pof, blues are top features among the significant features

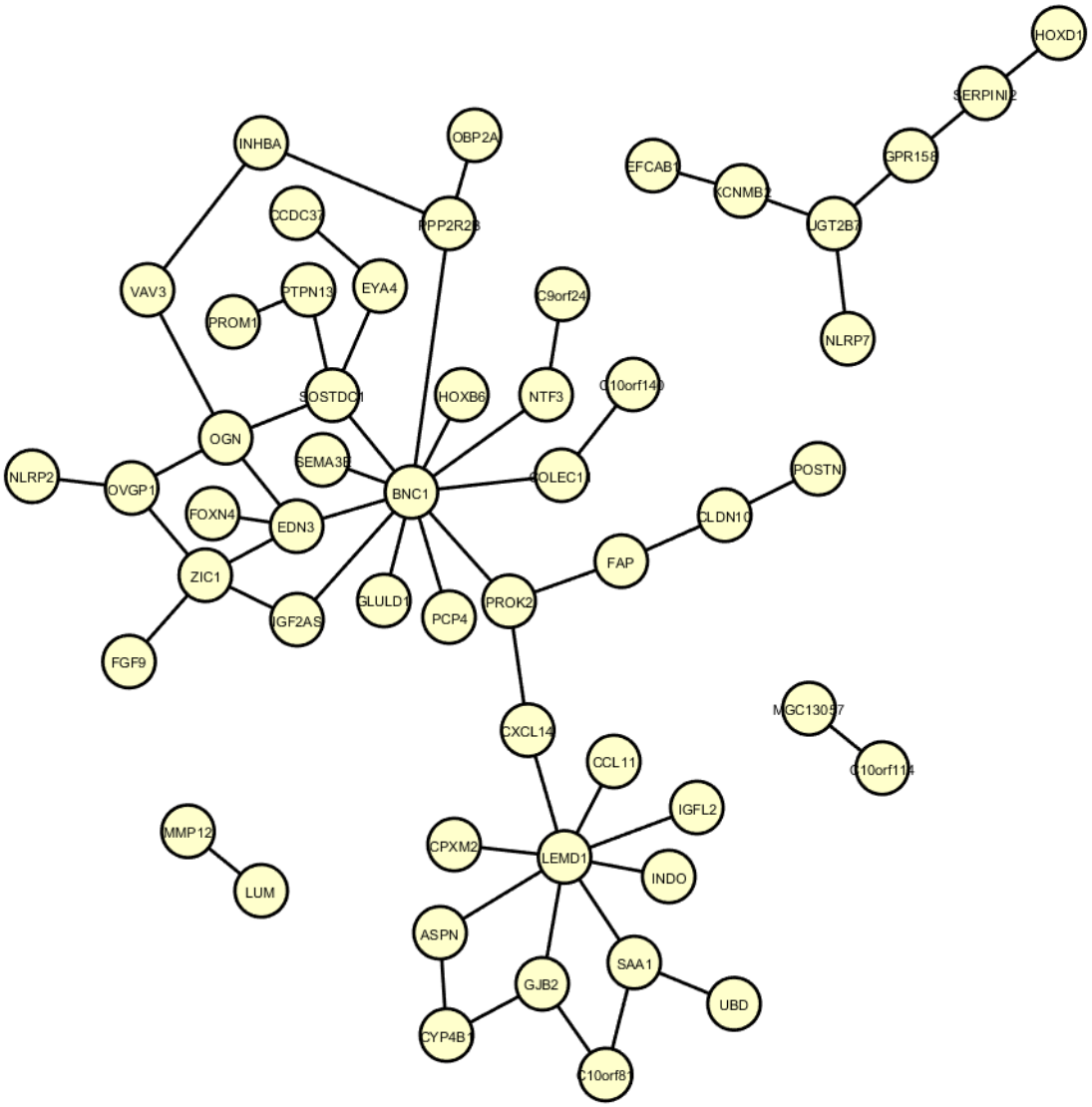

Figure 3.5: Graph of significant mRNA-pof, edges represent POF while nodes are mRNAs

Figure 3.6: Variable importance of significant RPPA-pof, blues are top features among the significant features

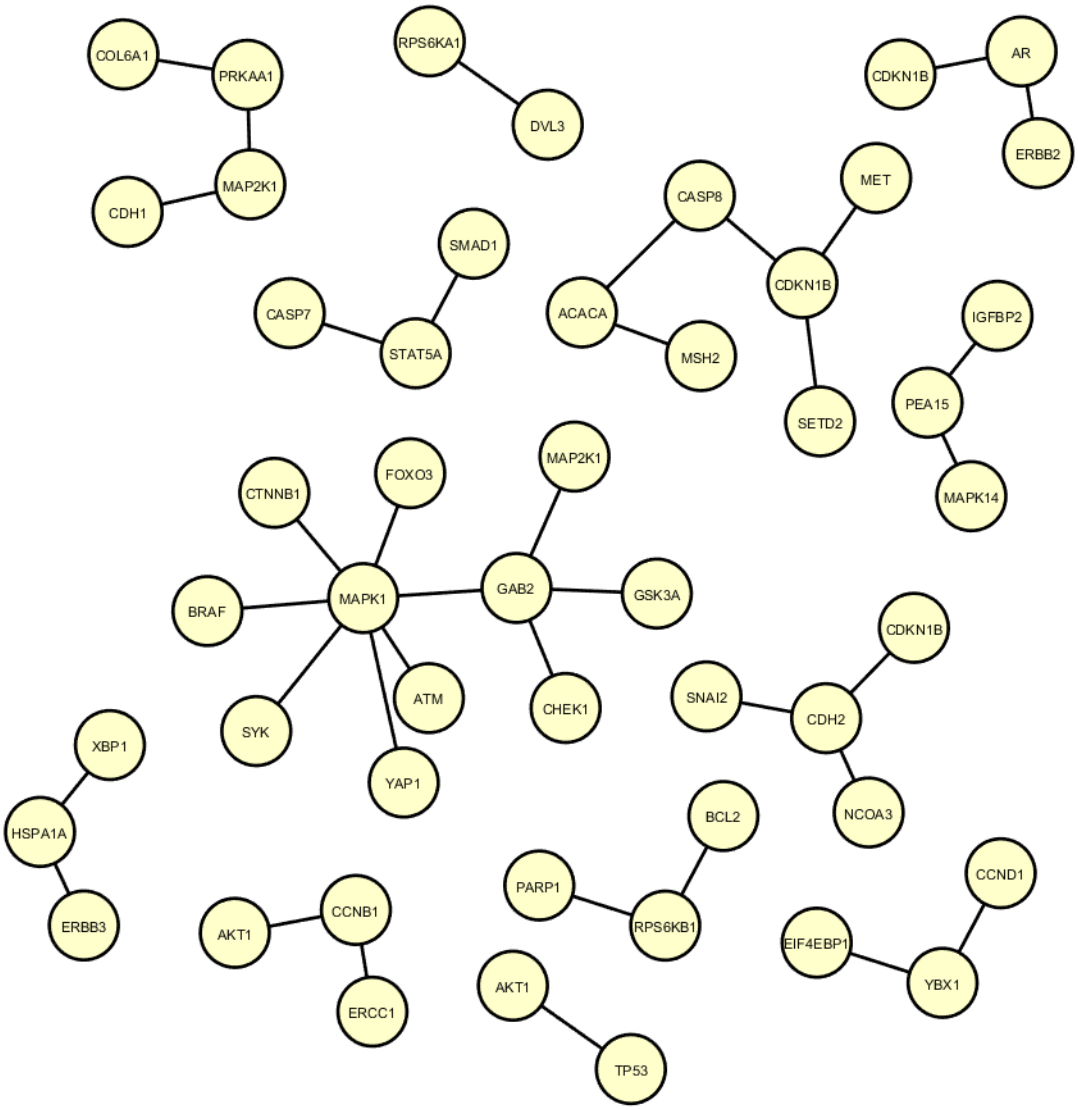

Figure 3.7: Graph of significant RPPA-pof, edges represent POF while nodes are RPPAs

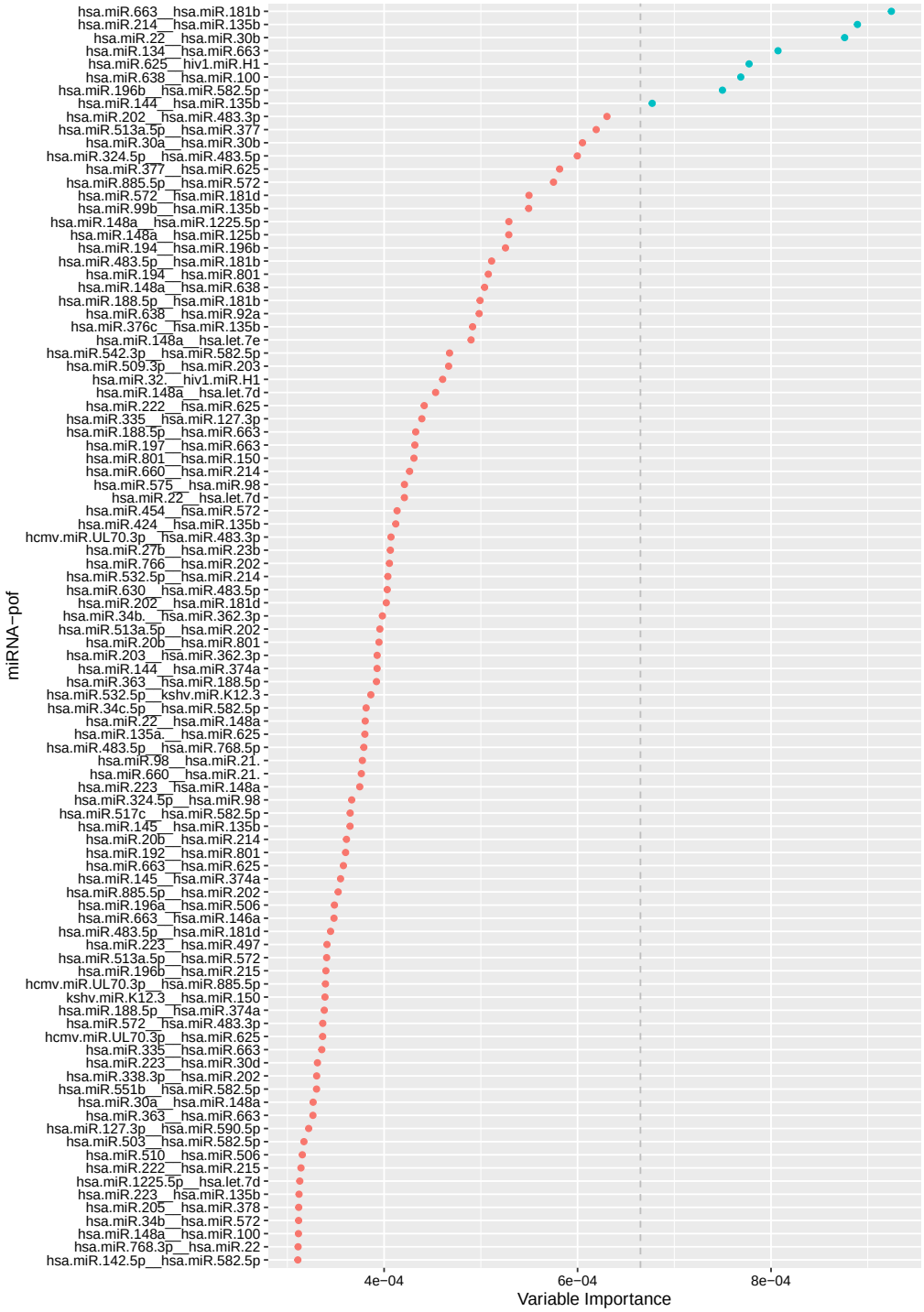

Figure 3.8: Variable importance of significant miRNA-pof, blues are top features among the significant features

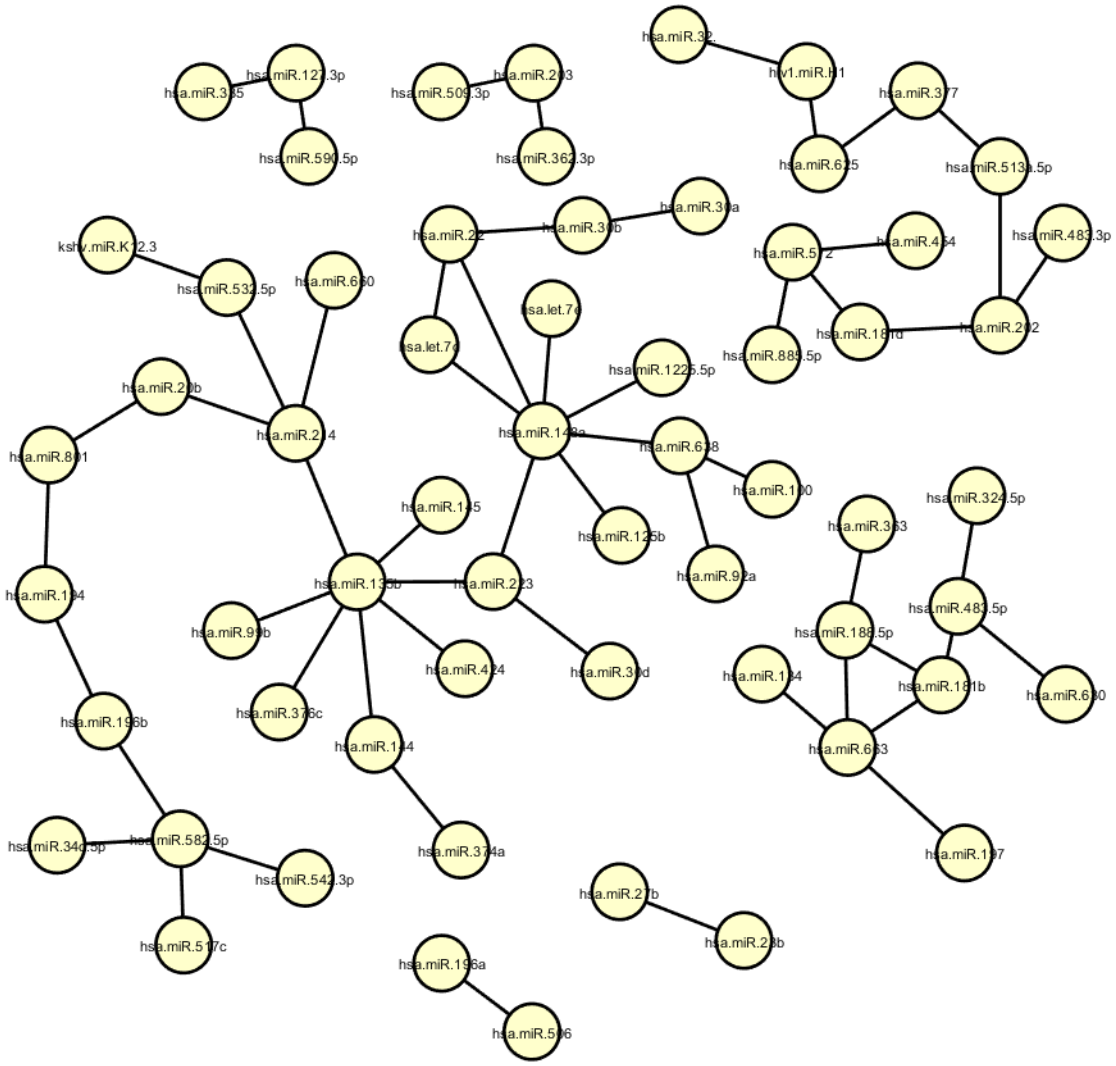

Figure 3.9: Graph of significant miRNA-pof, edges represent POF while nodes are miRNAs

3.3.2 Top Features

In this section, the top features extracted from POF are discussed in light of the available information culled from an extensive literature search. We observe that most of the top features have already been associated with ovarian cancer or other cancers.

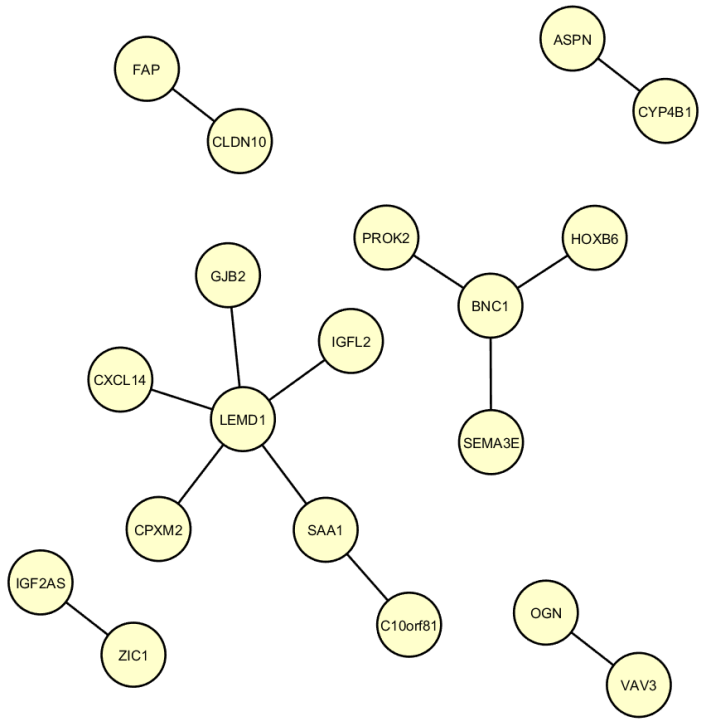

Figure 3.10: Graph of top mRNA-pof, edges represent POF while nodes are mRNA's

As a notable LEMD1 gene comes up as an important feature in the models where mRNA-pof is used. Interestingly, it has been reported to be one of the cancer testis antigens and its overexpression was reported in colorectal, prostate cancer and lymohoma. LEMD1 has also been labelled as a cancer biomarker and a promising target for immunotherapy[56, 57]. Other studies show that VAV3 is

overexpressed in cancer stem cells and found to be a poor prognostic indicator in ovarian cancer(OV) patients[58]. ZIC1 and OGN have been proposed to play a role in the progression of OV through its role in oxygen transport and embryonic development and offered to serve as a therapeutic target for OV[59]. Significantly increased CLDN10 expression in cancerous ovaries in comparison to normal ovaries have been reported and CLDN10 is proposed as a novel biomarker of OV[60]. FAP encodes the fibroblast activation protein, which has been associated to regulation of tumor-associated fibroblasts and epithelial cancer cells. Targeting FAP has been proposed as a potential therapeutic strategy [61]. SEMA3E has been highly expressed in human high-grade ovarian endometrioid carcinoma and found to be related to tumor progression[62]. PROK2 is thought to influence cancer growth by regulating the host defense and angiogenesis[63].

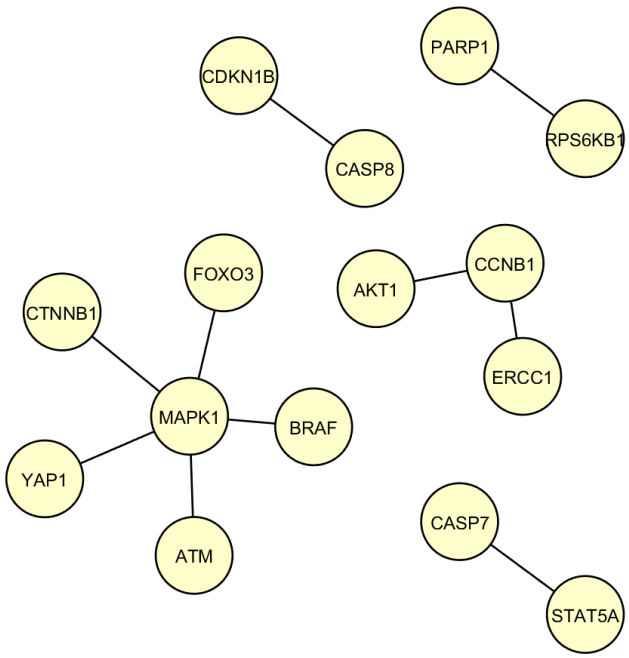

Figure 3.11: Graph of top RPPA-pof, edges represent POF while nodes are mRNA's

Among some of the top features in RPPA-pof models were Mitogen activated protein kinase 1(MAPK1), also known as extracellular signal regulated kinase-2 (ERK2), is involved in many cellular processes including cell proliferation, differentiation, programmed cell death, angiogenesis and migration[64]. Besides its normal physiological functions, this protein is also important for oncogenic transformation and progression in several cancers[65, 66]. Amplification in MAPK1 gene has been shown to shorten progression free survival in high grade OV[67]. MAPK1 phosphorylates FOXO3, a transcription factor which is responsible for the down regulation of FOXO3 in a proteosomal degradation dependent manner and tumorigenesis[68, 69]. BRAF, an upstream kinase in MAPK pathway is mutated in two thirds of melanoma patients, and less frequently for patients diagnosed with serous ovarian adenocarcinoma, colon cancer, papillary thyroid cancer[70]. MAPK and Wnt/B catenin signaling pathways interact with each other positively and negatively in different types of cancer cells[71]. YAP is important for establishing the stemness in OV initiating cells and high expression of it adversely affects survival and chemotherapy response[72]. ATM was also reported to be associated with poor prognosis since expression level of ATM correlates to adverse clinical outcomes in OV [73]. Cyclin B1 makes complex with CDK1 (cyclin dependent kinase 1) and controls G2-M checkpoint in cell cycle. Although high expression of Cyclin B1 has been found in low grade ovarian tumors, increase in its level is related to poor prognosis and advanced malignancy in breast, non-small cell lung, renal cancer and others[74, 75]. High Akt activation is associated with poor clinical outcome and resistance to treatment in cancer[76]. Elevated Cyclin B1 protein decreases response to cisplatin treatment in esophageal cancer cells through the induction of PTEN/Akt pathway[77]. DNA excision repair protein ERCC-1 status is a predictor of chemotherapy response among OV patients[78]. P70 S6 kinase regulates actin microfilaments and migration which is one of the steps forerunning metastasis so it could be a potent target to control the spread of disease[79, 80]. PARP-1 inhibitor had been approved for high grade serous ovarian adenocarcinoma with BRCA mutation which rendered progression free survival possible[81]. Caspase 8 and p27 are promising molecules to estimate the length of OV survival[82, 83].

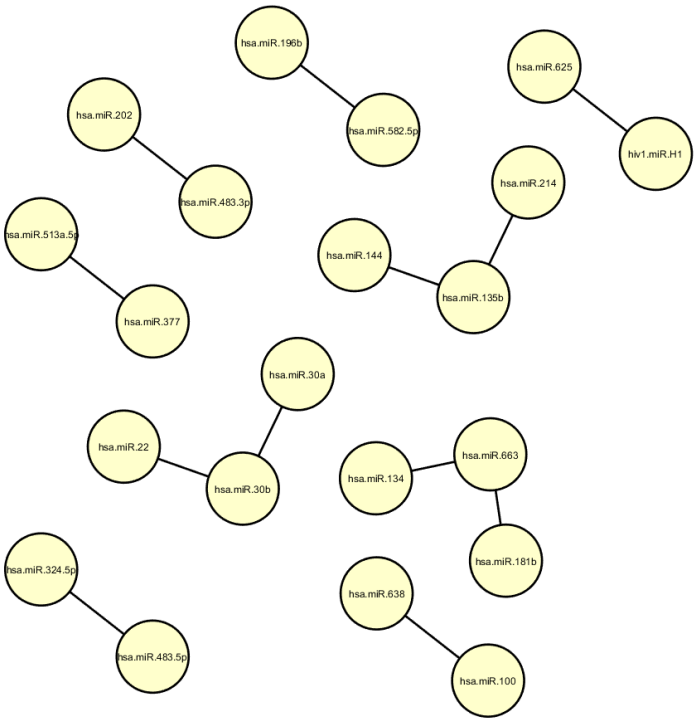

Figure 3.12: Graph of top miRNA-pof, edges represent POF while nodes are miRNA's

Among the top miRNA-pof, miR-135b is involved in the progression of breast, colon, lung and prostate cancers[84, 85, 86]. miR-135b-3p expression is significantly higher in OV tissue than in the normal ones[87]. miR-144 disrupts the balance between glycolysis and oxidative phosphorylation in OV cells and induces rapid cellular growth[88]. miR-214 increases OV cell survival and resistance to platinum based therapies via activating the Akt pathway[89]. While miR-663 is overexpressed in taxol resistant OV cells, overexpression of miR-134 is related to chemo-sensitivity[90, 91]. miR-22 functions as a tumor suppressor; inducing apoptosis while inhibiting proliferation, migration and invasion in OV[92]. miR-22 and miR-100 have significant prognostic value in OV[93, 94]. Strong expression of miR-30a is associated with ovarian clear cell adenocarcinoma[95].

3.4 Clustering with POF

In this section, POF and molecular data are compared in an unsupervised manner in terms of whether POF can perform well in producing subgroups of patients with different survival distributions. Clustering is a widely used methodology to spot gene patterns[96, 41, 44, 97] and identify sub groups[98, 99, 100] or survival groups[101, 102, 103, 104]. Clustering is another way to assess the quality of feature representation. However, clustering is harder to validate.

We use NMF consensus[46] clustering to assess whether POF result in any gene patterns. To validate the availability of the patterns on POF, we consider consensus matrix heat map[45] with cophenetic correlation(CPCC)[105]. After clustering, whether produced patterns are meaningful or not should be validated carefully. We compare patterns produced by POF and molecular data with log-rank test and seek for features that produce the best divergence in survival distribution of groups.

3.4.1 NMF Consensus Clustering

Non-negative matrix factorization decomposes a matrix into two matrices with positive entries. Various well tested algorithms exist in the literature that carry out such a decomposition[106]. NMF finds uses in machine learning in different contexts. For instance Lee and Seung exploit NMF[107] in order to learn parts (decompose) of objects. In their formulation, NMF produces a meaningful factorizations from which facial features such as eyes, and the nose obtained. Text data can also be decomposed to capture semantic polysemy.

Brunet at al.[46] adapts this idea to gene pattern discovery. They propose to use NMF for decomposing molecular data into two matrices in order to describe high dimensional molecular data in terms of a few metagenes: $\mathbf{A} \sim \mathbf{WH}$ where $\mathbf{A} = \mathbf{N} \times \mathbf{M}$, $\mathbf{W} = \mathbf{N} \times \ \mathbf{k}$ and $\mathbf{H{=}k} \times \mathbf{M}$. W maps N genes to k metagenes,

where w_{ij} is the coefficient of gene i in the linear combination for metagene j. Similarly, H maps k metagenes to M samples. After factorization, H groups M samples into k clusters just considering most highly expressed gene for each sample to place one of k clusters; that is, given that h_{ij} is the largest entry in the j'th column, sample j will be clustered to cluster i. As discussed in the paper[46], NMF consensus provides robust predictions while retaining their advantage in becoming less dependent to priori information such as feature selection.

We cluster samples into $k = 2, 3, 4$ clusters with POF and molecular data and compare the quality of clustering and survival patterns.

3.4.2 NMF Consensus with POF and Results

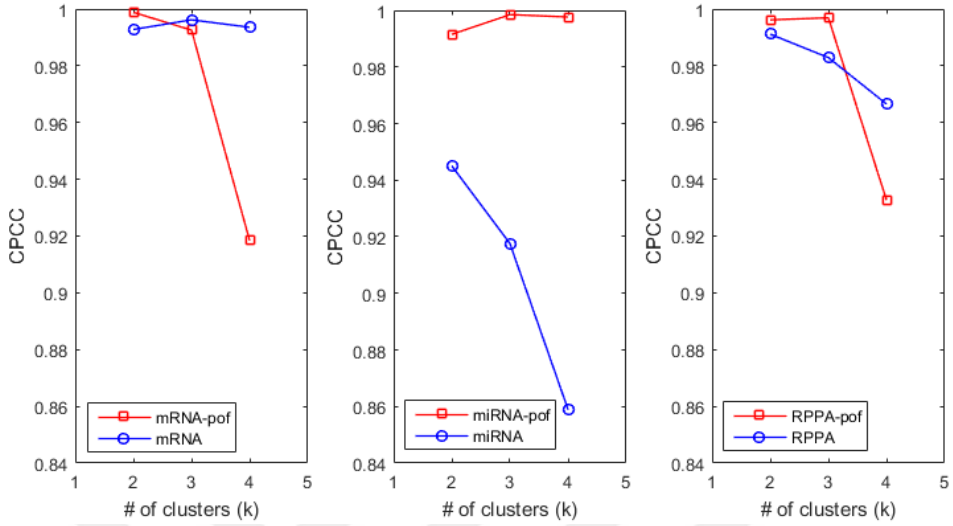

Figure 3.13: Cophenetic correlation coefficients of molecular data and POF for clustering (k=2,3,4).

Partial ordering is more successful in capturing the underlying molecular patterns. For all molecular data, most obvious patterns are clustered with POF. Figure 3.13 shows that partial ordered features produces the better clustering with mRNA, miRNA and RPPA. It can be also validated by the heat maps (see figures 3.14, 3.16 and 3.18.) Additionally, POF suggests a different number of clusters for each molecular data indicating that POF can find molecular patterns which are not detectable solely by molecular data. The most successful patterns related to survival are revealed by POF in mRNA and RPPA while miRNA-pof is slightly below the significance level of miRNA.

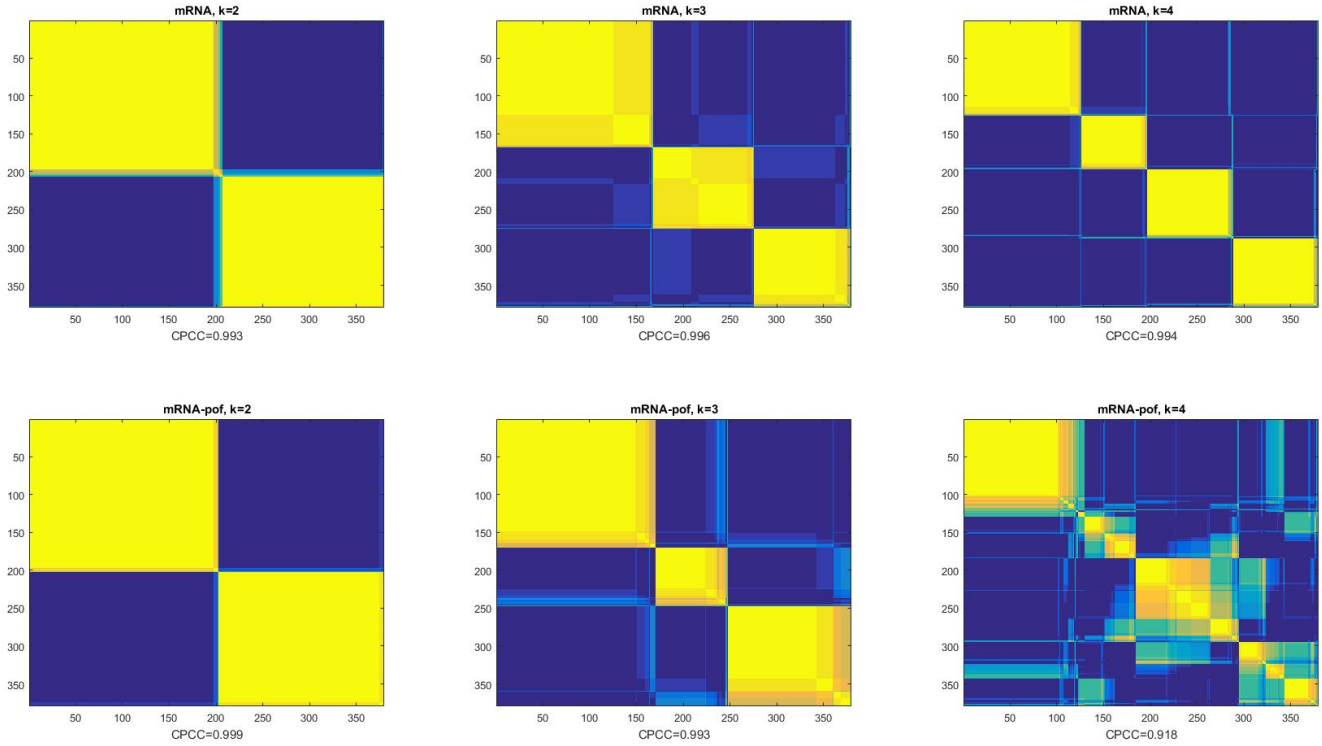

Figure 3.14: Consensus matrix heatmap of mRNA and mRNA-pof for clustering (k=2,3,4), colored from 0 (deep blue, samples are never in the same cluster) to 1 (dark yellow, samples are always in the same cluster).

Table 3.3: mRNA cophenetic correlation coefficients(CPCC) and p-values for k=2,3,4.

	mRNA			
	Molecular		POF	
	CPCC	*p-value*	CPCC	*p-value*
k=2	.993	.260	.999	.031
k=3	.996	.260	.993	.098
k=4	.994	.102	.918	.079

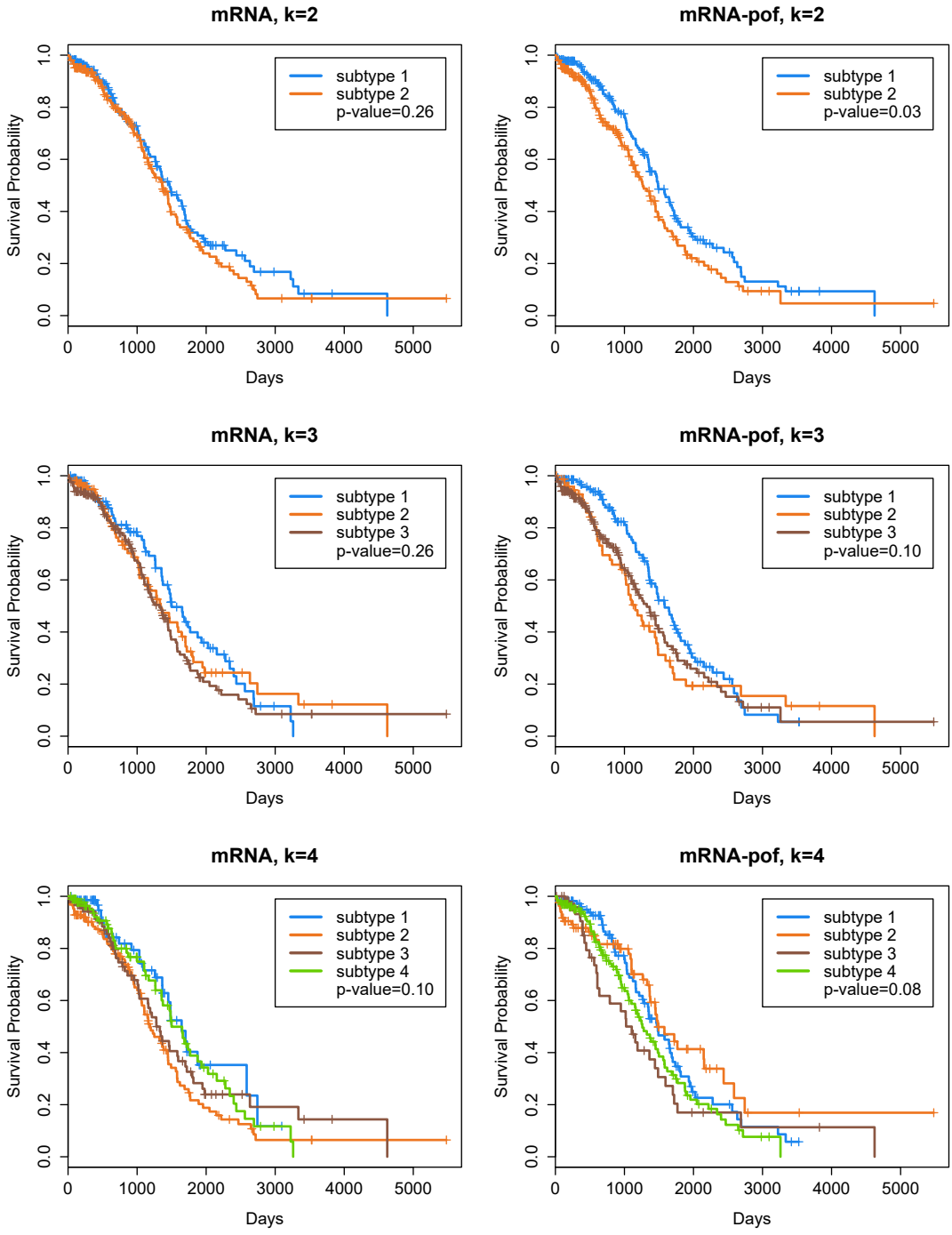

Figure 3.15: Kaplan Meier plots of clusters (k=2,3,4) for mRNA and mRNA-pof.

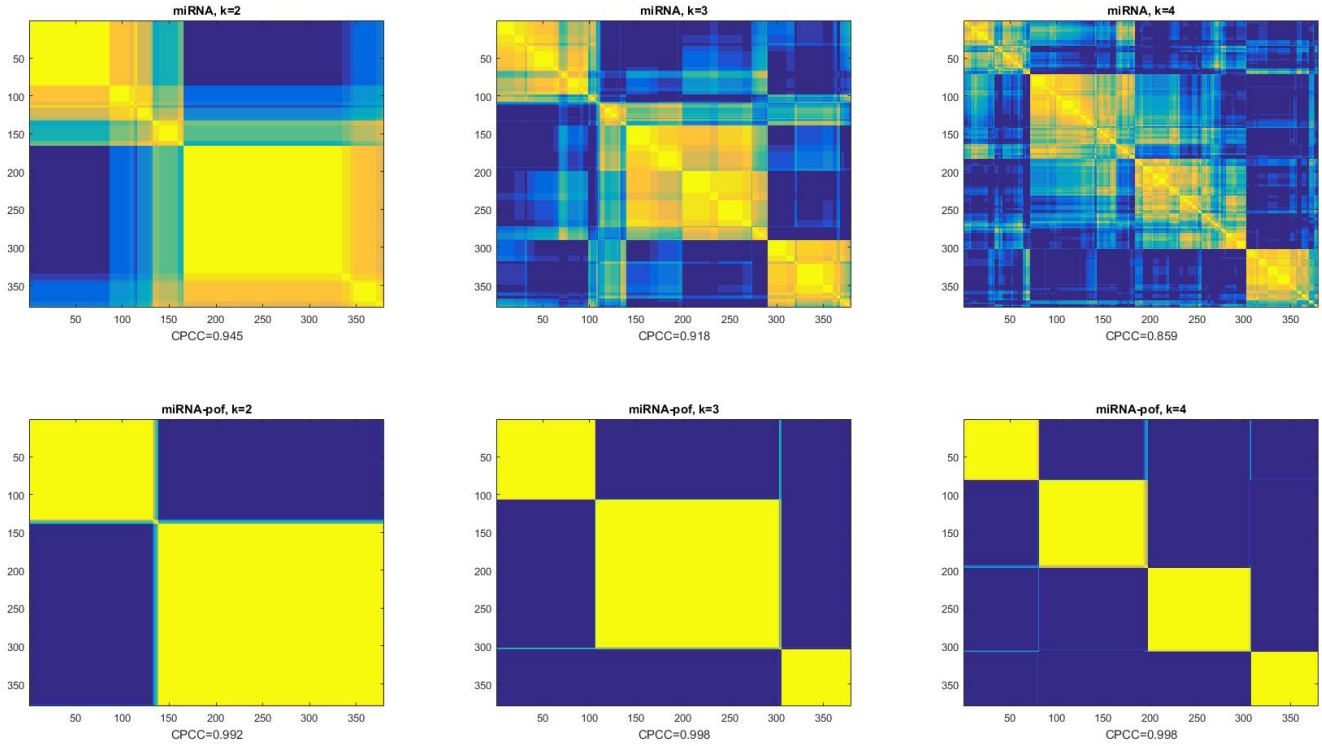

Figure 3.16: Consensus matrix heatmap of miRNA and miRNA-pof for clustering (k=2,3,4), colored from 0 (deep blue, samples are never in the same cluster) to 1 (dark yellow, samples are always in the same cluster).

Table 3.4: miRNA cophenetic correlation coefficients(CPCC) and p-values for k=2,3,4.

	miRNA			
	Molecular		POF	
	CPCC	*p-value*	*CPCC*	*p-value*
k=2	.945	.048	.992	.250
k=3	.917	.073	.998	.074
k=4	.859	.689	.998	.412

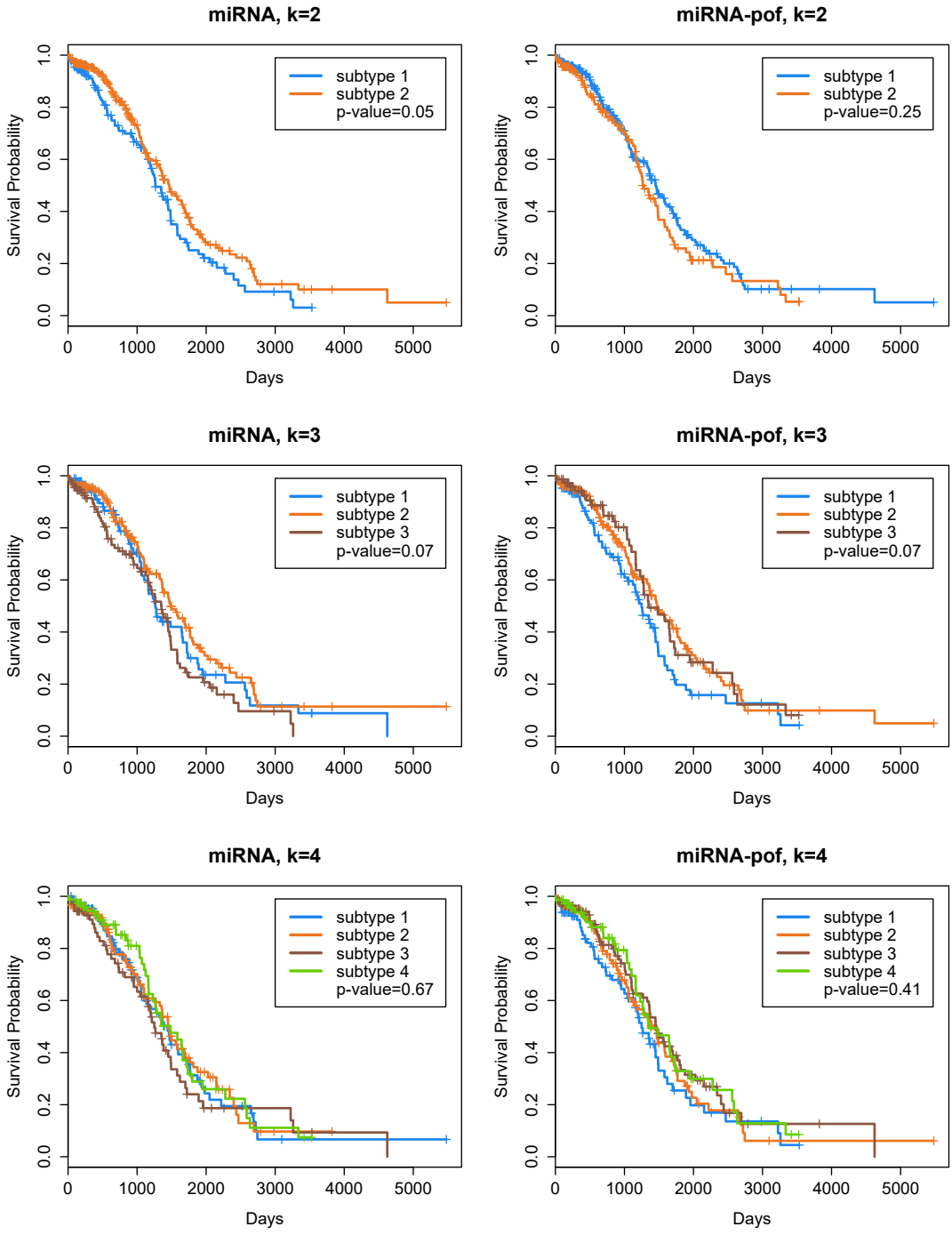

Figure 3.17: Kaplan Meier plots of clusters (k=2,3,4) for miRNA and miRNA-pof.

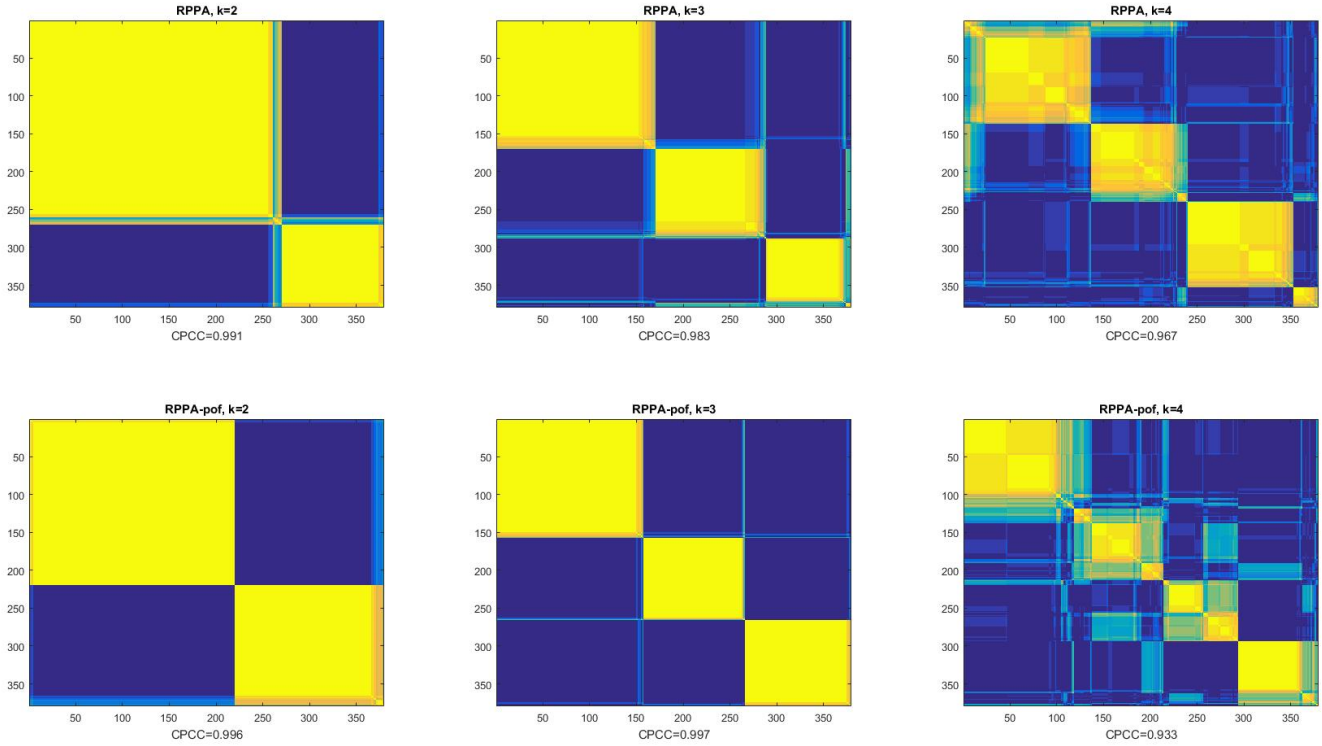

Figure 3.18: Consensus matrix heatmap of RPPA and RPPA-pof for clustering (k=2,3,4), colored from 0 (deep blue, samples are never in the same cluster) to 1 (dark yellow, samples are always in the same cluster).

Table 3.5: RPPA cophenetic correlation coefficients(CPCC) and p-values for k=2,3,4.

	RPPA			
	Molecular		POF	
	CPCC	*p-value*	*CPCC*	*p-value*
k=2	.991	.401	.996	.818
k=3	.983	.005	.997	.021
k=4	.967	.011	.933	.001

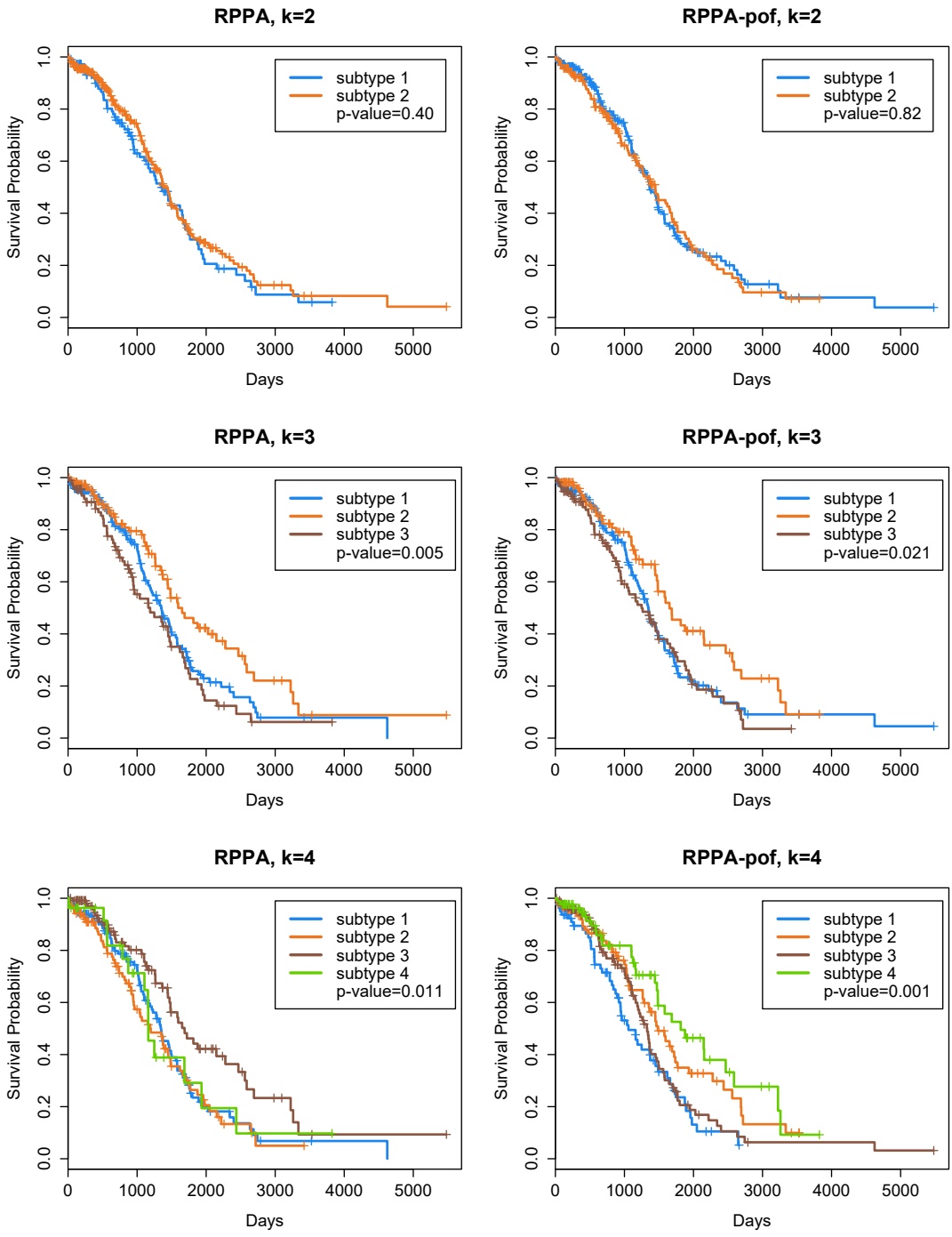

Figure 3.19: Kaplan Meier plots of clusters (k=2,3,4) for RPPA and RPPA-pof.

mRNA-pof suggests two clusters (CPCC=0.999, almost perfect clustering) and mRNA suggests three clusters (see figure 3.14 and table 3.3). While none of the patterns resulting from molecular data shows significant differences in survival, mRNA-pof produces a clear divergence on the survival as a result of $k = 2$ clustering (see 3.15). With the log-rank test mRNA yields $p = 0.26$ for $k = 3$, which is the best clustering attempt by mRNA. In contrast, mRNA-pof yields $p = 0.031$, which points out a clear difference in survival distributions according to log-rank test. Moreover, the mRNA-pof for all k. mRNA does not produce any significant divergence among the survival distributions. On the other hand, mRNA-pof succeeds by either producing a very significant ($p < 0.5$) divergence or by approximating the level of significance ($0.05 < p < 0.1$).

miRNA-pof offers better clustering than miRNA for all k values as it can be deduced from figure 3.16. miRNA ascertains two distinct molecular patterns and achieves clear divergence on the survival distributions of these patterns (see figure 3.17 and table 3.4). Also it approximates the significance level with $k = 3$. miRNA-pof suggests three patterns whose survival distributions are approximately significant with $p = 0.074$.

RPPA-pof clusters the samples better just like the cases with other data types for which using POF produces the best results (see figure 3.18). Additionally RPPA-pof separates the survival groups better. Firstly, the best clustering attempt suggested by RPPA($k = 2$) does not attain diverging samples with respect to survivals. It gives $p = 0.40$ in the case of $k = 2$. In contrast, RPPA-pof groups samples into statistically significantly three groups, which shows that RPPA-pof results in the best clustering (see table 3.5). RPPA yields significant difference in the survivals of groups for k=3,4. However, best divergence in the survivals is achieved by RPPA-pof with $k = 4$ resulting in $p = 0.001$(see figure 3.19).

POF reveals clearer patterns while maintaining better survival distribution divergence. For most of the cases, the quality of clustering matches the survival

difference achieved on grouped samples, when POF is resorted to with the exception of miRNA. miRNA is found to be the less convenient molecular data to build models for survival. This result coincides with the results obtained by Cox-ph and RSF. RPPA-pof stands out as the best candidate to cluster samples into different survival groups. RPPA-pof gives the best results while clustering samples to four different groups, which is the hardest case for log-rank test. Obtaining significant p-value as a result of log-rank test is harder if more groups are being compared. In fact, RPPA-pof offers more specific grouping and results in better divergence on the survival distributions. More importantly, mRNA cannot find patterns that significantly differ in terms of survival, if POF are not utilized.

Indeed, this analysis does not present a fair comparison between POF and the molecular data. As a result, some of the features are eliminated by MAD filtering to curb the number of features. This may indicate the fact that POF are able to reveal better molecular patterns than molecular data with less number of features. For example, mRNA-pof is created with 165 genes whilst mRNA has 17813 genes. This conclusion coincides with the results obtained by RPPA-pof wherein all RPPA features are utilized. Thus, POF offers a better way of feature representation than individual representation.

Chapter 4

A Ranking Approach to Survival Prediction: RSurVM

A common approach in survival prediction is modeling the survival rates of the patients and then evaluating the performance of the methods using concordance index. As explained in chapter 2, concordance index basically is dependent on correct partial ordering of the two patients. Thereby, optimizing this metric can be achieved through learning a survival model that predicts the partial order of the patients directly. In this chapter, we present a method that takes a pairwise ranking approach for survival prediction that can handle censored data appropriately.

4.1 Survival Prediction upon Pairwise Ranking

Ranking is a different task than classification and regression. In learning to rank the goal is to find a function that would rank the items correctly based on a score. Owing to its wide applicability learning to rank is studied a lot in the information retrieval domain. The most common application is to rank order documents based on their relevance to an input query. There are three main approaches in to learning to rank: pointwise[108], listwise[109] and pairwise[14] . In the pointwise approach, a score that ranks the examples is learned directly. The listwise methods consider all participating examples to get a correct ordering of all examples and they train on a loss function based on all examples score. Pairwise approach focuses on each pairs of examples and aim to get the correct ordering of the pairs.

In survival prediction, the common evaluation metric is concordance index. Concordance index by definition is related to pairwise ordering of the patients. Here we adapt Support Vector Machine ranking methodology (RSVM)[14]. The Ranking Survival Vector Machines (RsurVM) is presented as an adaptation of RSVM to survival prediction.

One difficulty in survival prediction is that data can be censored. Common survival prediction methods handle the censored data differently. Cox proportional hazard model assumes proportionality in the hazard rates if proportionality does not hold, Cox will not be applicable. Random Survival Forest incorporates the censored data in making split decision by using statistical tests such as long-rank test than can work with censored data, these estimates the survival of a patient according to survival distribution of the node that pertains to that patient. If ratio of censored patient is too high, Random Survival Forest would not work robustly.

Here our model will learn as much as possible from the censored cases without making any assumptions and without using the statistics of censored cases is proposed. First, a general framework of pairwise learning to rank task from the information retrieval literature is described, the RSVM before making a transition

to survival prediction context.

Let us describe $\langle q_i, q_j \rangle$ as a pair of queries in training set, (Q_{tr}, Y_{tr}) where Q_{tr} represents the feature vectors of queries and Y_{tr} stores the rankings or relevance scores. Let us denote π_{tr} as all pairwise ranking of queries in the training set. $PR(...)$ is a ranking model that trains on π_{tr}, and gives the ranking score, $\hat{r}_k$, for the query k such that $\hat{r}_k = PR(x_k|\pi_{tr}, Q_{tr}, Y_{tr})$. In this formulation, π is called preferences or rank orders and created within the ranking scores of queries. π denotes preferences as $\pi = \{\langle q_i, q_j \rangle = (-1, 1) : y_i < y_j, \forall i, j\}$ indicating that I is preferred to J.

4.1.1 Ranking Support Vector Machines (RSVM)

Let I is preferred to J be denoted as $I > J$, main task is to learn from all preferences or rankings in the training set to predict ranks of test queries. RSVM [14] adapts support vector machines (SVM) formulation for binary classification [110] to ranking task.

For a given training set (x_i, y_i) with $x_i \in \Re^N$ and $y_i \in \{-1, +1\}$ in regular SVM, binary classification problem with classification rule $\hat{y} = sign(\mathbf{w}^T x + b)$ tries to solve the following optimization problem:

$$\min_{\mathbf{w}, \xi_\mathbf{i} \geq 0} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i \tag{4.1}$$

where ξ is slack variables and C is penalization term. This optimization problem is well studied in the literature and it can be solved efficiently using quadratic programming[111]. RSVM is a pairwise method whose ranking methodology is mainly based on pairwise comparisons. Assume that $X_{tr} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ denotes the training set for ranking SVM where

$x_i \in \Re^N$ and y_i is the ranking of i'th query. The goal of RSVM is to learn function F from R satisfying $F(x_i) > F(x_j)$ for all pairs of $\{(x_i, x_j) : y_i < y_j \in R\}$,

$$F(x_i) > F(x_j) \Rightarrow \mathbf{w}x_i > \mathbf{w}x_j \tag{4.2}$$

$$\Rightarrow \mathbf{w}(x_i - x_j) > 0 \tag{4.3}$$

$$\Rightarrow F(x_i - x_j) > 0 \tag{4.4}$$

RSVM is being trained by pairwise ordering in X_{tr}, set of all constraints (pairwise rankings) is π_{tr}. If m is the size of X_{tr}, size of π_{tr} would be the number of all pairwise combinations: $\frac{m(m-1)}{2}$. For a given preference set π_{tr}, RSVM is built by the minimizing following objective function:

$$\min_{\mathbf{w}, l \geq 0} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(i,j) \in P} l(\mathbf{w}^T x_i - \mathbf{w}^T x_j) \tag{4.5}$$

where l is a suitable loss function such as hinge loss: $l(t) = max(0, 1-t)^2$. In order to minimize this objective function efficiently, many methods using dual form[112] or primal form with Newton optimization[113] have been proposed.

4.1.2 Ranking Survival Vector Machines (RsurVM)

We adapt the previous formulation to predict partial ranking order of the patients. Ranking Survival Vector Machines (RsurVM) is a method that estimates the survivals of patients based on the principle of RSVM. Pairwise rankings in the formulation above are considered as comparison between survival times of two patients.

Let A denote a patient. If the patient survival time is censored, $\delta_a = 1$, we will denote it with p_{a^1}, otherwise p_{a^0}. For censored patients, only observed (censored) survival time, S_a is known whereas the actual survival time, T_a is unknown,

$S_a < T_a$. All patients in the dataset,D, can be partitioned into four groups with respect to p_{a^1} as D_{-a^1}, D_{-a^0} and D_{+a^1}, D_{+a^0} representing that $p_{i^\delta} \in D_{-a^\delta}$ if $S_i < S_a$, similarly $p_{i^\delta} \in D_{+a^\delta}$ if $S_i > S_a$ where $\delta = 0, 1$. Please note that the information inferred from pairwise comparison between p_{a^1} and p_{i^δ}, $\langle p_{i^\delta}, p_{a^1} \rangle$, is certain if and only if $p_{i^\delta} \in D_{-a^0}$. In this regard, valid set of all pairwise survival comparisons in the cohort,π^S, would consist of all pairwise comparison among uncensored cases plus all certain comparisons between censored and uncensored cases. After determining valid survival comparison set,π^S:

$$\pi^S = \{\{\langle p_{j^0}, p_{i^1} \rangle \, and \, S_j < S_i, \forall i, j\} \cup \{\langle p_{j^0}, p_{k^0} \rangle \, \forall i, k\}\} \tag{4.6}$$

RSVM can train on a molecular dataset $D_{tr} = (X_{tr}, Y_{tr})$ and can return the ranking scores of patients, R_{ts}, as an estimation of survivals of each patient in the test set, X_{ts}:

$$RsurVM(X_{ts} \,|\, \pi^S_{tr}, X_{tr}, Y_{tr}) \simeq RSVM(X_{ts} \,|\, \pi^S_{tr}, X_{tr}, Y_{tr}) = R_{ts} \tag{4.7}$$

In this formulation, these scores indicate how likely a patient will survive based on the pairwise comparisons of all patients in the training set.

$$\lambda_{ts} = RsurVM(X_{ts} \,|\, \pi^S_{tr}, X_{tr}, Y_{tr}) = \frac{1}{R_{ts}} \tag{4.8}$$

where λ_{ts} is hazards of patients, which are inferred from the survival times.

Let $S(t)$ be the survival function that characterizes the probability that an individual is still alive at time t, $S(t) = Pr[T > t]$ for $t > 0$ and p be the density function of T. $\lambda(t) = \lim_{\triangle t \to 0} Pr[t < T \leq t + \triangle t | T > t] / \triangle t$ is the hazard function:

$$\lambda(t) = p(t)/S(t) \tag{4.9}$$

RsurVM utilizes the pairwise comparisons of survival times. Let's assume $\langle p_1, p_2 \rangle$ is a comparable pair, where $t_1 \geq t_2$ and $S(t = t_1) < S(t = t_2)$. Note that $p(t = t_1) \geq p(t = t_2)$ and $S(t = t_i) \simeq R_i$ where R_i is ranking score of p_i

obtained by RsurVM. Therefore, $\lambda(t = t_1) > \lambda(t = t_1)$ since p is non-decreasing because of the fact that survival to a later time is only possible if all younger ages are attained.

The framework proposed herein allows a transition from pairwise rankings to survival prediction and it is applicable for all pairwise ranking methods. For solving the learning to rank problem, RSVM is chosen as it is a well-studied and a tested method in diverse domains. There are also several efficient training methodologies for SVM to handle ranking tasks. In this work, Chapella's RSVM is used in the implementation of RsurVM[111, 113], which performs with linear time complexity. Similarly, Joachims solves the optimization problem of ranking SVM in linear time[112]. Moreover, new versions[114, 115] and embedded feature selection methodologies[116, 117], which are proposed for SVM for ranking, can be directly applied to survival prediction. However, to be able to compare the proposed method with Cox-ph and RSF on fair ground, only the core implementation of RSVM is used without integrating any feature selection methods.

4.2 Results of RsurVM

Empirical tests are conducted with all molecular data types on RsurVM and the resulting performance is compared against that of Cox-ph and RSF. In this part the same 100 bootstrapping training-test set pairs are used. Comparisons are based on C-index. CNV and RPPA are tested without any feature selection step while DNAmethy, miRNA and mRNA are tested with the features obtained from Cox-screening to avoid the high dimensionality of molecular data.

The current formulation of RsurVM necessitates a penalization parameter C to be set. C represents the trade-off between training error and model complexity. C parameter adjusts the amount of misclassifying allowed for each training sample. For large values of C, the optimization problem will try to classify all samples correctly, so there is an inherent risk of overfitting. Conversely, very small values will result in an overly simplified model, which might accept misclassified training samples thus making the model prone to underfitting[118].

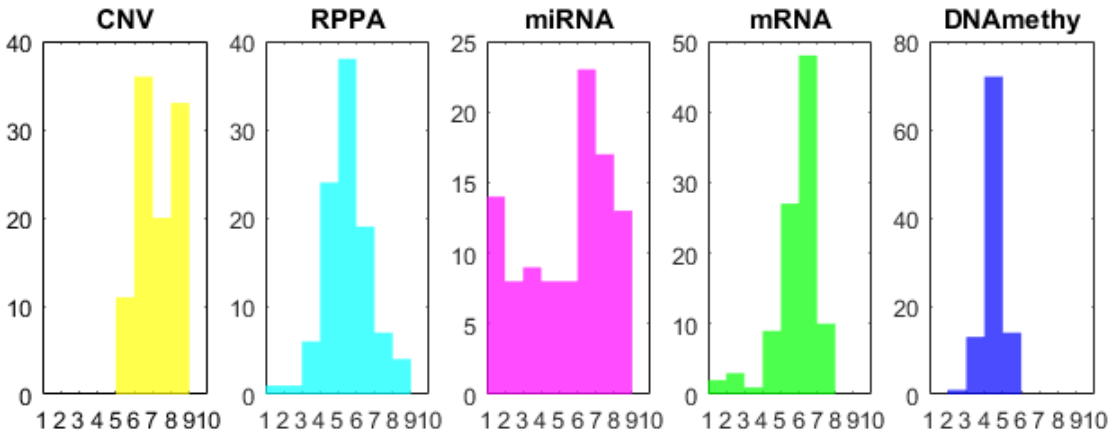

Figure 4.1: Distributions of estimated penalization parameters, C, resulted in 100 bootstrapping, for each molecular data. Indices $1, 2 \ldots, 9$ correspond to $C = 10^0, 10^{-1} \ldots, 10^{-8}$.

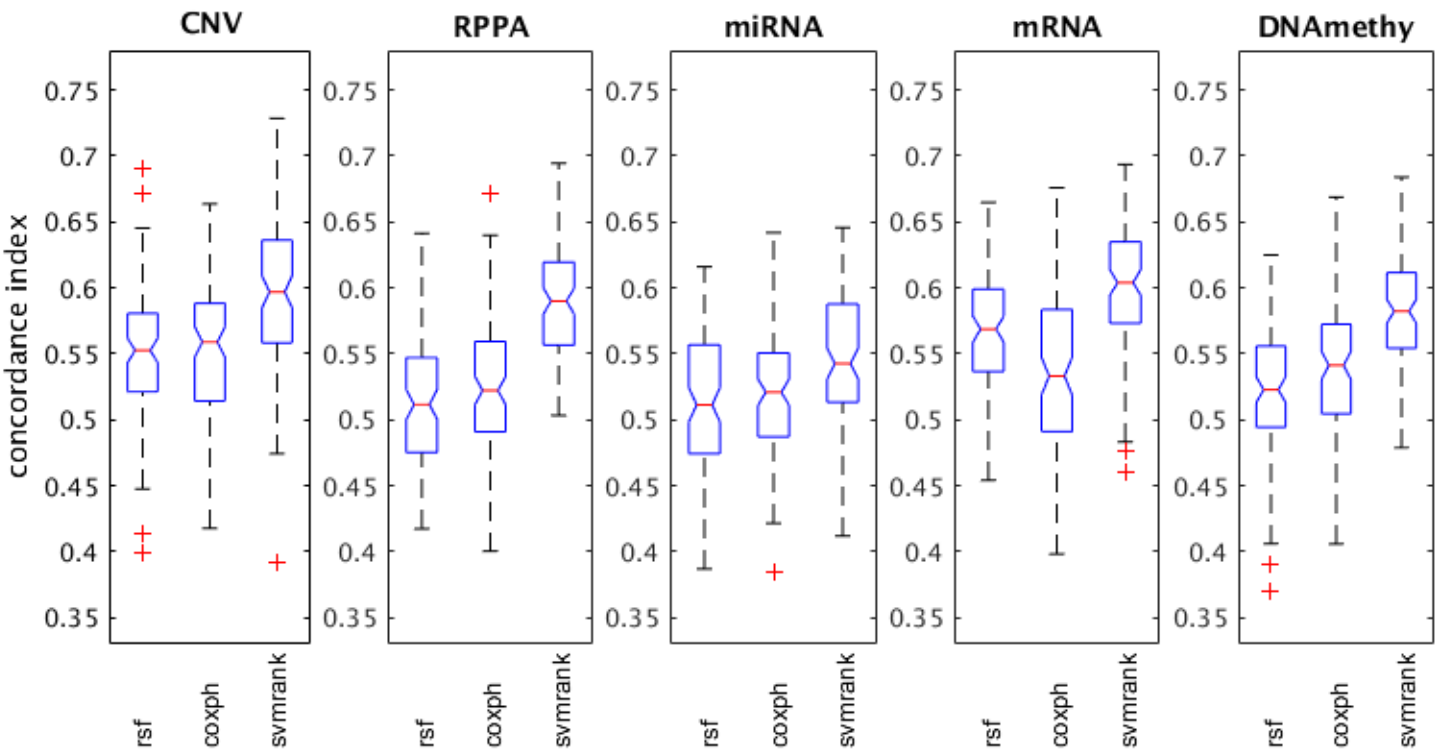

Figure 4.2: Comparison of RsurVM with Cox-ph and RSF

Table 4.1: Quantiles of RSF,Cox-ph and RsurVM resulted by 100 bootstrappings

Quantiles	CNV			RPPA			miRNA			mRNA			DNAmethy		
	Cox-ph	*RSF*	***RsurVM***	*Cox-ph*	*RSF*	***RsurVM***	*Cox-ph*	*RSF*	***RsurVM***	*Cox-ph*	*RSF*	***RsurVM***	*Cox-ph*	*RSF*	***RsurVM***
1^{st}	.42	.40	.39	.40	.42	.50	.39	.39	.41	.40	.45	.46	.41	.37	.48
2^{nd}	.51	.52	.56	.49	.47	.56	.49	.47	.51	.49	.54	.57	.50	.49	.55
median	**.56**	**.55**	**.60**	**.52**	**.51**	**.59**	**.52**	**.51**	**.54**	**.53**	**.57**	**.60**	**.54**	**.52**	**.58**
3^{rd}	.59	.58	.64	.56	.55	.62	.55	.56	.59	.58	.60	.64	.57	.55	.61
4^{th}	.66	.69	.73	.67	.64	.70	.64	.62	.65	.68	.66	.69	.67	.62	.68

RsurVM C values are selected with 5-fold censored stratified cross validation. Stratifying folds that takes into account censored cases enables RsurVM to estimate C more robustly. For each fold, an optimum C value is obtained as a result. RsurVM assigns the median of these C values to the final model. Figure 4.1 illustrates the distribution of C values estimated for each molecular data indicating that training CNV with soft margin SVM (less complex model) is more plausible while $DNAmethy$ is the more generalizable molecular data; therefore, it is less prone to overfitting. $mRNA$ and $RPPA$ ranks between CNV and $DNAmethy$ in terms of model complexity. It is hard to infer a conclusion for $miRNA$ relating to C, since it varies more according to bootstrapping samples.

RsurVM outperforms Cox-ph and RSF in all molecular data. Wilcoxin signed rank test results are not reported in this work for RsurVM since p-value $\simeq 0$ for all molecular data. In conjunction with the previous results, $miRNA$ is the less plausible data to estimate hazards of samples. This result consequently coincides with the results obtained by Cox-ph, RSF and NMF. Moreover, the least improvement is observed in $miRNA$; however, RSurVM more accurately predicts the survivals than other methods.

RsurVM shows a notable improvement in $DNAmethy$. Although $DNAmethy$ is not as promising as other molecular data, RsurVM still increases its performance to a level higher than the performances of other molecular data handled with Cox-ph or RSF. $RPPA$, $mRNA$ and CNV are the two best molecular data in estimating the survivals. RsurVM best fits to $RPPA$. $RPPA$ is among the least predictive data for Cox-ph and RSF whereas it still yields better results than $mRNA$, which performs best in most of the cases. This may indicate that pairwise relation between patients is more observable in protein expressions. $mRNA$ again yields more promising performance due to the direct influence of its expression upon the in cell mechanism. Interestingly, CNV yields one of the best results for all methods including RsurVM.

In conclusion, RsurVM optimizes the concordance index metric and handles the

censored data without making any assumptions. With extensive empirical tests, we demonstrate that RsurVM overperforms the predictive ability of all molecular data significantly without any exceptions. It is possibly stemming from the robustness of RsurVM adapted from the RSVM. Apart from the prediction accuracy of RsurVM discussed earlier, very fast training run time in RsurVM is observed. Another advantage of RsurVM is that the model obtained by this method is amenable for biological interpretation unlike other SVM models since it utilizes a linear kernel, which is very simple.

4.2.1 Variable Importance in RsurVM

For the linear kernel SVM, variable importance can be inferred from the weights, $\hat{w}$. In linear SVM, weights formulate the hyperplane, which separates the classes, by indicating the coordinates of an orthogonal vector to the hyperplane. The dot product of $\hat{w}$ can tell on which side a point is to classify samples. Thus, absolute weight of each feature can tell us the importance of each variable[119]. Using this approach, we calculated the variable importance of features in RsurVM for each bootstrapping sample and averaged this variable importance. Figures 4.3, 4.4, 4.5, 4.6, 4.7 show the significant features of molecular data obtained by RsurVM.

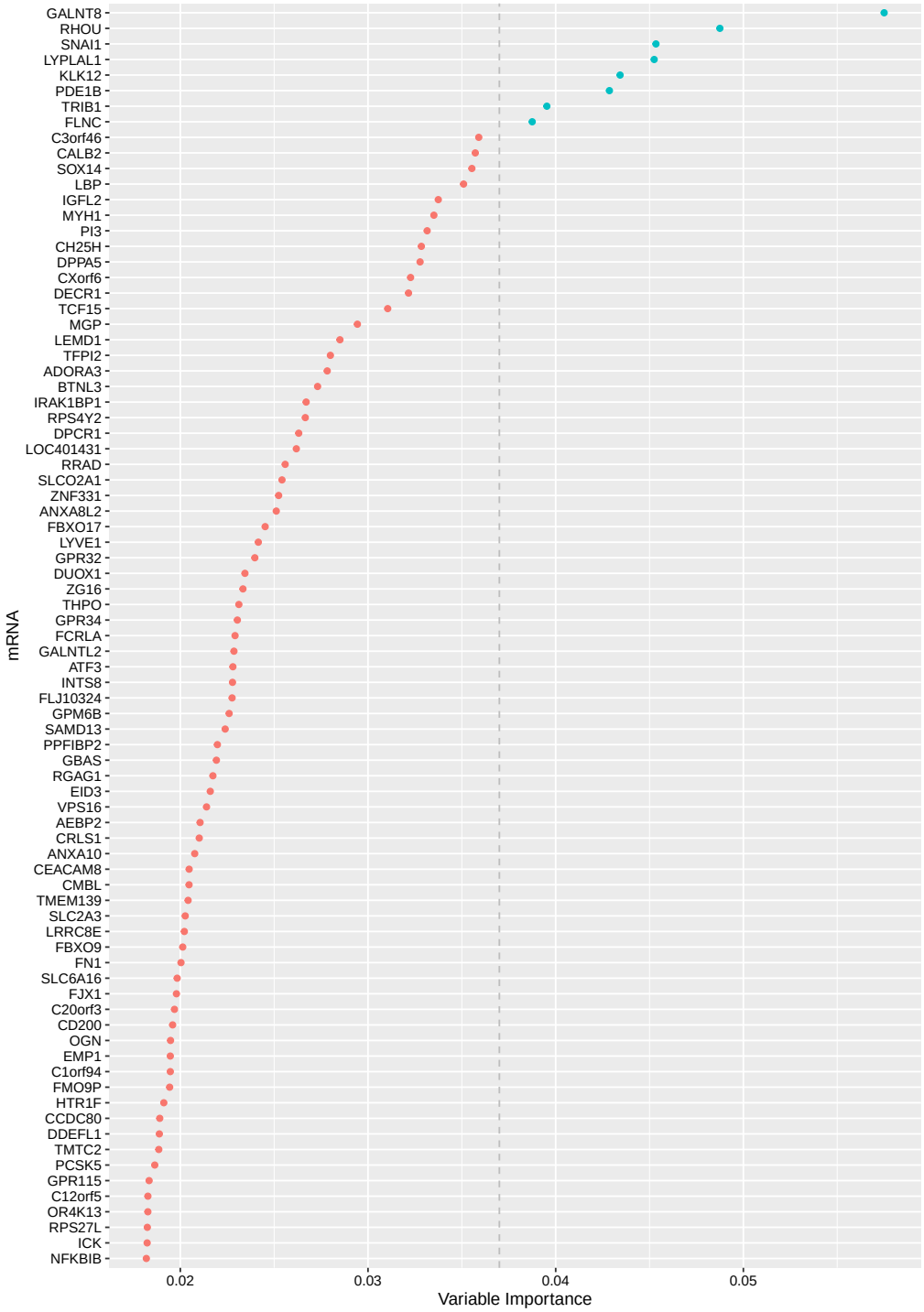

Figure 4.3: Significant mRNA features obtained by RsurVM, blues are top ranked features.

Figure 4.4: Significant RPPA features obtained by RsurVM, blues are top ranked features.

Figure 4.5: Significant miRNA features obtained by RsurVM, blues are top features ranked features.

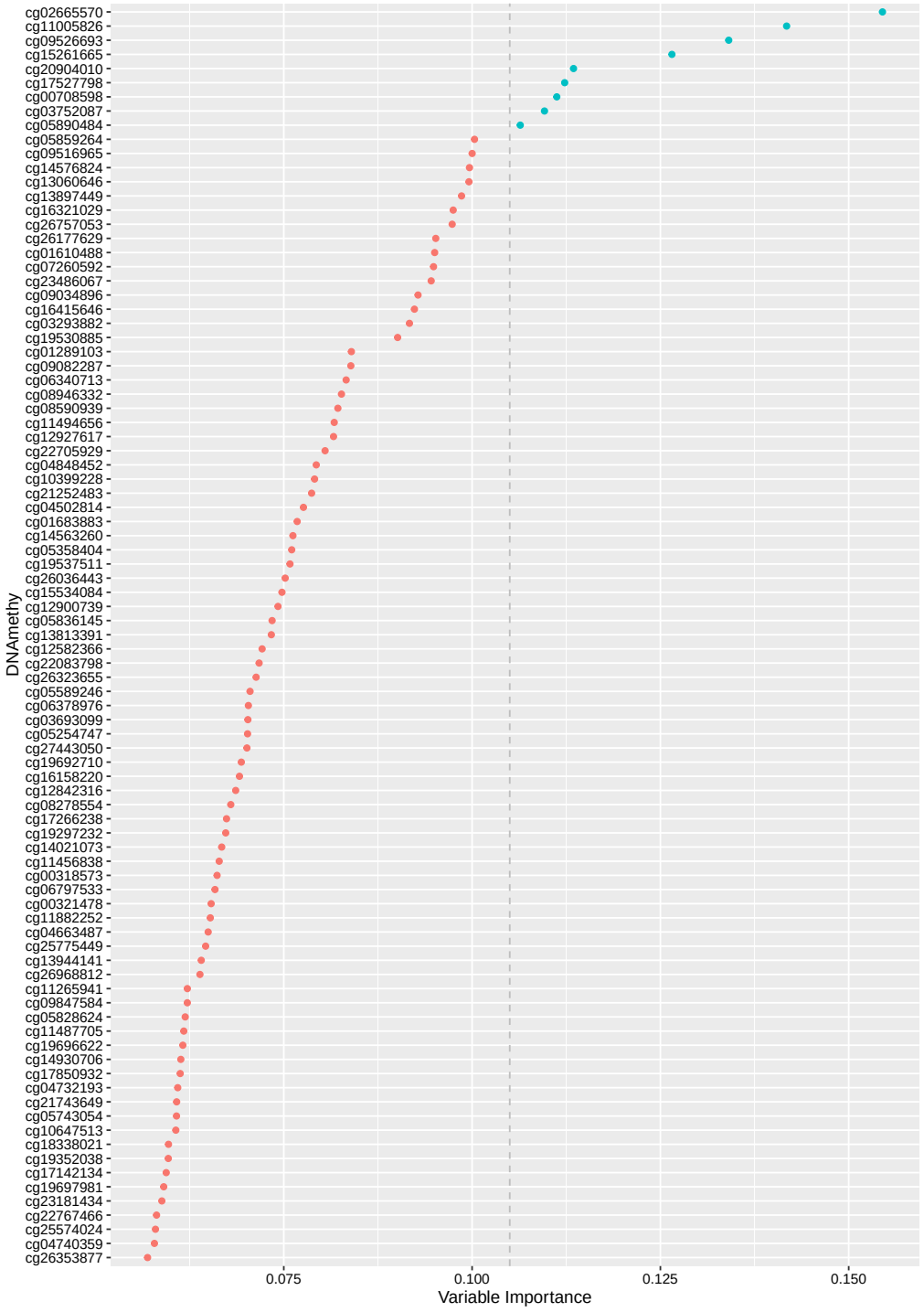

Figure 4.6: Significant DNA-methylation features obtained by RsurVM, blues are top ranked features.

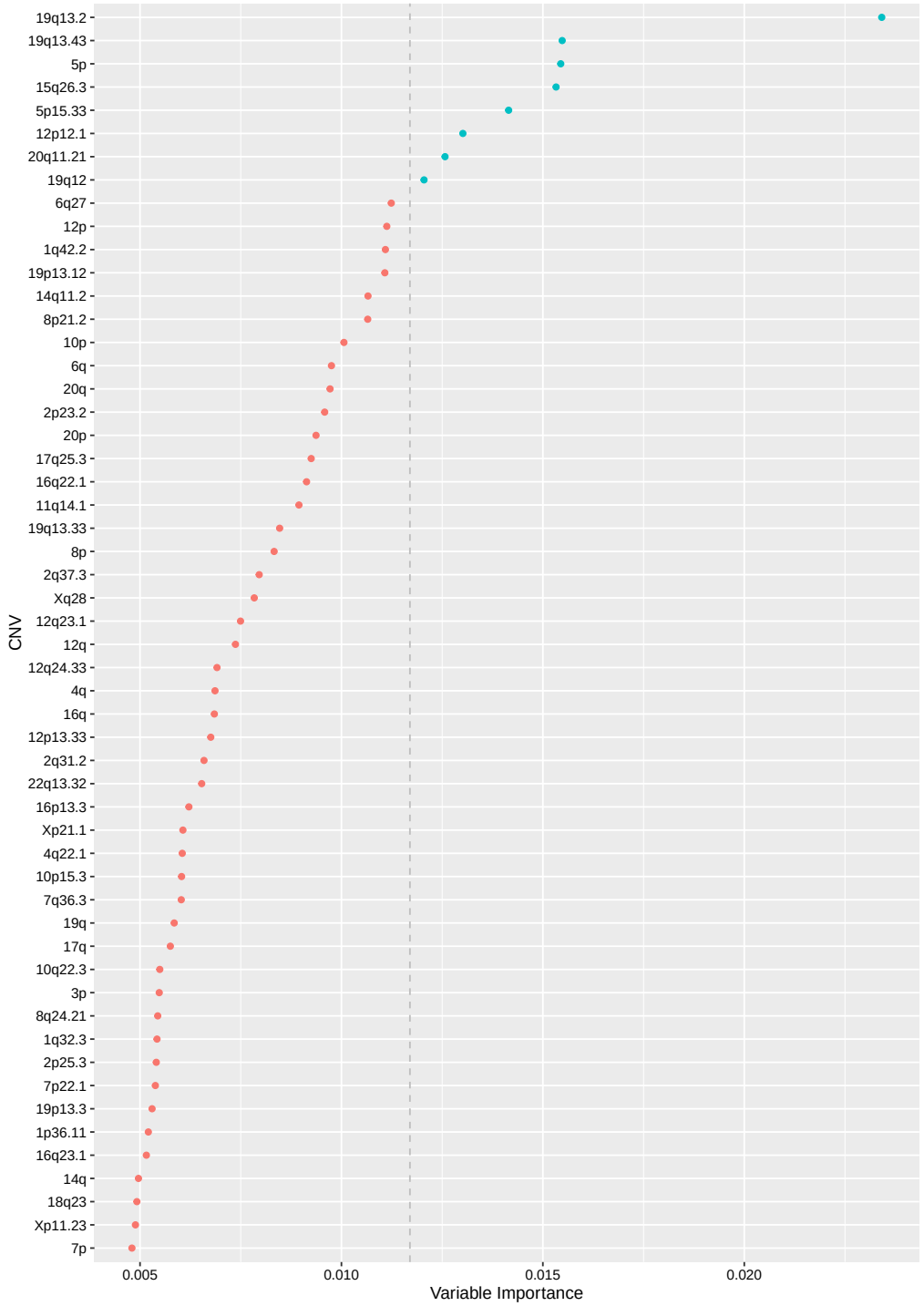

Figure 4.7: Significant CNV features obtained by RsurVM, blues are top ranked features.

Amplification of GAB2 (GRB2-associated binding protein 2) gene and increased expression of this gene have role in tumorigenesis process and progression in ovarian cancer[120, 121] However, high expression of it is related with better clinical outcome[122]. Ovarian cancer stem cells responsible for high recurrence rate of disease are enriched in c-Kit, a tyrosine kinase which is not a prognostic factor for ovarian cancer yet[123, 124]. In the future, MEK1 (mitogen-activated protein kinase kinase 1) may be used to choose ovarian cancer patients with poor response to platinum based therapy for application of alternative therapies[125]. Prognostic value of miR-452 miR-BART3 miR-BART15 miR-920 miR-380 and effect of these microRNAs in ovarian cancer have not been studied as far as we investigated.

Chapter 5

Conclusion and Future Work

In this thesis, we first proposed a method that attains partial ordering in feature space and abbreviated it as POF. Intensive empirical experiments made on Cox-ph model and RSF demonstrated that POF representation outperforms modes trained with individual molecular data in each model. For all molecular data types, POF yielded statistically significant improvements. Particularly, RSF with POF representation achieved the best performance for all cases. Besides being better in continuous survival prediction, it also illustrated more significant survival subgroup clustering. For all molecular data types, POF found more significantly clustered sub groups. While clustering quality mostly coincides with survival distribution difference in POF, molecular data did not produce significant correlations between clusters and the associated survival distributions. The results obtained from mRNA and RPPA were especially noticeable. Without POF, mRNA could not cluster significantly for any cases whereas POF yielded better clustering. In RPPA, while molecular data and POF yielded different number of clusters, the clustering suggested by POF was more significant. Candidate models proposed by POF seem promising since most of the suggested molecules have already annotated by recent studies in the literature.

Secondly, we proposed a pairwise ranking in sample space to estimate survival of patients more precisely and a ranking based survival prediction model, RsurVM. We outlined a general framework that suggests the applicability of for pairwise ranking methods in information literature to survival estimation. We implemented this framework on ranking SVM. RsurVM was shown to outperform two well established methods Cox-ph and RSF. Finally we described variable importance in RsurVM and reported significant features of each molecular data.

There are several directions to pursue in a future work that builds on the present study which can be listed as follows:

- POF can be improved by establishing a new methodology for feature elimination. Due to feature explosion, we do not consider all the features to create POF. A new pre-selection step can be devised in lieu of univariate Cox screening. Our preliminary work (data not shown) that involves log-rank screening gives promising results in this direction.
- Post feature filtering can be applied in order to investigate the correlations between candidate features and eliminated features. If some eliminated features are found to be highly correlated, the model can be modified accordingly.
- By introducing a normalization step, POF can be enhanced both for ordering within a single molecular data type and for ordering between different molecular data.
- RsurVM can be enriched with embedded feature selection strategies originally proposed for SVM. A particular advantage of RsurVM is that all the methods applicable for linear kernel SVM can be readily applied to RsurVM. Finally, other pairwise ranking methods can be motivated from the framework proposed.

Bibliography

[1] "Cancer facts and figures 2014." http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/. Accessed: 2016-05-03.

[2] S. Sharma, T. K. Kelly, and P. A. Jones, "Epigenetics in cancer," *Carcinogenesis*, vol. 31, no. 1, pp. 27–36, 2010.

[3] C. S. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," *Journal of applied genetics*, vol. 52, no. 4, pp. 413–435, 2011.

[4] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, M. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, 2010.

[5] C. G. A. R. Network *et al.*, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[6] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature methods*, vol. 10, no. 11, pp. 1108–1115, 2013.

[7] J. Galon, A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, C. Lagorce-Pagès, M. Tosolini, M. Camus, A. Berger, P. Wind, *et al.*, "Type, density, and location of immune cells within human colorectal tumors predict clinical outcome," *Science*, vol. 313, no. 5795, pp. 1960–1964, 2006.

[8] J. Haybittle, R. Blamey, C. Elston, J. Johnson, P. Doyle, F. Campbell, R. Nicholson, and K. Griffiths, "A prognostic index in primary breast cancer.," *British journal of cancer*, vol. 45, no. 3, p. 361, 1982.

[9] G. L. Lu-Yao and S.-L. Yao, "Population-based study of long-term survival in patients with clinically localised prostate cancer," *The Lancet*, vol. 349, no. 9056, pp. 906–910, 1997.

[10] H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde, "Predicting survival from microarray data?a comparative study," *Bioinformatics*, vol. 23, no. 16, pp. 2080–2087, 2007.

[11] M. Zou, Z. Liu, X.-S. Zhang, and Y. Wang, "Ncc-auc: an auc optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data," *Bioinformatics*, vol. 31, no. 20, pp. 3330–3338, 2015.

[12] P. C. Boutros, "The path to routine use of genomic biomarkers in the cancer clinic," *Genome research*, vol. 25, no. 10, pp. 1508–1513, 2015.

[13] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[14] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, ACM, 2002.

[15] B. Aslan, P. Monroig, M.-C. Hsu, G. A. Pena, C. Rodriguez-Aguayo, V. Gonzalez-Villasana, R. Rupaimoole, A. S. Nagaraja, S. Mangala, H.-D. Han, *et al.*, "The znf304-integrin axis protects against anoikis in cancer," *Nature communications*, vol. 6, 2015.

[16] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in neural information processing systems*, pp. 1209–1216, 2008.

[17] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part iv: further concepts and methods in survival analysis," *British Journal of Cancer*, vol. 89, no. 5, p. 781, 2003.

[18] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary oncology*, vol. 19, no. 1A, p. A68, 2015.

[19] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society. Series A (General)*, pp. 185–207, 1972.

[20] M. J. Bradburn, T. G. Clark, S. Love, and D. Altman, "Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods," *British journal of cancer*, vol. 89, no. 3, pp. 431–436, 2003.

[21] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[22] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, pp. 841–860, 2008.

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

[24] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[25] F. E. Harrell, K. L. Lee, and D. B. Mark, "Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, pp. 361–387, 1996.

[26] D. R. Cox and D. Oakes, *Analysis of survival data*, vol. 21. CRC Press, 1984.

[27] R. L. Prentice and J. D. Kalbfleisch, *The statistical analysis of failure time data*. New York: John Wiley, 2011.

[28] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] V. Van Belle, K. Pelckmans, J. A. Suykens, and S. Van Huffel, "Survival svm: a practical scalable algorithm.," in *ESANN*, pp. 89–94, 2008.

[32] F. E. Ahmed, "Artificial neural networks for diagnosis and survival prediction in colon cancer," *Molecular cancer*, vol. 4, no. 1, p. 29, 2005.

[33] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," *PLoS Biol*, vol. 2, no. 4, p. E108, 2004.

[34] S. M. Pagnotta and M. Ceccarelli, "An algorithm for finding gene signatures supervised by survival time data," in *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 568–578, Springer, 2011.

[35] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer," *Journal of biomedical informatics*, vol. 56, pp. 220–228, 2015.

[36] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via adaboost algorithms," in *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80*, pp. 55–64, Australian Computer Society, Inc., 2008.

[37] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[38] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, "Survival-time classification of breast cancer patients," *Computational Optimization and Applications*, vol. 25, no. 1-3, pp. 151–166, 2003.

[39] J.-S. Lee, I.-S. Chu, J. Heo, D. F. Calvisi, Z. Sun, T. Roskams, A. Durnez, A. J. Demetris, and S. S. Thorgeirsson, "Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling," *Hepatology*, vol. 40, no. 3, pp. 667–676, 2004.

[40] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[41] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[42] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[43] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[44] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression

with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.

[45] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[46] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the national academy of sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[47] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, *et al.*, "Assessing the clinical utility of cancer genomic and proteomic data across tumor types," *Nature biotechnology*, vol. 32, no. 7, pp. 644–652, 2014.

[48] H. Gómez-Rueda, E. Martínez-Ledesma, A. Martínez-Torteya, R. Palacios-Corona, and V. Trevino, "Integration and comparison of different genomic data for outcome prediction in cancer," *BioData mining*, vol. 8, no. 1, p. 1, 2015.

[49] L. Xu, L. Fengji, L. Changning, Z. Liangcai, L. Yinghui, L. Yu, C. Shanguang, and X. Jianghui, "Comparison of the prognostic utility of the diverse molecular data among lncrna, dna methylation, microrna, and mrna across five human cancers," *PloS one*, vol. 10, no. 11, p. e0142433, 2015.

[50] D. Kim, J.-G. Joung, K.-A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie, and J. H. Kim, "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 109–120, 2015.

[51] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," *PLoS One*, vol. 6, no. 11, p. e24709, 2011.

[52] M. Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.

[53] T. Therneau, "A package for survival analysis in s. r package version 2.37-4," *URL http://CRAN. R-project. org/package= survival. Box*, vol. 980032, pp. 23298–0032, 2013.

[54] H. Ishwaran *et al.*, "Variable importance in binary regression trees and forests," *Electronic Journal of Statistics*, vol. 1, pp. 519–537, 2007.

[55] H. Ishwaran and U. Kogalur, "Package ?randomforestsrc.? 2015. ht tp," *cran. r-project. org/web/packages/randomForestSRC/randomForestSRC. pdf. Accessed March*, vol. 23, 2015.

[56] S. Ghafouri-Fard, Z. O. Ashtiani, B. S. Golian, S.-M. Hasheminasab, and M. H. Modarressi, "Expression of two testis-specific genes, spata19 and lemd1, in prostate cancer," *Archives of medical research*, vol. 41, no. 3, pp. 195–200, 2010.

[57] D. Yuki, Y.-M. Lin, Y. Fujii, Y. Nakamura, and Y. Furukawa, "Isolation of lem domain-containing 1, a novel testis-specific gene expressed in colorectal cancers," *Oncology reports*, vol. 12, no. 2, pp. 275–280, 2004.

[58] A.-Y. Kwon, G.-I. Kim, J.-Y. Jeong, J.-Y. Song, K.-B. Kwack, C. Lee, H.-Y. Kang, T.-H. Kim, J.-H. Heo, and H. J. An, "Vav3 overexpressed in cancer stem cells is a poor prognostic indicator in ovarian cancer patients," *Stem cells and development*, vol. 24, no. 13, pp. 1521–1535, 2015.

[59] L. Du, X. Qian, C. Dai, L. Wang, D. Huang, S. Wang, and X. Shen, "Screening the molecular targets of ovarian cancer based on bioinformatics analysis.," *Tumori*, vol. 101, no. 4, pp. 384–389, 2015.

[60] H. W. Seo, D. Rengaraj, J. W. Choi, S. E. Ahn, Y. S. Song, G. Song, and J. Y. Han, "Claudin 10 is a glandular epithelial marker in the chicken model as human epithelial ovarian cancer," *International Journal of Gynecological Cancer*, vol. 20, no. 9, pp. 1465–1473, 2010.

[61] D. Lai, L. Ma, and F. Wang, "Fibroblast activation protein regulates tumor-associated fibroblasts and epithelial ovarian cancer cells," *International journal of oncology*, vol. 41, no. 2, pp. 541–550, 2012.

[62] C.-H. Tseng, K. D. Murray, M.-F. Jou, S.-M. Hsu, H.-J. Cheng, and P.-H. Huang, "Sema3e/plexin-d1 mediated epithelial-to-mesenchymal transition in ovarian endometrioid cancer," *PloS one*, vol. 6, no. 4, p. e19396, 2011.

[63] S. Lauttia, H. Sihto, H. Kavola, V. Koljonen, T. Böhling, and H. Joensuu, "Prokineticins and merkel cell polyomavirus infection in merkel cell carcinoma," *British journal of cancer*, vol. 110, no. 6, pp. 1446–1455, 2014.

[64] A. Handra-Luca, H. Bilal, J.-C. Bertrand, and P. Fouret, "Extra-cellular signal-regulated erk-1/erk-2 pathway activation in human salivary gland mucoepidermoid carcinoma: association to aggressive tumor behavior and tumor cell proliferation," *The American journal of pathology*, vol. 163, no. 3, pp. 957–967, 2003.

[65] J. A. McCubrey, L. S. Steelman, W. H. Chappell, S. L. Abrams, E. W. Wong, F. Chang, B. Lehmann, D. M. Terrian, M. Milella, A. Tafuri, *et al.*, "Roles of the raf/mek/erk pathway in cell growth, malignant transformation and drug resistance," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1773, no. 8, pp. 1263–1284, 2007.

[66] P. G. Santamaria and A. R. Nebreda, "Deconstructing erk signaling in tumorigenesis," *Molecular cell*, vol. 38, no. 1, pp. 3–5, 2010.

[67] M. T. Rahman, K. Nakayama, M. Rahman, H. Katagiri, A. Katagiri, T. Ishibashi, M. Ishikawa, E. Sato, K. Iida, N. Nakayama, *et al.*, "Kras

and mapk1 gene amplification in type ii ovarian carcinomas," *International journal of molecular sciences*, vol. 14, no. 7, pp. 13748–13762, 2013.

[68] J.-Y. Yang, C. S. Zong, W. Xia, H. Yamaguchi, Q. Ding, X. Xie, J.-Y. Lang, C.-C. Lai, C.-J. Chang, W.-C. Huang, *et al.*, "Erk promotes tumorigenesis by inhibiting foxo3a via mdm2-mediated degradation," *Nature cell biology*, vol. 10, no. 2, pp. 138–148, 2008.

[69] R. Ley, K. Balmanno, K. Hadfield, C. Weston, and S. J. Cook, "Activation of the erk1/2 signaling pathway promotes phosphorylation and proteasome-dependent degradation of the bh3-only protein, bim," *Journal of Biological Chemistry*, vol. 278, no. 21, pp. 18811–18816, 2003.

[70] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch, "Map kinase signalling pathways in cancer," *Oncogene*, vol. 26, no. 22, pp. 3279–3290, 2007.

[71] D. Guardavaccaro and H. Clevers, "Wnt/β-catenin and mapk signaling: allies and enemies in different battlefields," *Sci. Signal.*, vol. 5, no. 219, pp. pe15–pe15, 2012.

[72] Y. Xia, Y.-L. Zhang, C. Yu, T. Chang, and H.-Y. Fan, "Yap/tead co-activator regulated pluripotency and chemoresistance in ovarian cancer initiated cells," *PloS one*, vol. 9, no. 11, p. e109575, 2014.

[73] T. M. Abdel-Fatah, A. Arora, P. Moseley, C. Coveney, C. Perry, K. Johnson, C. Kent, G. Ball, S. Chan, and S. Madhusudan, "Atm, atr and dna-pkcs expressions correlate to adverse clinical outcomes in epithelial ovarian cancers," *BBA clinical*, vol. 2, pp. 10–17, 2014.

[74] K. Aaltonen, R.-M. Amini, P. Heikkilä, K. Aittomäki, A. Tamminen, H. Nevanlinna, and C. Blomqvist, "High cyclin b1 expression is associated with poor survival in breast cancer," *British journal of cancer*, vol. 100, no. 7, pp. 1055–1060, 2009.

[75] H. Zheng, W. Hu, M. T. Deavers, D.-Y. Shen, S. Fu, Y.-F. Li, and J. J. Kavanagh, "Nuclear cyclin b1 is overexpressed in low-malignant-potential

ovarian tumors but not in epithelial ovarian cancer," *American journal of obstetrics and gynecology*, vol. 201, no. 4, pp. 367–e1, 2009.

[76] J. Luo, B. D. Manning, and L. C. Cantley, "Targeting the pi3k-akt pathway in human cancer: rationale and promise," *Cancer cell*, vol. 4, no. 4, pp. 257–262, 2003.

[77] Y. Ou, L. Ma, L. Ma, Z. Huang, W. Zhou, C. Zhao, B. Zhang, Y. Song, C. Yu, and Q. Zhan, "Overexpression of cyclin b1 antagonizes chemotherapeutic-induced apoptosis through pten/akt pathway in human esophageal squamous cell carcinoma cells," *Cancer biology & therapy*, vol. 14, no. 1, pp. 45–55, 2013.

[78] F.-Y. Li, X.-B. Ren, X.-Y. Xie, and J. Zhang, "Meta-analysis of excision repair cross-complementation group 1 (ercc1) association with response to platinum-based chemotherapy in ovarian cancer," *Asian Pac J Cancer Prev*, vol. 14, no. 12, pp. 7203–6, 2013.

[79] C. Ip, A. Cheung, H. Ngan, and A. Wong, "p70 s6 kinase in the control of actin cytoskeleton dynamics and directed migration of ovarian cancer cells," *Oncogene*, vol. 30, no. 21, pp. 2420–2432, 2011.

[80] C. K. Ip and A. S. Wong, "Exploiting p70 s6 kinase as a target for ovarian cancer," *Expert opinion on therapeutic targets*, vol. 16, no. 6, pp. 619–630, 2012.

[81] F. W. Liu and K. S. Tewari, "New targeted agents in gynecologic cancers: Synthetic lethality, homologous recombination deficiency, and parp inhibitors," *Current treatment options in oncology*, vol. 17, no. 3, pp. 1–15, 2016.

[82] M. Kim, L. Hernandez, and C. Annunziata, "Caspase 8 expression may determine the survival of women with ovarian cancer," *Cell Death & Disease*, vol. 7, no. 1, p. e2045, 2016.

[83] M. Lu, Y. Wang, F. Xu, J. Xiang, and D. Chen, "The prognostic of p27kip1 in ovarian cancer: a meta-analysis," *Archives of gynecology and obstetrics*, vol. 293, no. 1, pp. 169–176, 2016.

[84] C.-W. Lin, Y.-L. Chang, Y.-C. Chang, J.-C. Lin, C.-C. Chen, S.-H. Pan, C.-T. Wu, H.-Y. Chen, S.-C. Yang, T.-M. Hong, *et al.*, "Microrna-135b promotes lung cancer metastasis by regulating multiple targets in the hippo pathway and lzts1," *Nature communications*, vol. 4, p. 1877, 2013.

[85] A. J. Lowery, N. Miller, A. Devaney, R. E. McNeill, P. A. Davoren, C. Lemetre, V. Benes, S. Schmidt, J. Blake, G. Ball, *et al.*, "Microrna signatures predict oestrogen receptor, progesterone receptor and her2/neu receptor status in breast cancer," *Breast Cancer Res*, vol. 11, no. 3, p. R27, 2009.

[86] N. Valeri, C. Braconi, P. Gasparini, C. Murgia, A. Lampis, V. Paulus-Hock, J. R. Hart, L. Ueno, S. I. Grivennikov, F. Lovat, *et al.*, "Microrna-135b promotes cancer progression by acting as a downstream effector of oncogenic pathways in colon cancer," *Cancer cell*, vol. 25, no. 4, pp. 469–483, 2014.

[87] L. Wang, M.-J. Zhu, A.-M. Ren, H.-F. Wu, W.-M. Han, R.-Y. Tan, and R.-Q. Tu, "A ten-microrna signature identified from a genome-wide microrna expression profiling in human epithelial ovarian cancer," *PLoS One*, vol. 9, no. 5, p. e96472, 2014.

[88] J.-Y. Fan, Y. Yang, J.-Y. Xie, Y.-L. Lu, K. Shi, and Y.-Q. Huang, "Microrna-144 mediates metabolic shift in ovarian cancer cells by directly targeting glut1," *Tumor Biology*, pp. 1–6, 2015.

[89] H. Yang, W. Kong, L. He, J.-J. Zhao, J. D. O'Donnell, J. Wang, R. M. Wenham, D. Coppola, P. A. Kruk, S. V. Nicosia, *et al.*, "Microrna expression profiling in human ovarian cancer: mir-214 induces cell survival and cisplatin resistance by targeting pten," *Cancer research*, vol. 68, no. 2, pp. 425–433, 2008.

[90] Y.-W. Kim, E. Y. Kim, D. Jeon, J.-L. Liu, H. S. Kim, J. W. Choi, and W. S. Ahn, "Differential microrna expression signatures and cell type-specific association with taxol resistance in ovarian cancer cells," 2014.

[91] T. Shuang, M. Wang, C. Shi, Y. Zhou, and D. Wang, "Down-regulated expression of mir-134 contributes to paclitaxel resistance in human ovarian cancer cells," *FEBS letters*, vol. 589, no. 20PartB, pp. 3154–3164, 2015.

[92] J. Li, S. Liang, H. Yu, J. Zhang, D. Ma, and X. Lu, "An inhibitory effect of mir-22 on cell migration and invasion in ovarian cancer," *Gynecologic oncology*, vol. 119, no. 3, pp. 543–548, 2010.

[93] D.-X. Peng, M. Luo, L.-W. Qiu, Y.-L. He, and X.-F. Wang, "Prognostic implications of microrna-100 and its functional roles in human epithelial ovarian cancer," *Oncology reports*, vol. 27, no. 4, pp. 1238–1244, 2012.

[94] W. Wan, Y. Zhang, X. Wang, Y. Liu, Y. Zhang, Y. Que, W.-j. Zhao, and P. Li, "Down-regulated mir-22 as predictive biomarkers for prognosis of epithelial ovarian cancer," *Diagnostic pathology*, vol. 9, no. 1, p. 178, 2014.

[95] A. Gadducci, C. Sergiampietri, N. Lanfredini, and I. Guiggi, "Micro-rnas and ovarian cancer: the state of art and perspectives of clinical research," *Gynecological Endocrinology*, vol. 30, no. 4, pp. 266–271, 2014.

[96] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[97] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[98] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, *et al.*, "Gene expression

patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10869–10874, 2001.

[99] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology*, vol. 6, no. 3-4, pp. 281–297, 1999.

[100] N. A. Makretsov, D. G. Huntsman, T. O. Nielsen, E. Yorida, M. Peacock, M. C. Cheang, S. E. Dunn, M. Hayes, M. van de Rijn, C. Bajdik, *et al.*, "Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma," *Clinical cancer research*, vol. 10, no. 18, pp. 6143–6151, 2004.

[101] A. R. Brannon, A. Reddy, M. Seiler, A. Arreola, D. T. Moore, R. S. Pruthi, E. M. Wallen, M. E. Nielsen, H. Liu, K. L. Nathanson, *et al.*, "Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns," *Genes & cancer*, vol. 1, no. 2, pp. 152–163, 2010.

[102] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, *et al.*, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[103] J. Jiang, Y. Gusev, I. Aderca, T. A. Mettler, D. M. Nagorney, D. J. Brackett, L. R. Roberts, and T. D. Schmittgen, "Association of microrna expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival," *Clinical Cancer Research*, vol. 14, no. 2, pp. 419–427, 2008.

[104] W. A. Freije, F. E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. M. Liau, P. S. Mischel, and S. F. Nelson, "Gene expression profiling of gliomas strongly predicts survival," *Cancer research*, vol. 64, no. 18, pp. 6503–6510, 2004.

[105] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, pp. 33–40, 1962.

[106] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[107] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[108] D. Cossock and T. Zhang, "Subset ranking using regression," in *Learning theory*, pp. 605–619, Springer, 2006.

[109] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, ACM, 2007.

[110] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[111] O. Chapelle, "Training a support vector machine in the primal," *Neural computation*, vol. 19, no. 5, pp. 1155–1178, 2007.

[112] T. Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 217–226, ACM, 2006.

[113] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, vol. 13, no. 3, pp. 201–215, 2010.

[114] J. Xu, Y. Cao, H. Li, and Y. Huang, "Cost-sensitive learning of svm for ranking," in *Machine Learning: ECML 2006*, pp. 833–840, Springer, 2006.

[115] H. Yu, Y. Kim, and S. Hwang, "Rv-svm: An efficient method for learning ranking svm," in *Advances in Knowledge Discovery and Data Mining*, pp. 426–438, Springer, 2009.

[116] Q. Zhou, W. Hong, G. Shao, and W. Cai, "A new svm-rfe approach towards ranking problem," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, vol. 4, pp. 270–273, IEEE, 2009.

[117] L. Laporte, R. Flamary, S. Canu, S. Déjean, and J. Mothe, "Nonconvex regularizations for feature selection in ranking with sparse svm," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, no. 6, pp. 1118–1130, 2014.

[118] N.-E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of svm kernels," *Pattern Recognition*, vol. 38, no. 10, pp. 1733–1745, 2005.

[119] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[120] G. P. Dunn, H. W. Cheung, P. K. Agarwalla, S. Thomas, Y. Zektser, A. M. Karst, J. S. Boehm, B. A. Weir, A. M. Berlin, L. Zou, *et al.*, "In vivo multiplexed interrogation of amplified genes identifies gab2 as an ovarian cancer oncogene," *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 1102–1107, 2014.

[121] C. Duckworth, L. Zhang, S. Carroll, S. Ethier, and H. Cheung, "Overexpression of gab2 in ovarian cancer cells promotes tumor growth and angiogenesis by upregulating chemokine expression," *Oncogene*, 2015.

[122] S. J. Davis, K. E. Sheppard, M. S. Anglesio, J. George, N. Traficante, S. Fereday, M. P. Intermaggio, U. Menon, A. Gentry-Maharaj, J. Lubinski, *et al.*, "Enhanced gab2 expression is associated with improved survival in high-grade serous ovarian cancer and sensitivity to pi3k inhibition," *Molecular cancer therapeutics*, vol. 14, no. 6, pp. 1495–1503, 2015.

[123] W. Chau, C. Ip, A. Mak, H.-C. Lai, and A. Wong, "c-kit mediates chemoresistance and tumor-initiating capacity of ovarian cancer cells through activation of wnt/β-catenin–atp-binding cassette g2 signaling," *Oncogene*, vol. 32, no. 22, pp. 2767–2781, 2013.

[124] M. Medinger, M. Kleinschmidt, K. Mross, B. Wehmeyer, C. Unger, H.-E. Schaefer, R. Weber, and M. Azemar, "c-kit (cd117) expression in human tumors and its prognostic value: an immunohistochemical analysis," *Pathology & Oncology Research*, vol. 16, no. 3, pp. 295–301, 2010.

[125] Z. Pénzváltó, A. Lánczky, J. Lénárt, N. Meggyesházi, T. Krenács, N. Szoboszlai, C. Denkert, I. Pete, and B. Győrffy, "Mek1 is associated with carboplatin resistance and is a prognostic biomarker in epithelial ovarian cancer," *BMC cancer*, vol. 14, no. 1, p. 1, 2014.