GRAPHICAL MODELS FOR DE NOVO AND PATHWAY-BASED NETWORK PREDICTION OVER MULTI-MODAL HIGH-THROUGHPUT BIOLOGICAL DATA

by

Andrew James Sedgewick

B.A., Computer Science, Princeton University, 2010

Submitted to the Graduate Faculty of

The School of Medicine in partial fulfillment

of the requirements for the degree of

Ph.D. in Computational Biology

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Andrew James Sedgewick

It was defended on

June 29, 2016

and approved by

Greg Cooper, PhD, Professor, Department of Biomedical Informatics

Charles Vaske, PhD, Nantomics LLC

Hussein Tawbi, MD, PhD, Associate Professor, Department of Medicine

Larry Wasserman, PhD, Professor, Department of Statistics, Carnegie Mellon University

Advisor: Takis Benos, PhD, Professor, Department of Computational and Systems Biology

GRAPHICAL MODELS FOR DE NOVO AND PATHWAY-BASED NETWORK PREDICTION OVER MULTI-MODAL HIGH-THROUGHPUT BIOLOGICAL DATA

Andrew James Sedgewick, PhD

University of Pittsburgh, 2016

It is now a standard practice in the study of complex disease to perform many high-throughput -omic experiments (genome wide SNP, copy number, mRNA and miRNA expression) on the same set of patient samples. These multi-modal data should allow researchers to form a more complete, systems-level picture of a sample, but this is only possible if they have a suitable model for integrating the data. Due to the variety of data modalities and possible combinations of data, general, flexible integration methods that will be widely applicable in many settings are desirable. In this dissertation I will present my work using graphical models for *de novo* structure learning of both undirected and directed sparse graphs over a mixture of Gaussian and categorical variables. Using synthetic and biological data I will show that these models are useful for both variable selection and inference. Selecting the regularization parameters is an important challenge for these models so I will also cover stability based methods for efficiently setting these parameters, and for controlling the false discovery rate of edge predictions. I will also show results from a biological application to data from metastatic melanoma patients where our methods identified a PARP1 slice site variant that is predictive of response to chemotherapy. Finally, I present work incorporating miRNA into a pathway based graphical model called PARADIGM. This extension of the model allows us to study patient-specific changes in miRNA induced silencing in cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I want to thank Samantha for moving to Pittsburgh with me so I could pursue this degree, and for supporting me throughout the entire program, but particularly during the last few months of crunch time. I couldn't have done it without you! I also want to thank my parents and my family for all of their support, and for instilling a desire to learn in me.

Thanks to my collaborators on these projects, they would not have been possible without you. On the PARP1 project, thanks to Irina Abecassis for her tireless work in the wet lab and Marjorie Romkes for her guidance. Much of the software written for this thesis is built on the TETRAD project which was largely constructed and maintained by Joe Ramsey. Many thanks to Joe for his quick and friendly responses to questions and his openness with regards to sharing code in development and to incorporating my code into the project. Thanks to Joe, Clark Glymour and Peter Spirtes for all of their guidance on the causal learning project.

Thank you to my committee. Each member has contributed to this dissertation through a combination of their teaching, mentorship, previous research that I've attempted to build upon, and their insightful comments. Finally I'd like to thank my advisor, Takis, for his guidance and support throughout the years, and for the freedom he allowed me to pursue internships and new fields of study.

This dissertation is dedicated to the memory of David Jaeger.

1.0 INTRODUCTION

In the Era of Big Data researchers are presented with many new problems that standard statistical methods and models are not equipped to handle. The statistics and machine learning literature is ripe with novel approaches that may fit the needs of a given biological problem, but there are often hurdles to adapt these approaches to a given problem, including parameter selection, computational requirements, and normalization issues. This dissertation sits at the intersection of statistics and biology and therefore I will endeavor to present methods and models that are adaptable to a wide variety of biological problems and systems while presenting potential users with guidelines for the use of these tools.

Integrative analysis of data of different modalities and different sources is a common task in biomedical research. In particular, finding relationships between continuous and categorical variables with several levels can be challenging as it often requires non-intuitive methods such as conversion of categorical variables to binary dummy variables. We explore this mixed variable type setting in depth in this dissertation, and present strategies for learning both undirected and directed graphical models over these data.

Using directed modes to predict causal relationships between variables is especially desirable in the study of human disease, as we would like to be able to treat the aberrant biology that is causing the disease rather then the downstream variables that are merely side effects of the change in the causal variable. We focus on the difficult problem of predicting causality from observational data here which is common for exploratory studies in modern biology. Relationships predicted by our models are candidates for validation in directed laboratory experiments. Recently, high-throughput knockdown experiments have been performed to study causal relationships at genomic scale. While this data is invaluable for causal modeling in biology, these types of experiments are still expensive and relatively new, so we aim to be widely applicable to the enormous body of observational data currently available. We anticipate that as these large-scale perturbation studies become more wide-spread they will be a valuable source of data for both validation of and integration with causal search algorithms.

In this dissertation I will present several tools for integrative analysis of multiple types of data. Chapters 2 and 3 focus on finding networks of interactions between these multimodal data. Although there are large-scale databases of known interactions between DNA, RNA, protein and other epigenetic factors, we initially focus on *de novo* methods which search for interactions without any prior knowledge as a proof of concept, to allow for validation with prior knowledge and to allow for integration of variables that do not have good coverage in the literature such as clinical tests, patient traits, and newer biological data types. Chapter 2, adapted from an article we recently published (Sedgewick et al. 2016), presents work on learning undirected networks over mixed variable types. Chapter 3 uses the methods from chapter 2 as a starting point for directed causal search algorithms. In chapter 4 (adapted from (Abecassis et al. 2015)) I present an in depth look at a successful application of these network search algorithms to a cohort of patients with metastatic melanoma where we were able to identify a single nucleotide polymorphism (SNP) that predicts how patients will respond to treatment. Finally, in chapter 5 I tackle the problem of integrative network analysis from the other direction by extending

PARADIGM, an algorithm that depends heavily on prior knowledge of biological pathways, to handle a new type of data, miRNA.

2.0 UNDIRECTED GRAPHICAL MODELS WITH MIXED VARIABLES

Mixed graphical models (MGMs) are graphical models learned over a combination of continuous and discrete variables. These models provide both a network structure and a parameterized joint probability density over these heterogeneous variables, which are common in biomedical datasets. The network structure reveals the direct associations between variables and the joint probability density allows one to ask arbitrary probabilistic questions on the data. This information can be used for feature selection, classification and other important tasks. We studied the properties of MGM learning and applications of MGMs to high-dimensional data (biological and simulated). Our results show that MGMs reliably uncover the underlying undirected graph structure, and, when used for classification, their performance is comparable to popular univariate methods (lasso regression and support vector machines). We also show that imposing separate sparsity penalties for edges connecting different types of variables significantly improves edge recovery performance. To choose these sparsity parameters, we propose an efficient model selection method based on an existing method, stability approach to regularization selection (StARS). We call this approach Stable Edge-specific Penalty Selection (StEPS). MGMs produced by StEPS outperform models selected using standard techniques (including AIC, BIC and cross-validation) in edge recovery. In addition, our method uses a heuristic search that is linear in size of the sparsity value search space as opposed to the cubic grid search required by other model selection methods. An MGM learned over mRNA expression and clinical data from the Lung Genomics Research Consortium correctly recovered connections between the diagnosis of obstructive or interstitial lung disease, two diagnostic breathing tests, and cigarette smoking history. Our model also suggested biologically relevant mRNA markers that are linked to these three clinical variables.

2.1 BACKGROUND

Integrating biomedical datasets from different data streams (e.g., -omics, clinical) and of different types (continuous, discrete) is of utmost importance and has become an analysis bottleneck in biomedical research. Ideally, one would like to be able to uncover all direct associations between variables and/or perform feature selection and classification tasks using all data. The first task can reveal disease mechanisms and the second can be used to select variables characteristic of disease status, therapy outcome or any other variable of clinical importance. Graphical models have been used in the past for both of these tasks, but they are often limited to datasets with discrete-only or continuous-only variables. Traditional univariate approaches for feature selection exist as well, but they also often operate on a single data type. In addition, due to the high dimensionality and co-linearity of biological data, markers selected by these standard feature selection algorithms can be unstable and lack biological relevance (Abeel et al. 2010), a problem that has recently been addressed directly (Huang et al. 2014). Many existing models that do integrate different data types make heavy use of prior knowledge (Sedgewick et al. 2013; Wang et al. 2013) and as such are not easily extendable to clinical and other data that are not well studied. As a result, although numerous biomedical data sets exist with genomic, transcriptomic, epigenetic and phenotypic data for each sample, a general framework for integrative analysis of these heterogeneous data is lacking.

In this chapter, we study several strategies for learning the structure of graphical models over mixed data types (discrete and continuous) to produce statistically and biologically meaningful predictive models. We measure the performance of these strategies in synthetic data (via true edge recovery) and biological data (via functional enrichment and performance on classification tasks).

The major contributions of this work are threefold. First, we apply an MGM, proposed by Lee and Hastie (Lee and Hastie 2013), to simulated and biological datasets. These datasets have higher dimensionality and are derived from more complicated network structures than datasets used in previous work with this model. Second, we propose the use of a separate sparsity penalty for each edge type in the MGM, which significantly improves performance. Third, to assist with setting the sparsity parameters we use a heuristic search, StEPS, based on an existing model selection method (StARS) (Liu et al. 2010), that outperforms standard methods.

2.1.1 Prior work

Graphical models are a natural tool for decoding the complex structure of heterogeneous data and allow for integration of many data types. They learn a network of statistical dependencies subject to a joint probability distribution over the data. Mixed graphical models (MGMs) are graphical models learned over a mixture of continuous and discrete features.

A fully specified conditional Gaussian MGM, as characterized by Lauritzen & Wermuth (Lauritzen and Wermuth 1989), would require different continuous distribution parameters for every possible setting of the discrete variables. Restricting ourselves to “homogeneous” models, which use a common covariance matrix for continuous variables independent of the discrete variable values, is therefore necessary to avoid trying to learn a parameter space that is exponential in the number of variables. Similar to pairwise Markov Random Fields over only discrete variables, the main hurdle to the calculation of likelihood in MGMs is calculation of the partition function. This computation is intractable with a large number of discrete variables because it requires summing over all possible discrete variable settings. Two approaches to get around this partition function calculation are: (1) learn separate regressions of each variable given all of the others (Fellinghauer and Bühlmann 2011; Chen et al. 2014; Yang et al. 2014), and (2) maximize a tractable pseudolikelihood function instead of the actual likelihood (Lee and Hastie 2013).

Performing separate regressions is a common approach to the MGM learning problem. This class of methods learns a conditional distribution for each node given the rest. Examples of this strategy include estimation of the sparse inverse covariance matrix of a multivariate Gaussian by Meinshausen and Bühlmann (Meinshausen and Buehlmann 2006), and estimation of mixed variable networks via random forests (Fellinghauer and Bühlmann 2011) or exponential families (Chen et al. 2014; Yang et al. 2014). Alternatively, the pseudolikelihood, proposed by Besag (Besag 1975), is a consistent estimator of the likelihood, and is defined as the product of the conditional distributions of each node given the rest. Both of these approaches thus avoid calculation of the partition function for the joint distribution by substituting the conditional distributions of each node into the optimization problem. Separate regressions offer flexibility and are easily parallelized, but in both the continuous (Friedman et al. 2008) and mixed cases (Lee and Hastie 2013) estimating the parameters by maximizing the likelihood or pseudolikelihood, respectively, has the advantage of better empirical performance. Because of this we chose the focus our efforts on the MGM learning approach via pseudolikelihood, as proposed in Lee and Hastie (Lee and Hastie 2013).

Although Lee and Hastie do not test their algorithm on high-dimensional data, we find that their model is well suited for high-dimensional learning due to their inclusion of a sparsity penalty on the parameters. An important issue that we ran into in our experiments was that the model would often select too many continuous-continuous edges and too few edges involving discrete variables. This is likely a combination of the phenomenon observed in (Chen et al. 2014) where linear regressions have better edge prediction performance than logistic regression between the same nodes and the fact that Lee and Hastie use the same sparsity penalty on all edges regardless of the type(s) of nodes they connect. Lee and Hastie use a weighting scheme to take into account discrete variables with differing numbers of categories, but this does not solve this problem. Therefore, in this paper we introduce a new regularization method for the Lee and Hastie's model that uses a different penalty for each type of edge: continuous-continuous, continuous-discrete, and discrete-discrete. In addition, because this approach creates more parameters for the user to set, we present an edge stability based method for selecting the three sparsity parameters. We call the combination of using separate sparsity penalties with our heuristic search Stable Edge-specific Penalty Selection (StEPS).

2.2 METHODS

2.2.1 Mixed Graphical Models

Lee and Hastie (Lee and Hastie 2013) parameterize a mixed graphical model over p Gaussian variables, x, and q categorical variables, y, as a pairwise Markov Random Field. Here we briefly summarize their model:

$$p(x, y, \Theta) \propto \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\right)$$

In this model β_{st} represents the interaction between two continuous variables, x_s and x_t, $\rho_{sj}(y_j)$ is a vector of parameters that correspond to the interaction between the continuous variable x_s and the categorical variable y_j indexed by the levels (i.e. categories) of the variable y_j, and $\phi_{rj}(y_r, y_j)$, is a matrix of parameters indexed by the levels of the categorical variables y_j, and y_r. In the continuous only case, this model reduces to a multivariate Gaussian model where the β_{st} parameters are entries in the precision matrix. In the categorical only case, this model is the popular pairwise Markov random field with potentials given $\phi_{rj}(y_r, y_j)$; and it could parameterize an Ising model as in the binary-only case, for example. Thus the model serves as a generalization of two popular uni-modal models to the multi-modal regime.

In order to avoid the computational expense of calculating the partition function of this model, Lee and Hastie optimize the negative log pseudolikelihood, which is:

$$\tilde{l}(\Theta|x, y) = -\sum_{s=1}^{p}\log p(x_s|x_{\backslash s}, y; \Theta) - \sum_{r=1}^{q}\log p(y_r|x, y_{\backslash r};\ \Theta)$$

where $x_{\setminus s}$ is short hand for the set of all x_i where $i \neq s$. To ensure a sparse model, $\tilde{l}$ is minimized with respect to a sparsity penalty, λ:

$$\text{minimize}_{\Theta}\ \tilde{l}(\Theta) + \lambda\left(\sum_{t<s}|\beta_{st}| + \sum_{s,j}\|\rho_{sj}\|_2 + \sum_{r<j}\|\phi_{rj}\|_F\right)$$

where Θ is a shorthand for all of the model parameters. The parameter matrices β and ϕ are symmetric, so only half of each matrix is penalized. Lee and Hastie use an accelerated proximal gradient method to solve this optimization problem.

A standard way of handling a categorical variable with L levels is to convert the variable to L-1 indicator variables where the last level is encoded by setting all indicators to zero, this is necessary to ensure the linear independence of variables in the regression problem. This can lead to some ambiguity about the choice of the last level and how to interpret the regression coefficients. In contrast, Lee and Hastie's MGM approach uses L indicator variables (i.e. the elements of $\rho_{sj}(y_j)$ and $\phi_{rj}(y_r, y_j)$) to improve interpretability of the model, and enforces a group penalty to ensure the indicator coefficients sum to zero.

To perform our experiments we adapted the Matlab code provided by Lee and Hastie (available at http://web.stanford.edu/~jdl17/learningmgm.html).

2.2.2 Separate Sparsity Penalties

Our main modification to the Lee and Hastie model itself is that we use different sparsity penalties for the three edge types: edges connecting two continuous nodes (*cc*), edges connecting a continuous and discrete node (*cd*) and edges connecting two discrete nodes (*dd*). With these penalties, the new optimization problem becomes:

$$\text{minimize}_{\Theta}\ \tilde{l}(\Theta) + \lambda_{cc}\sum_{t<s}|\beta_{st}| + \lambda_{cd}\sum_{s,j}\left\|\rho_{sj}\right\|_2 + \lambda_{dd}\sum_{r<j}\left\|\phi_{rj}\right\|_F$$

2.2.3 Methods for Model Selection

K-fold cross-validation (CV) (Efron 1982) splits the data into K subsets and holds each set out once for validation while training on the rest. We use K=5 and average the negative log-pseudolikelihood of the test sets given the trained models.

The Akaike information criterion (AIC) (Akaike 1998) and Bayes information criterion (BIC) (Schwarz 1978) are model selection methods that optimize the likelihood of a model based on a penalty on the size of the model represented by degrees of freedom. To calculate the AIC and BIC, we substitute the pseudolikelihood for the likelihood and we define the degrees of freedom of the learned network as follows.

In the standard lasso problem, the degrees of freedom is simply the number of non-zero regression coefficients (Zou et al. 2007). So, in the continuous case, the degrees of freedom of a graphical lasso model is the number of edges in the learned network. In the mixed case, edges incident to discrete variables have additional coefficients corresponding to each level of the variable. Lee and Hastie's MGM uses group penalties on the edge vectors, ρ, and matrices, ϕ, to ensure that all dimensions sum to zero. So, in the model, an edge between two continuous variables adds one degree of freedom, and edge between a continuous variable and a categorical variable with L levels adds L-1 degrees of freedom, and an edge between two discrete variables with L_i and L_j levels adds $(L_i – 1)(L_j - 1)$ degrees of freedom.

We compare these model selection methods to an oracle selection method. For the oracle model, we select the sparsity parameters that minimize the number of false positives and false negatives between the estimated graph and the true graph. While we do not know the true graph in practice and none of the other methods use the true graph, this method shows us the best possible model selection performance under our experimental conditions.

AIC, BIC, and CV all require calculating the pseudolikelihood from a learned model so to optimize over separate sparsity penalties for each edge type, we perform a cubic grid search of λ_{cc}, λ_{cd}, and λ_{dd} over $\{.64, .32, .16, .08, .04\}$.

2.2.4 Stability for Model Selection

Here we briefly present the StARS procedure (Liu et al. 2010) reformulated in terms of λ rather than $\Lambda = 1/\lambda$ as was originally described. Given a dataset with *n* samples, StARS draws *N* subsamples of size *b* without replacement from the set of $\binom{n}{b}$ possible subsamples. An MGM network is learned for each subsample over a user specified set of values and a single sparsity parameter, λ. The adjacency matrices from these learned models are used to calculate , $\hat{\theta}_{st}(\lambda)$, the fraction of subsample networks that predict an edge from node *s* to node *t*. Using this value we can then calculate edge instability, $\hat{\xi}_{st}(\lambda) = 2\hat{\theta}_{st}(\lambda)(1 - \hat{\theta}_{st}(\lambda))$, which is the empirical probability of any two subsample graphs disagreeing on each possible edge at each value of λ. Liu *et al* define total instability of the graph, $\widehat{D}(\lambda)$, as the average of $\hat{\xi}_{st}(\lambda)$ over all edges: $\widehat{D}(\lambda) = \frac{\sum_{s<t} \hat{\xi}_{st}(\lambda)}{\binom{p+q}{2}}$. Very low values of λ will result in very dense but stable graph, which is not desirable. To avoid this, StARS monotonizes the instability: $\overline{D}(\lambda) = \sup_{\lambda \leq t} \widehat{D}(t)$ and selects $\hat{\lambda} = \inf\{\lambda : \overline{D}(\lambda) \leq \gamma\}$ where γ is a user defined threshold (called β in (Liu et al. 2010)). In other words, starting with a large value of λ that produces an empty graph, we reduce λ until the total instability hits the given threshold.

2.2.5 Stable Edge-specific Penalty Selection (StEPS)

We modified the StARS procedure to accommodate selection of separate λ for each edge type. We now define the total instability over each edge type instead of the entire graph: $\widehat{D}_{cc}(\lambda) = \frac{\sum_{cc} \hat{\xi}_{st}(\lambda)}{\binom{p}{2}}$, $\widehat{D}_{cd}(\lambda) = \frac{\sum_{cd} \hat{\xi}_{st}(\lambda)}{pq}$, $\widehat{D}_{dd}(\lambda) = \frac{\sum_{dd} \hat{\xi}_{st}(\lambda)}{\binom{q}{2}}$. Given these separate estimates of total instability, we then perform the rest of the StARS algorithm for λ_{cc}, λ_{cd}, and λ_{dd} independently. This approach does not require any additional model learning, the only extra computations in this approach compared to the standard, single penalty StARS are the additional averages, which are trivial to calculate. Because the subsample network learning uses the single penalty MGM, this procedure is linear in the size of the parameter search space. Based on the suggestions in (Liu et al. 2010), and the default parameters in the R implementation of StARS (Zhao et al. 2012), we use $N = 20, b = 10\sqrt{n}$, and $\gamma = .05$.

2.2.6 Simulated Network Data

We generated 20 scale-free networks of 100 variables each, based on the framework of Bollobás *et al* (Bollobás et al. 2003) but ignoring edge direction. So, given a number of nodes to connect, we start with an edge between two nodes and the rest of the nodes unconnected, we iteratively add edges until all nodes are connected. At each edge addition, we connect two non-zero degree nodes with probability .3; and we connect a node *i* with degree 0 to a node *j* with non-zero degree with probability 0.7. In each case, the non-zero degree nodes are selected randomly with probability proportional to their degree: $\frac{\text{degree}(j)}{\sum_{k \in V} degree(k)}$.

For each network we simulated two datasets of 500 samples with 50 continuous and 50 categorical variables. Each categorical variable had 4 levels. The parameters in one dataset were set so that discrete-continuous and discrete-discrete edges had approximately linear interactions, while the other dataset did not have this constraint. Each edge, from node *s* to node *t* is given a weight, w_{st}, drawn uniformly from [.5, .8]. For continuous-continuous edges we chose a sign with even probability and set $\beta_{st} = w_{st}$ or $\beta_{st} = -w_{st}$. To ensure the β matrix is positive definite, we set the diagonal elements the largest value of the sum of the absolute value of the edge weights over each node. For continuous-discrete edges, in the linear dataset we set $\rho_{st} = [-w, -.5w, .5w, w]$ (so if the levels of a discrete variable are coded as adjacent integers it can be treated as a continuous variable and will have a linear relationship with neighboring nodes) and in the non-linear data we set $\rho_{st} = perm([-w, -.5w, .5w, w])$, where *perm* is a random permutation of the elements in the vector. For discrete-discrete edges we set the diagonal of $\phi_{rj}(y_r, y_j)$ to w_{st} and the rest to -w_{st}, while in the non-linear data we randomly set one parameter in each column and row to w_{st} and the rest to -$w_{st.}$.

2.2.7 Lung Chronic Disease Data

The Lung Genomics Research Consortium (LGRC) contains multiple genomic datasets and clinical variables for two chronic lung diseases: chronic obstructive pulmonary disease (COPD) and interstitial lung disease (IDL). We used two data types from LGRC: gene expression profiles (15,261 probes) and clinical data for 457 patients (COPD N=215; ILD N=242). To expedite the execution time and avoid sample size problems, we only used the 530 most variant expression probes and 8 clinical variables: age, height, weight, forced expiratory volume in one second (FEV1), forced vital capacity (FVC), gender, cigarette history, and diagnosis (COPD or ILD).

Age, height, weight and the spirometry variables (FEV1 and FVC) were divided into tertiles. Diagnosis was used for classification experiments.

2.2.8 Graph Estimation Performance

Non-zero MGM edge parameters correspond to a prediction of the presence of that edge. For edges with multiple parameters, (i.e. $\rho_{sj}(y_j)$ and $\phi_{rj}(y_r, y_j)$) if any of the parameters are non-zero we predict the edge is present. We use accuracy, precision and recall to evaluate edge recovery in our predicted graphs: precision is the ratio of true edge predictions to all edge predictions; recall is the ratio of true edge predictions to all edges in the true graph; accuracy is the ratio of true predictions to all predictions (in this case true prediction includes the correct predictions of the presence or absence of an edge); and the F1 score is the harmonic mean of precision and recall. In addition we consider the Matthews' correlation coefficient (MCC) (Matthews 1975) which provides a correlation between the presence of edges in the true and predicted graphs. MCC is formulation of Pearson's correlation for two binary variables so values of 1 correspond to perfect agreement between the variables, -1 to all disagreements, and 0 to random guessing. This measure is robust to unbalanced nature of the problem where in the true, sparse graph edge absence is much more frequent than edge presence.

2.2.9 Functional Enrichment and Classification

For evaluation of the performance of various MGMs and other models on real data we used functional enrichment analysis of external databases and classification analysis over specific variables in the network, including disease diagnosis (for clinical datasets).

Gene annotations were retrieved from the Gene Ontology (GO) database (Consortium 2015) and we used the hypergeometric test to determine if sets of selected genes were overrepresented for any of these annotations (i.e. more occurrences of a given annotation were observed than we would expect from randomly selected genes).

Given the parameters learned from training data, $\hat{\Theta}_{train}$, we make predictions on any categorical variable, y_{target}, in a testing dataset given the rest of the variables by selecting the category minimizes the negative log pseudolikelihood of the test data given the trained model:

$$\hat{y}_{target} = argmin_{L_{target}} \tilde{l}\,(\hat{\Theta}_{train}; x_{test}, y_{test\setminus target}, y_{target} = L_{target})$$

We use this approach to predict lung disease diagnosis in a test dataset with an MGM trained with a training dataset.

We used 8-fold cross validation to determine the optimal classification settings of λ for MGM and Lasso, and which kernel to use for support vector machines (SVMs). We used the built-in Matlab implementations of Lasso and SVMs for these experiments.

2.3 RESULTS

2.3.1 Separate Sparsities versus Single Sparsity Parameter

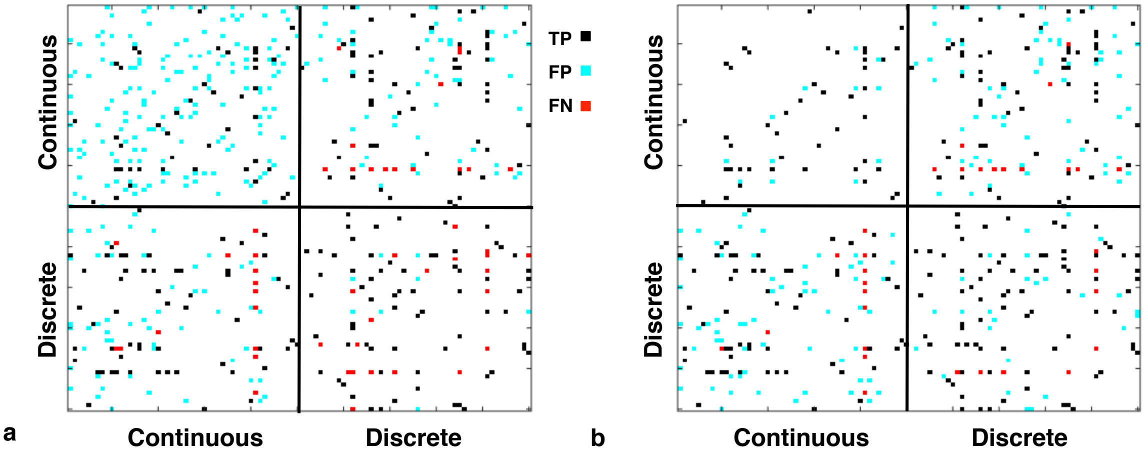


Figure 2.1 Example adjacency matrices predicted by an MGM, with sparsity selected using the oracle. **a**. Single sparsity penalty $\lambda = .19$ **b**. Split sparsity penalties $\lambda_{cc} = .64$, $\lambda_{cd} = .19$, $\lambda_{dd} = .13$

We applied Lee and Hastie's method for learning an MGM to datasets simulated from a scale-free network. Initial experiments found that using a single sparsity penalty for all edge types produced many false positive continuous-continuous edge predictions, while missing many true discrete-discrete edges. We first present an example of this behavior on a single dataset of 500 samples over 50 four-level discrete variables and 50 continuous variables generated from a scale free network structure. **Figure 2.1a** shows the adjacency predictions of the learned MGM compared to the true graph using a λ selected by the oracle to minimize the number of edges present in one graph but not the other. This observation leads us to introduce separate sparsity penalties for each edge type. **Figure 2.1b** shows the adjacencies learned by an MGM with separate sparsity penalties for each edge type. For the sparsity parameters, the oracle searched over a range of 13 values evenly spaced on a log scale from .08 to .64.

Figure 2.2 shows the Matthews correlation of the edge predictions over the range of sparsity parameters, both overall and separated by edge type. For this example dataset, edge recovery of discrete-discrete edges had the highest MCC at $\lambda = .13$ while correlation of recovery of continuous-discrete edges was maximized at $\lambda = .19$ and continuous-continuous edges at $\lambda = .64$.

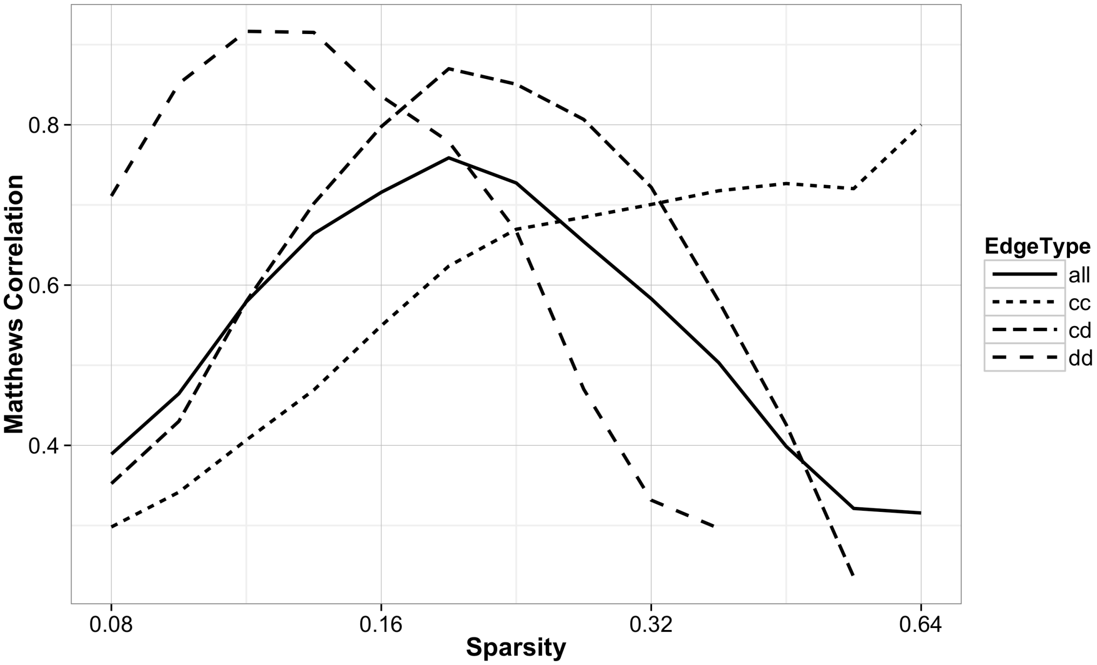


Figure 2.2 Matthews correlation between edge predictions and the true graph versus sparsity for the example dataset from Fig. 2.1. Calculated for each edge type, *cc* for continuous-continuous, *cd* for continuous-discrete, *dd* for discrete-discrete, and over all edge predictions.

Selecting an optimal value for a single λ can be challenging, and the addition of two more sparsity parameters made it necessary to develop an efficient selection strategy. Other methods with multiple sparsity parameters search over a grid of models learned on all possible combinations of the parameters (Zhang and Kim 2014), but for our model the complexity of this selection would be cubic in the number of parameter values tested. Many model selection methods rely on calculating some likelihood over the training data, and it is not clear how to divide up this calculation by edge type. We do expect the presence of edges to remain relatively constant for a given edge sparsity parameter setting, so we extended a recent subsampling technique for model selection, StARS (Liu et al. 2010), to select three edge-type specific sparsity penalties by assuming independence between edge types. This assumption allows for a linear rather than cubic search over possible sparsity parameters. Thus, our method, StEPS, selects three sparsity penalties for Lee and Hastie's MGM learning using a modified StARS approach for subsampling over different edge types.

2.3.2 StEPS Outperforms Other Methods for Model Selection

Table 2.1 Comparison of model selection methods. Mean (and standard error) of classification performance over 20 datasets simulated from scale-free networks. The entry for the method that performs best (excluding the oracle) in each category is bolded. AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; CV: cross-validation; Oracle: best possible prediction performance (maximize accuracy using true graph).

Methods	Precision	Recall	F1-score	Matthews CC	Accuracy
AIC	0.1104 (0.002)	**0.9698 (0.004)**	0.1982 (0.003)	0.2882 (0.003)	0.7952 (0.003)
BIC	0.4588 (0.028)	0.8633 (0.007)	0.5890 (0.025)	0.6098 (0.022)	0.9652 (0.004)
CV	0.1530 (0.003)	0.9694 (0.004)	0.2640 (0.005)	0.3539 (0.004)	0.8587 (0.003)
Oracle	0.9149 (0.015)	0.7868 (0.021)	0.8397 (0.009)	0.8416 (0.008)	0.9923 (0.000)
StARS – 1 λ	0.8988 (0.018)	0.4993 (0.010)	0.6408 (0.011)	0.6632 (0.011)	0.9854 (0.001)
StEPS – 3 λ	**0.9159 (0.014)**	0.6720 (0.009)	**0.7731 (0.007)**	**0.7787 (0.007)**	**0.9897 (0.000)**

Table 2.1 Summarizes graph prediction results for MGMs trained using sparsity penalties chosen with different model selection procedures over the 20 simulated non-linear datasets. Oracle, AIC, BIC, and CV evaluated models over a three dimensional grid of all possible combinations of $\lambda_{cc}, \lambda_{cd}, \lambda_{dd} \in \{.64, .32, .16, .08, .04\}$. For StARS, models were trained using a single sparsity penalty over the same range of values, and then either a single λ was selected based on the average instability over all edges or $\lambda_{cc}, \lambda_{cd}$ and λ_{dd} were selected based on the average instability of each edge type.

Our results show that AIC, BIC and CV produce overly dense models in the high-dimensional setting. Even when restricted to the single sparsity model, StARS significantly outperforms these traditional model selection methods. These results agree with what Liu *et al* observed in their model selection experiments with the graphical lasso (Liu et al. 2010). In addition, our modification of StARS with separate sparsities outperforms StARS with a single sparsity. Neither StARS model selection with 3 penalties nor the oracle model selection output a model where all three sparsities were equal in any of these experiments. Both methods always set $\lambda_{dd} = .16$ while the other parameters were always in the set $\{.64, .32, .16\}$. These results confirm the effectiveness of separating the MGM sparsity penalty into three λ values.

The original StARS procedure uses a subsampled dataset to make final edge predictions because the instability calculations are made on subsamples. We found, however, that in all cases the final edge prediction performance is higher if we use all samples compared to predictions from a model using a subsampled dataset. This improved performance is observed for all three metrics: accuracy, MCC, and F1. So, for all results presented below we used all samples to learn the MGM and make edge predictions with StARS selected sparsities.

It is important to note that because our method, StEPS, selects each sparsity parameter independently, it incorrectly assumes that the instability of each edge type is independent of the parameters of the other edge types. Without this assumption, we would have to perform stability experiments on all combinations of the sparsity parameters. To test if this assumption is reducing the edge recovery performance of StEPS, we ran StARS on the non-linear datasets using all 125 possible settings of $\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}} \in \{.64,\ .45,\ .32,\ .23,\ .16\}$. This search space was chosen because all of the values selected by the Oracle or either of the other StARS methods fell in the set $\{.64, .32, .16\}$ and additional intermediate values were needed to compare the relative performance of these methods. Although StARS occasionally selected values of .64 and .16 which are on the boundary of this test range, we did not include higher or lower values because .32 was selected most of the time, and the cubic growth made it expensive to search over more than 5 penalty values. This experiment posed a new problem of how to monotonize and select the total instability over three dimensions rather than one. In addition, this experiment showed that the number of predicted edges in the graph does not always increase when one of the λ parameters decreases, even when the other two are held constant. We found that simply choosing the model with monotonized total instability closest to the user-specified γ threshold produced poor results. Taking into account the number of edges predicted across all subsamples for each parameter setting, as described below, was essential to producing usable results.

We first looked at the total instability of the whole graph with all edge types pooled together, $\widehat{D}_{all}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}})$. We monotonized this 3-dimensional matrix across each dimension: $\overline{D}_{all}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}) = \ \sup_{\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}} \leq t_1, t_2, t_3} \widehat{D}(t_1, t_2, t_3)$ and selected the setting of $\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}$ that produced subsampled networks with the most edges such that $\overline{D}_{all}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}) \leq \gamma =$ .05. Surprisingly, this approach performed worse than StEPS on all measures. MCC, for example, was significantly worse (mean of .845 for the heuristic versus .718 for this method, t-test p = 1.4e-4). We found that the networks produced by this method were too dense in the continuous-continuous edges and too sparse in continuous-discrete edges (results not shown). This is the result of averaging the instability of all edge types: the selected models were too stable for some edge types and too unstable for other types. To fix this, we separated the instability as before into $\widehat{D}_{cc}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}})$, $\widehat{D}_{cd}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}})$ and $\widehat{D}_{dd}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}})$, and monotonized as before. Then we choose $\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}$ that produced networks with the most edges such that $max(\overline{D}_{cc}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}), \overline{D}_{cd}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}}), \overline{D}_{dd}(\lambda_{\mathrm{cc}}, \lambda_{\mathrm{cd}}, \lambda_{\mathrm{dd}})) \leq \gamma = .05$. On 17 of the 20 datasets tested, this approach selected the same sparsity parameters as our proposed linear parameter search method. For the three runs where the two methods selected different parameters, the cubic search made better choices than the heuristic. Averaging over all runs the cubic search performed better than the heuristic but these results are not significant (e.g., mean MCC for the cubic search was .850 versus .845 for StEPS, p = 0.56). These results indicate that the independence assumption made by our heuristic is reasonable and that StEPS performs only slightly worse than a more theoretically sound cubic search while requiring much less computation.

2.3.3 Comparison to SCGGM

An important potential application of MGMs is in identifying expression quantitative trait loci (eQTLs) based on the predicted dependencies between single nucleotide polymorphisms (SNPs) and mRNA expression. The sparse conditional Gaussian graphical model (SCGGM) (Zhang and Kim 2014) is a method that addresses this problem specifically. Like many methods for finding eQTLs, the SCGGM assumes a linear relationship between the number of variant alleles and the mRNA expression level. Thus, the SCGGM is not technically a mixed graphical model because it treats the SNP allele counts as continuous variables. Another difference is that SCGGM does not predict discrete-discrete edges, which is also common among methods for finding eQTLs. Like StEPS, SCGGM also adopts a strategy of using a separate sparsity penalty for each edge type. SCGGM uses cross-validation to search over a two dimensional grid of parameter values in order to optimize prediction of continuous values given the discrete values.

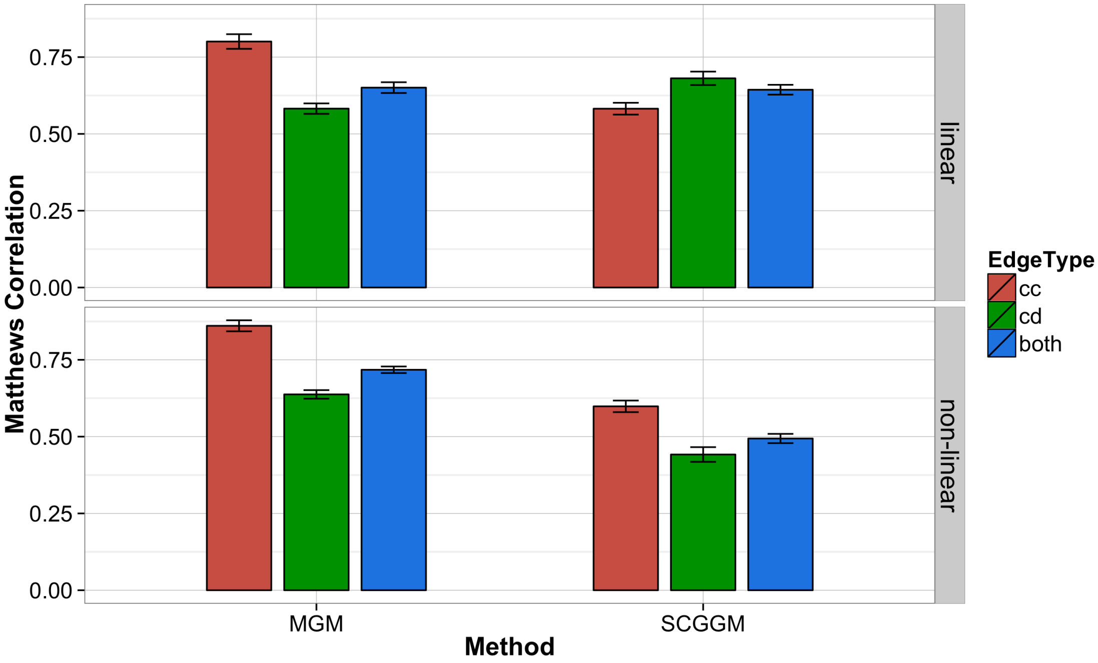


Figure 2.3 Comparison of edge recovery performance of MGM and SCGGM on continuous-continuous (cc), continuous-discrete (cd) and both edge types. Matthews correlation is averaged over 20 simulated datasets with linear continuous-discrete interactions and 20 datasets with non-linear interactions with error bars ± one standard error. Sparsity parameters for both methods selected by StEPS. Figure created in collaboration with Ivy Shi.

First, we examined how our stability method can be used in SCGGM parameter selection instead of cross validation on our synthetic data and we found that StEPS resulted in significantly higher MCC ($p < .01$) for recovery of both continuous-continuous and continuous-discrete edge types. To perform a comparison between MGM and SCGGM edge predictions we used two sets of 20 mixed datasets generated from the same set of 20 scale-free networks but with parameters that resulted in either linear or non-linear interactions between discrete and continuous variables. **Figure 2.3** shows the results of this experiment with StEPS selected sparsity parameters. As expected, MGM learning performed similarly on the linear and non-linear datasets because it does not assume linearity. The SCGGM had similar performance on continuous-continuous edge recovery with both datasets, but significantly worse performance on continuous-discrete edge recovery in the data with non-linear *cd* interactions, which resulted in worse overall performance in that setting.

For these tests we found that when allowing the selection of (different) edge type specific sparsity penalties, SCGGM chose the same penalty for the *cc* and *cd* edges in 36 out of the 40 datasets; and StEPS chose the same penalty for the *cc* and *cd* edges in 38 out of the 40 datasets, but a different *dd* penalty in all 40 cases.

2.3.4 Performance of MGM on Lung Disease Data

It is difficult to evaluate the edge recovery performance of MGM in real clinical datasets since the ground truth (all associations between variables) is not generally known. Alternatively, we evaluate MGM performance indirectly, by (1) recovering the small number of interactions that are known, (2) using external datasets (GO categories) to see if connected genes have similar function, (3) performing classification on a target variable in the network (disease diagnosis).

We applied our MGM learning approach to the LGRC biomedical data (described above). On this data StEPS selected the same value of $\lambda_{cc}, \lambda_{cd} = .2$ for an average instability threshold of $\gamma = .05$ and $\lambda_{cc}, \lambda_{cd} = .1$ for $\gamma = .1$. The selection of λ_{dd} proved more problematic. Even with $\gamma = .1$, λ_{dd} was selected to be so high that only one edge was selected (FEV1-FVC). This issue is likely caused by the fact that there are only 28 possible edges between the 8 clinical variables, and we expect that many of these variables are connected. Because of this and the fact that the experiments we perform below depend more on the continuous-discrete edges, we set all three penalties to the same value for our parameter searches in this section.

2.3.5 Recovering Known Interactions

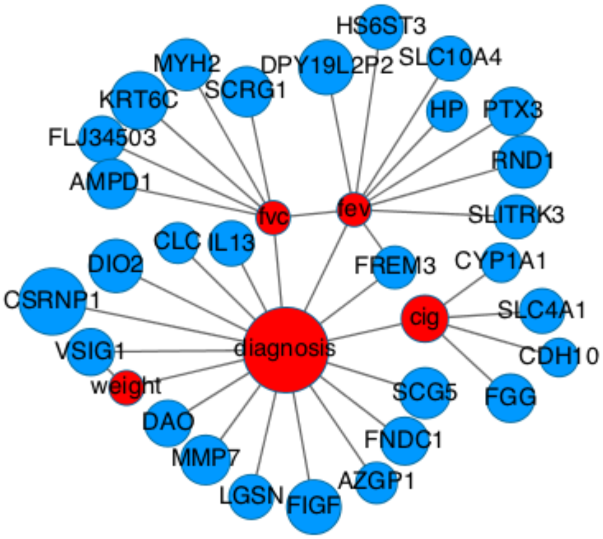


Figure 2.4 Learned sub-network of gene expression and clinical features connected to lung disease diagnosis, lung tests and cigarette smoking. Nodes are colored by data type, blue for gene expression, red for clinical variables. Edges were filtered by weight with a threshold of .05. Node size is proportional to the diagonal of the β matrix for continuous variables and $\|\phi_{yy}\|_F$ for each categorical variable, y.

Figure 2.4 shows part of the network learned over the lung (LGRC) dataset with $\lambda_{cc}, \lambda_{cd}, \lambda_{dd} =$ .1. We only show the nodes adjacent to the clinical variables most relevant for lung disease: diagnosis, spirometry tests and cigarette smoking. This model found a very strong connection between the FEV1 and FVC variables. A number of relevant gene expression variables are linked to diagnosis in this network. IL13 is part of the family of interleukin signaling molecules, which are associated with inflammatory response to tissue damage, and COPD is an inflammatory disease. We also see a link between diagnosis and MMP7, a previously discovered biomarker for idiopathic pulmonary fibrosis which is categorized as ILD (Rosas et al. 2008). A link between diagnosis and AZGP1, another previously studied marker for COPD (Mazur et al. 2012), was also recovered. FGG and CYP1A1 were found to be linked to cigarette smoking history. CYP1A1 is known to convert polycyclic aromatic hydrocarbons, found in cigarette smoke, into carcinogens (Walsh et al. 2013), and FGG codes for fibrinogen, a marker for inflammation, which is positively correlated with risk of mortality and COPD severity (Mannino et al. 2012).

2.3.6 Recovering Functional Relationships

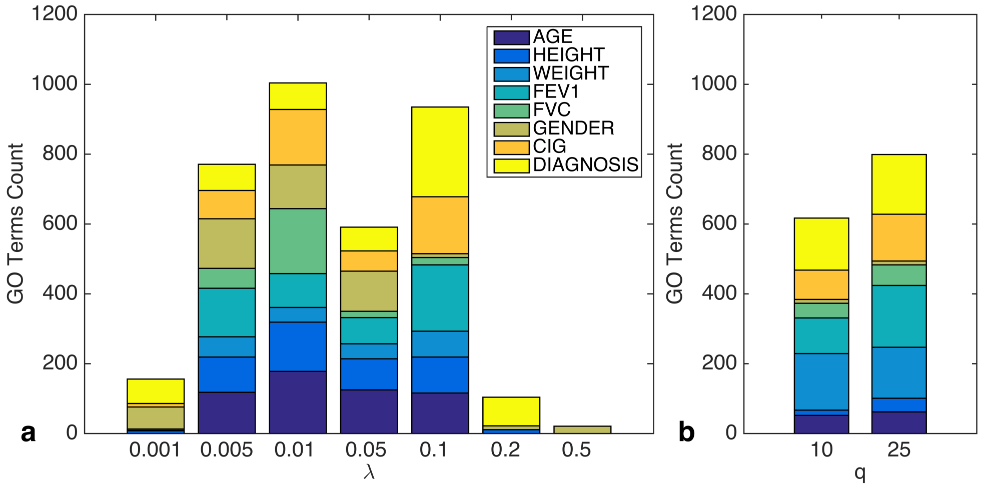


Figure 2.5 Counts of GO terms with uncorrected p < .05 for groups of genes with expression variables linked to each discrete clinical variable in: **a.** MGM networks at different values of λ and **b.** qp-graphs at different values of q. Edge thresholds for qp-graphs were chosen to select similar numbers of connected genes to an MGM network with λ = .1

We also compared the functional relevance of MGM networks learned with StEPS and those learned by qp-graphs (Tur and Castelo 2012), another method for learning networks over mixed data. Like SCGGM, qp-graphs do not attempt to learn edges between two discrete variables, but qp-graphs do not make a linearity assumption about the discrete variables. To assess the biological relevance of networks learned at different levels of sparsity, we performed enrichment analyses on genes with expression variables linked to each clinical variable. For each group of genes linked to each clinical variable we counted GO terms with an uncorrected enrichment p < .05 (via Fisher's exact test). These counts are shown in **Figure 2.5.** Since each clinical variable represents a phenotype, we would hope that genes linked to those variables share similar biological function as measured by functional enrichment. We would like to choose a value of λ that maximizes the number of enriched GO terms.

The setting of $\lambda = .1$ recovers the most annotations for diagnosis, FEV1 and FVC, and also corresponds to an instability threshold of $\gamma = .1$. qp-graphs output a “non-rejection rate” for each edge, which corresponds to the number of different conditional independence tests that rejected the presence of each edge. To predict edges, this output needed to be thresholded, so we chose thresholds that produced similar numbers of edge predictions to $\lambda = .1$. While qp-graphs perform comparably well to MGMs in this test, we found the learning procedure to be very computationally expensive. On a quad-core laptop, learning a qp-graph with q = 25 took over 3 hours (running time scales linearly with q) while learning an MGM took 4.4 minutes on average when the iteration limit was reached.

2.3.7 Evaluating MGM in Classification Tasks

We also evaluated MGMs for predicting the status of a given target variable. We chose the lung disease diagnosis as a clinically relevant target variable. The MGM was compared to SVM and lasso. We optimized the settings of SVM, lasso and our mixed models to maximize the 8-fold cross-validation accuracy of predicting lung disease diagnosis using the 530 expression variables and 7 clinical variables. For SVM, we found that a linear kernel worked best on this data. For lasso and MGM, the parameter scan found that λ = 0.05 maximized this accuracy. **Figure 2.6a** shows a comparison between the optimized classification accuracies of these three methods. For MGM classification, we expected similar results to lasso because the conditional distribution of a discrete variable in the mixed model reduces to a (multivariate) logistic regression. It is interesting to see that the generative MGMs are not significantly different from discriminative lasso and SVM models in this experiment. While $\lambda = .05$ maximized the cross-validation accuracy for MGMs, **Figure 2.6b** shows that the StARS selected sparsity values of λ =.1 and .2 do not perform significantly worse than $\lambda = .05$. In addition, we ran experiments using StEPS with settings of $[\lambda_{cc}, \lambda_{cd}, \lambda_{dd}] = [.1, .1, .2]$ and $[\lambda_{cc}, \lambda_{cd}, \lambda_{dd}] = [.2, .2, .3]$ which correspond to instability thresholds of $\gamma = .1$ and $\gamma = .05$, respectively, and found that these changes did not significantly alter classification performance.

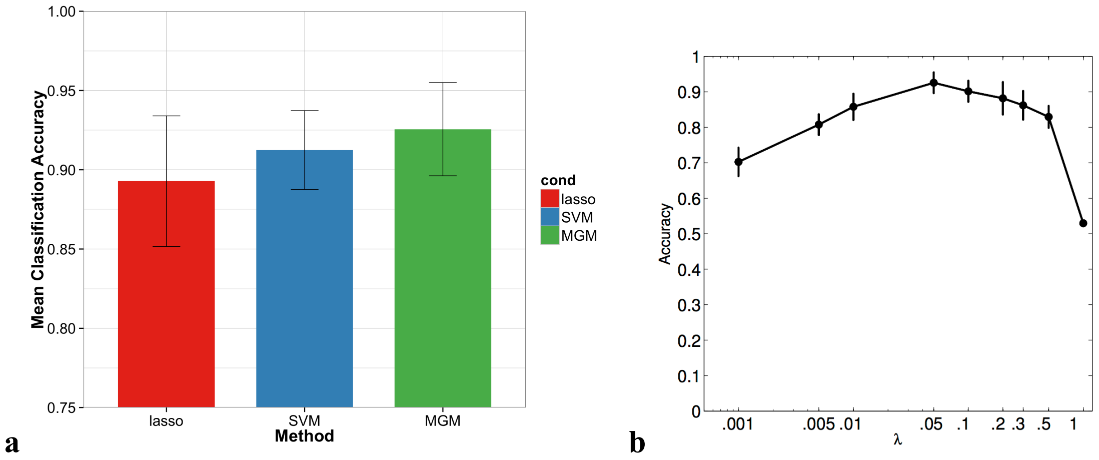


Figure 2.6 a. 8-fold cross validation accuracies for COPD/ILD classification using different methods **b.** Regularization effects on classification accuracy (with error bars of 1 standard deviation).

2.4 DISCUSSION

Learning graphical models over variables of mixed type is very important for biomedical research. The most widely used types of genomic data include continuous (gene expression, methylation, and protein data) and discrete (polymorphism and mutation data) variables. Similarly, clinical variables can be either continuous or discrete (numerical, categorical, boolean). We are interested in learning graphical models from these heterogeneous data to identify significant interactions between variables and uncover important biological pathways.

As an added advantage, a learned network and joint probability can be used to ask an arbitrary number of classification questions over the data without the need for retraining each time (Tsamardinos et al. 2003). These models would be broadly applicable to biological network inference, biomarker selection and patient stratification. Although calculating the MGM requires certain distributional assumptions about the data, the two distributions that make up the model in this work, a multivariate Gaussian for the continuous variables and a pairwise Markov random field for discrete variables, are well studied and have been successfully applied to many types of data. Additionally, using Gaussian copula methods (Liu et al. 2009) in conjunction with MGM learning would allow users to relax the normality assumption for the continuous data.

Our simulation study strongly supports the need for separate sparsity penalty for each edge type when learning an MGM. In addition we show the effectiveness of our extension of the StARS procedure, StEPS, to select these penalty terms. By using instability estimates from the single sparsity parameter model to select parameters for the three parameter model, we are making the assumption that each edge type set is independent from the others. We showed that StEPS performance under this independence assumption is comparable to a stability selection procedure that does not make this assumption. The pay off for StEPS is that we can select three parameters in linear time (over the number of parameter values searched) rather than cubic time. StEPS is a general methodology, which can be applied to a variety of mixed distribution settings, and will be especially useful in problems with many different edge types.

One could argue that StEPS substitutes an arbitrary setting of λ for an arbitrary setting of the instability threshold, γ. As Liu *et al.* point out, γ has a more intuitive meaning than λ, and we feel that setting this threshold compares to the common practice of setting an arbitrary significance threshold for rejecting the null hypothesis. Our results from applying StEPS to

MGMs highlights the fact that the same setting of γ applies well to all edge types while different, edge type specific settings of λ are required for accurate edge recovery. Although it is possible to set the sparsity parameters based on some prior knowledge of the expected number of edges in the network, the data driven methods we present here allow for wide application of MGMs to domains where such knowledge is not available.

Furthermore, we show that our approach to MGM learning is competitive with a state-of-the-art eQTL learning method, SCGGMs. Although SCGGMs can be learned more quickly than our MGMs due to the fact that they treat all variables as continuous, we showed that MGMs have a clear advantage when the discrete variables have non-linear relationships with the continuous variables. The assumption of linearity is common in eQTL learning and it makes sense in the haploid yeast datasets (e.g. (Brem et al. 2005)) used in the SCGGM study. In more complex organisms, however, an MGM that can handle non-linear interactions may be necessary.

While we had difficulty setting the discrete-discrete edge penalty in the lung dataset, we were still able to show the utility of MGM based analysis on biological data. Also, results from our classification experiment were robust to variation in the setting of this parameter. We do not expect MGMs to perform better than standard classification methods because they minimize the prediction error of the classification problem directly while the pseudolikelihood optimization in the MGM must take into account the relationships between all of the variables, not just between the target variable and the rest. Our results show, however, that the MGM-based classification is comparable to standard methods while offering two key advantages: (1) the same trained MGM can be used to make predictions about any variable without additional learning, and (2) the graph structure allows us to look at the second neighbors of the target variable and beyond for possible functional significance.

2.5 CONCLUSIONS

Mixed graphical models are becoming popular in the statistics and machine learning literature, and there is a lot of potential for their application to high dimensional biological data. We have broached that potential in this study. We showed that MGMs can accurately learn undirected graphical models over a mixture of discrete and continuous variables in a high dimensional setting. In addition, we showed that using a separate sparsity parameter for each edge type in a graph can significantly improve edge recovery performance. These separate parameters can account for the differences in both the difficulty of learning such an edge and differences in the sparsity of edge types in the true graph. Finally, we showed that stability based methods are well suited for model selection in this setting and that our method StEPS allow us to perform a search over the sparsity penalties in linear time.

3.0 CAUSAL SEARCH WITH MIXED VARIABLES

In this chapter we shift from searching for undirected graph structures to directed structures. Directed graphical models are closely related to undirected models in that they both encode conditional probability relationships between variables, but they differ in their assumptions and ability to represent certain structures. The power of directed models come with their ability to encode causation, which is especially desirable in the study of biological systems. Of course, inferring causation from observational data is a difficult task, but these predictions can guide interventional studies to better understand these systems. We will show in this chapter that an undirected model learned over data that is generated from a directed graph is a superset of the true graph (specifically, a moralization of the true graph). This fact suggests the strategy that we adopt in this chapter of first learning an undirected graph using the methods we presented in chapter 2, and then filtering and orienting our predicted edges using a directed graph search algorithm.

3.1 CHAPTER SUMMARY

Graphical causal models are an important tool for biomedical research because they can simultaneously represent both the influence pathways and complex, multivariate probability distributions that are useful for modeling biological data. Learned models can be used for classification, biomarker selection, and functional analysis. These models are generally designed to handle only one type of data, however, and this limits their applicability to a large class of biological datasets with both continuous and discrete variables. To address this issue, we develop new methods that modify and combine existing methods for finding undirected graphs with methods for finding directed graphs. These hybrid methods are not only faster, but also perform better than the directed graph estimation methods alone for a variety of parameter settings and data set sizes. When applied to breast cancer data, our methods recovered relevant connections between gene expression variables and clinical variables for hormone receptor status and subtype label.

3.2 INTRODUCTION

3.2.1 Background

Commonly studied biological data are multi-modal: they include both discrete variables (polymorphisms, mutations) and continuous variables (gene expression, methylation, and protein data). The sizes of relevant databases containing these data have become enormous. In many problems, the number of potentially relevant variables and cellular pathways demands the aid of fast, accurate, computerized search methods for identifying causal relations. These methods produce network models, represented as directed graphs or collections of directed graphs, that can provide guidance to experimentalists and clinicians and are useful for classification and prediction of clinical outcomes. A number of such methods have been developed in the past, but they typically assume (for proof of asymptotic correctness) that all variables are of the same distribution type—categorical (multinomial), Gaussian, conditional Gaussian, or linear non-Gaussian—but not of the mixed types characteristic of biomedical data. In this paper, we test existing and develop new methods to learn directed graphs over mixed data types.

Regarding existing methods, we test PC-stable and CPC-stable (Colombo and Maathuis 2014). PC-stable is a modification of PC (Spirtes and Glymour 1991), the oldest correct algorithm for searching for directed acyclic graphs when there are no feedback relations and no unrecorded common causes and sampling is independent and identically distributed. PC-stable allows for parallelization and produces graph estimates that are independent of the ordering of the input variables. PC-stable relies on conditional independence tests, which can easily be varied according to the distributions of the variables. Similarly CPC-stable is the order independent variant of Conservative PC (CPC) (Ramsey et al. 2006), which modifies the edge orientation procedure of PC to make it robust to ambiguous conditional independence test results.

The main idea behind our two new algorithms is to first learn the undirected graph over mixed data types and then prune-and-orient this graph using methods derived from existing algorithms for directed graph learning. Undirected graphs are found using a modified version of the method of (Lee and Hastie 2013) (which we call MGM for Mixed Graphical Model). For the prune-and-orient step we use the strategies implemented in PC-stable (MGM-PCS) and CPC-stable (MGM-CPCS). The directed search algorithms rely on testing the conditional independence of pairs of nodes using larger and larger conditioning sets. The pairwise Markov property of undirected graphical models tells us that the absence of an edge in an undirected model implies that the two nodes incident to that absent edge are conditionally independent given all other variables. So, another way to think of our approach is that by learning an undirected graphical model, we are performing pairwise independence tests conditioned on all other nodes in the network. Edge orientation is only implied by absence from the conditioning set, so because all nodes are in the conditioning sets for the undirected search step, this procedure should not lose information needed by the directed search step to make edge direction predictions. In addition to our newly proposed algorithms, this chapter contributes an in-depth study of constraint-based causal learning algorithms on high-dimensional mixed data, which is a currently underdeveloped area in the causal learning literature.

3.2.2 Graphical Structures

Graphical models represent families of probability distributions restricted by conditional independence relations between their variables. Undirected graphical models represent distributions in which a variable, X, is independent of all other variables in a system given the set of variables directly connected to X by an edge (i.e., variables adjacent to X), and any two variables connected by an undirected path are associated. Directed graphical models often restrict the network structures to directed acyclic graphs (DAGs), and represent distributions in which a variable X is independent of all other variables conditional on its Markov Blanket: the parents of X (direct causes of X), the children (direct effects) of X, and the parents of the children of X. Two variables are dependent or conditionally dependent if they are represented by a d-connection relation (Geiger et al. 1990). The essential difference is that for an undirected graph a structure $X - Y - Z$ represents that X and Z are associated but X is independent of Z conditional on Y. As a shorthand for this last relationship we write X ⊥ Z | Y. For a DAG with the same adjacencies, the independence relations depend on the orientations of the edges: if $X \rightarrow Y \leftarrow Z$ (this is called a *v-structure*) then X and Z are independent but they become dependent when conditioned on Y; the independence relations for other orientations of the edges in the example are the same as for undirected graphs. When the distribution accords with the structure $X \rightarrow Y \leftarrow Z$, an undirected graph search will return an undirected graph in which X and Z are connected by an edge (this is called a *moralized graph*). These basic properties suggest a search strategy that finds an undirected moralized graph, prunes the edges introduced by the v-structures and directs the remaining edges Here we consider using the Lee and Hastie algorithm for undirected graph search and then pruning and orienting edges by PC-stable or CPC-stable. We contrast this strategy with using the original PC-stable or CPC-stable procedures directly on mixed data types.

3.2.3 Related Work

Recently, learning a sparse undirected graph structure over multi-modal datasets has attracted attention (Bøttcher 2001; Romero et al. 2006; Tur and Castelo 2011; Cheng et al. 2013; Fellinghauer et al. 2013; Lee and Hastie 2013; Chen et al. 2014; Yang et al. 2014). There is publically available software for several of these methods (Tur and Castelo 2011; Fellinghauer et al. 2013; Lee and Hastie 2013). The Tur and Castelo method is not able to learn connections between categorical variables, so their approach is appropriate for the study expression quantitative trait loci (eQTLs). But, their model does not allow for analysis of downstream discrete clinical variables, for example. A number of proposals suggest a nodewise regression approach for learning networks over a variety of distributions of continuous and discrete variables (Cheng et al. 2013; Fellinghauer et al. 2013; Chen et al. 2014), Lee and Hastie (Lee and Hastie 2013) propose optimizing the pseudolikelihood of a mixed distribution over Gaussian and categorical variables. We developed our algorithms using Lee and Hastie's method as a starting point, both because we will only look at Gaussian and categorical variables in this study and because their approach involves learning fewer parameters than with nodewise regression methods.

The idea of using an undirected method to estimate a superstructure of the true graph, and then restricting the search space of a directed search algorithm to the superstructure has previously been studied for continuous, possibly non-Gaussian data with linear interactions between nodes (Loh and Bühlmann 2014). Like our proposed method, Loh and Bühlmann first find an undirected graph which serves as an estimate of the moralization of the true graph, and then use this undirected graph as an estimate for a directed search method. The two primary differences between this study and our proposals are that Loh and Bühlmann only look at continuous data in their study, and that the directed search is a score-based method while we focus on constraint-based directed search methods here.

Adapting score-based methods to mixed data is a challenging problem that we are very interested in. The concept behind score-based methods is to efficiently search over the space of DAGs to find the structure that has the best score given the data. In general, these scores take advantage of the fact that joint probability distributions represented by DAGs are factorizable, so adding or subtracting edges from the estimated graph only require re-calculating scores of the incident nodes. Scores are usually related to likelihood calculations, for example, the Bayesian information criterion (BIC) is commonly used for continuous data and is calculated by penalizing the log-likelihood for the degrees of freedom and sample size. The challenge is to find a mixed score that is factorizable and efficient to compute. Preliminary experiments showed that a mixed BIC score based on Lee and Hastie's pseudolikelihood presented in chapter 2 could recover small graphs of ~10 variables, but was too computationally expensive to use on larger graphs. This is an open area of research, but should a useable score for mixed data be developed, these methods can take advantage of an initial undirected graph by using it to restrict the search space of the score-based search algorithm.

3.3 MATERIALS AND METHODS

3.3.1 Simulated Data

We simulated data from 2 sizes of networks with 50 directed graph structures each. Our low dimensional (LD) datasets consisted of 500 samples drawn from a network structure of 50 variables, 25 Gaussian and 25 3-level categorical. The high dimensional (HD) datasets consisted of 100 samples drawn from a network structure of 200 variables, 100 Gaussian and 100 3-level categorical. The structures are sampled uniformly from the space of all directed acyclic graphs (DAGs) with maximum node degree of 10 and a maximum of average node degree of 2.

The relationships between variables are set up in a similar fashion to (Lee and Hastie 2013). Here, for an edge $X \rightarrow Y$ we refer to X as the parent and Y as the child. Parents of the Gaussian variables contribute linearly to the mean of each child; the value of continuous parents is multiplied by an edge parameter and the value of discrete parents is associated with an edge parameter where a separate edge parameter is specified for each category of the discrete variable. Parents of discrete variables contribute log-linearly to the probabilities of each category, with separate parameters for each category of the child variable. With this set up, each edge connecting two continuous variables (*cc*) depends on 1 edge parameter, each edge connecting a continuous and a discrete variable (*cd*) depends on a vector of 3 parameters and edges connecting two discrete variables (*dd*) depend on a 3 by 3 matrix of 9 edge parameters. In order to ensure identifiability, the *cd* parameter vector, and the rows of the *dd* parameter matrix are constrained to sum to 0 leaving these edges with 2 and 6 degrees of freedom, respectively. Edge weights were drawn uniformly from the union of the regions [-1.5, -1] and [1, 1.5]. For *cc* edges the parameter is equal to the weight; for *cd* edge parameters we draw a vector three values uniformly from [0,1] and shift and scale the values so they sum to zero and the largest parameter is equal to the edge weight; for *dd* edge parameters we draw one vector of three values as with *cd* edges and set the rows of the matrix as the three permutations of this vector. Depending on the graph structure, there may be covariance between parents of a node, but since cycles are not allowed, this will take the form of a feed forward loop.

To generate data from these distributions we used TETRAD (version 5.3.0, https://github.com/cmu-phil/tetrad), a Java package for causal modeling that uses linear or non-linear structural equation models (SEMs) to generate data from network distributions. Our fork of TETRAD can be found at https://github.com/ajsedgewick/tetrad/. In the continuous case, zero-mean, Gaussian error terms with standard deviation uniformly drawn from the interval [1, 2], are drawn for every variable and then the variable means are resolved. In DAGs this resolution is trivial as we can start from root nodes with no parents and propagate downwards. To make this process accommodate categorical distributions, we use a uniform draw over [0, 1] as an error term for each discrete variable and this term is used to determine the value of the variable given the probabilities of each category. In generating simulated models, these probabilities that are then updated in the same way as are the means of the continuous variables. This approach ensures convergence of each discrete variable for each sample.

3.3.2 Biological Data

The breast cancer dataset was obtained from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas 2012). This data (BRCA, n=448 samples) included RNA-seq data normalized with RSEM for 20530 transcripts, and clinical variables (PAM50 subtypes, Progesterone, Estrogen and HER2 receptor status, and tumor and node stage codes). We used only the 500 genes with the most variant expression across all samples.

3.3.3 Undirected Graph Search Algorithms

We use Lee and Hastie's model (Lee and Hastie 2013) as the basis for both algorithms and we refer to the undirected graph produced by this method as an MGM. This model has a form similar to a pairwise Markov Random Field (MRF) and learns the undirected graph over mixed data by maximizing a penalized pseudo-likelihood over all mixed type variables. Since this objective function is convex but not smooth, Lee and Hastie use proximal gradient methods implemented in a Matlab library called TFOCS (Becker et al. 2011) for optimization. We implemented the MGM model and the accelerated proximal gradient method (Nesterov's 1983 method as described by (Becker et al. 2011)) in Java and incorporated it into the TETRAD project.

To improve speed, we made an important change to Lee and Hastie's optimization scheme: instead of waiting for the penalized pseudo-likelihood to converge, we keep track of edge changes between iterations of the accelerated proximal gradient method and we terminate the search when three iterations in a row have the same graph structure.

In our experiments with synthetic data we learned MGM graphs across a range of edge sparsity penalties for the Lee and Hastie algorithm, 7 values evenly spaced on the $\log_2$ scale over the range $.05 \leq \lambda \leq .4$. For the HD data we add two values to extend this range to $\lambda \leq .8$. For simplicity, in these experiments we use the same λ edges connecting different types of variables.

3.3.4 Directed Graph Search Algorithms

We compared two popular causal discovery methods, PC-stable and CPC-stable, implemented in TETRAD, with our proposed hybrid methods MGM-PCS and MGM-CPCS. In addition, as a proof of concept, we ran experiments with Complementary Pairs Stability Selection (CPSS) a stability based method that ensures the predicted network only includes a small proportion of low-probability edges that are unlikely to generalize well to small changes in the data. We present results from all algorithms using a range of values for the conditional independence test threshold: $\alpha \in \{.001, .01, .05, .1\}$.

PC-stable (Colombo and Maathuis 2014) is a graph search algorithm that is a modification of **PC** (Spirtes and Glymour 1991). PC assumes that the underlying graph is acyclic with no latent (unmeasured) variables. The PC algorithm and its descendants depend on conditional independence decisions that are made by a user-specified test and significance level, α (described below). PC starts with a complete graph and in step 1 it sequentially tests all edges for independence given conditioning sets of increasing size. Starting with the empty set, these conditioning sets are subsequently made up of every set (of the given size) of common neighbors of the two nodes incident to the edge being tested. Edges that are found to be conditionally independent are immediately removed and not considered in future tests. When an edge is removed, the conditioning set that lead to the independence decision is saved. Step 2 directs edges based on the fact that common neighbors of nodes incident to a removed edge that are not in the conditioning set must be in a v-structure. It is possible that two implied v-structures will induce conflicting edge directions. In the TETRAD implementation of PC-stable this direction conflict results in a bi-directed edge: $X \leftrightarrow Y$. Step 3 further directs edges based on a set of rules that ensure the directions will not induce any cycles or new v-structures (Spirtes et al. 2000). PC-stable modifies PC by waiting to update the edge removals in phase 1 until all tests for a given conditioning set size are completed. This leads to an output that is independent of variable ordering and allows for parallelization of the independence tests.

CPC-stable (Colombo and Maathuis 2014) is the variable order independent variant of **Conservative PC** (Ramsey et al. 2006) which revises step 2 of PC, described above to perform conditional independence tests with all possible conditioning sets between two nodes, A and C, that have had an edge between them removed. The conditioning sets are determined by taking subsets of neighbors of the two nodes found in the skeleton graph returned by step 1 of PC, described above. For any node, B that is incident to both A and C, the v-structure $A \rightarrow B \leftarrow C$ is only predicted if B is not in any separating set S such that A ⊥ C | S. Otherwise no direction is predicted from this triplet of nodes. If B participates in some sets that result in the conditional independence of A and C and some that result in a conditional dependence, the ambiguity is recorded. Since the change to the PC algorithm takes place after adjacency has been determined, PC and CPC algorithms will produce the same adjacency predictions.

MGM-PCS and **MGM-CPCS** first learns an MGM and then uses the predicted undirected graph instead of a full graph as the starting point for PC-stable and CPC-stable respectively.

CPSS (Shah and Samworth 2013) is an a variation of the Stability Selection (Meinshausen and Bühlmann 2010) that both loosens the assumptions on the selection procedure (i.e. our network prediction algorithms are "selecting" edges), and tightens the bounds on the error rate, allowing for a less stringent threshold. Besides the obvious benefit of tighter bounds, the loose assumptions are especially attractive to us, as we would like to be able to substitute a variety of algorithms without worrying about violating the theoretical framework of the method. As with StARS and StEPS in chapter two, this method works by learning networks over subsamples of the data and counting how many times an edge appears. Rather than calculating network instabilities from these empirical edge probabilities, edges are selected by simply thresholding the probabilities. The threshold is calculated from the number of subsamples, the average number of selected edges, and the number of variables using Shah and Samworth's procedure. The user specifies an error control rate where errors are defined as edges that have a lower than random probability of being selected in a given subsample. We use a heuristic to adapt this method to directed edge recovery: the empirical selection probability of each edge direction setting (two directions or undirected) is calculated and thresholded separately. We ran CPSS in conjunction with MGM-PCS and MGM-CPCS with $\alpha = .05$ and $\lambda = .1$ for the LD dataset, and with $\lambda = .2$ for the HD dataset with error rates $q \in \{.001, .01, .05, .1\}$.

3.3.5 Conditional Independence Tests

The following test, used by PC-stable and MGM-PCS, is a hypothesis test for conditional dependence of two variables, X and Y, given a conditioning set of variables, S. The null hypothesis is that X and Y are independent given S, or $X \perp Y \mid S$. By definition, if this null hypothesis is true:

$$P(X,Y|S) = P(X|S)P(Y|S)$$

Rearranging, we find:

$$P(X|S) = \frac{P(X,Y|S)}{P(Y|S)} = P(X|Y,S)$$

So, to test $X \perp Y \mid S$ it suffices to test if $P(X|S) = P(X|Y,S)$ which is done via likelihood ratio test (LRT) of two regressions:

$$2 \ln \left(\frac{L(\theta_{\mathrm{XYS}})}{L(\theta_{\mathrm{XS}})}\right) \sim \chi^2(d_X d_Y)$$

Where the θs represent the regression coefficients to model X given S with and without Y as an additional independent variable. The degrees of freedom, d_X and d_Y, of each variable are 1 if the variable is continuous and the number of categories minus 1 if the variable is categorical. Although this description uses regressions with X as the dependent variable, the same reasoning allows us to use Y as the dependent variable instead.

The regressions in this test allow us to formulate this test so that any of the variables can be continuous or categorical. We preform linear or multinomial logistic regressions if the dependent variable is continuous or categorical, respectively. Because of this, if X and Y are of different variable types, we have a choice of whether X or Y should be the independent variable that determines whether we perform logistic or linear regressions. Our own experiments and observations in previous studies (Chen et al. 2014) suggest that a linear regression will give a more accurate test result than a logistic regression for these continuous-discrete edges. To handle any independent categorical variables in the regression, use the standard practice of converting each k-level categorical variable to k-1 binary variables.

It is also possible to conduct these tests by regressing Y and S onto X, and using a t-test to determine if the regression coefficient of Y is significantly different from 0. If Y is categorical this procedure requires performing a test on each dummy variable associated with Y and then combining them using Fisher's method. The main advantage of using t-tests over the LRT that it only requires one regression instead of two, so it is significantly faster. The downside is that in our experiments we found that it had less power to detect true edges, and was less robust to low sample sizes, particularly on edges that required a logistic regression. Because of this we will work exclusively with the LRT based test here.

3.3.6 Edge Recovery Evaluation

To evaluate network estimation performance, we compare the Markov equivalence classes of the estimated and true networks. Markov equivalence classes represent the variable independence and conditional relationships for an acyclic directed graph by removing the direction from edges that are free to point in either direction without altering the independence relationships in the network. For example, directed graphs $X \rightarrow Y \rightarrow Z$, and $X \leftarrow Y \leftarrow Z$ both have the Markov equivalence class $X - Y - Z$ while the graph $X \rightarrow Y \leftarrow Z$ (v-structure) would remain the same when converted to a Markov equivalence class. Thus, Markov equivalent graphs share the same variables, have the same adjacencies, and imply the same independence and conditional independence relations among their variables. We also consider performance on skeleton estimation, (i.e. the set of node adjacencies, without edge orientations).

We use standard classification statistics to evaluate the recovery of the undirected adjacencies from the skeleton of the true graph. Precision, also known as true discovery rate or positive predictive value, is the proportion of predicted edges that are found in the true graph. Recall, also known as sensitivity or true positive rate, is the proportion of edges in the true graph that were found in the predicted graph. For direction recovery we use these same statistics applied to the recovery of only the directed edges in the Markov equivalence class of the true graph. So, in the context of direction recovery, precision is the number of directed edges in the predicted graph that are found in the true graph out of the total number of directed edges in the predicted graph. Bi-directed edges are treated as undirected edges for these statistics because they do not give an indication of which edge direction is more likely. These statistics can be easily calculated from confusion matrices, which are shown for the undirected and directed graph estimation in **Figure 3.1**.

		Estimated Edge	
		X —> Y X —- Y X <— Y	X ⦸ Y
True Edge	X—> Y X —- Y	True Positive	False Negative
	X ⦸ Y	False Positive	True Negative

a.

		Estimated Edge	
		X —> Y	X—- Y X <— Y X ⦸ Y
True Edge	X —> Y	True Positive	False Negative
	X —- Y X <— Y X ⦸ Y	False Positive	True Negative

b.

Figure 3.1 Confusion matrices for edge recovery on **a.** undirected graphs and **b.** directed graphs.

We use the Matthews correlation coefficient (MCC) (Matthews 1975) as a measure for overall recovery performance that strikes a balance between precision and recall. The MCC is a formulation of Pearson's product-moment correlation for two binary variables (i.e. true edge indicators and predicted edge indicators). In addition, we use the structural Hamming distance (SHD) (Tsamardinos et al. 2006) as a combined measure of adjacency and direction recovery. The SHD is the minimum number of edge insertions, deletions, and direction changes, where only undirected edges are inserted or deleted, to get from the true Markov equivalence class to the estimated equivalence class.

3.4 RESULTS AND DISCUSSION

3.4.1 Simulation Experiments

In order to determine which algorithms have the most general applicability, we performed experiments using two different dataset sizes and randomly drawn DAG structures. In addition since optimal parameter setting is a difficult problem that may depend on the needs and goals of the user, we studied a range of possible parameter settings to show the relationship between these settings and edge recovery performance.

3.4.1.1 Adjacency Recovery

Figure 3.2 shows the (undirected) adjacency recovery performance of PC-stable, MGM-PCS and CPSS on the HD dataset. CPC-stable and MGM-CPCS are not shown because they have the same adjacency predictions as the PC algorithms. Settings of $\lambda < .2$ for the MGM-PCS algorithm are omitted from the figure because they mostly overlap with the PC-stable curves. Despite the apparent overlap, these denser MGM structures do cause a slight decrease in the precision of MGM-PCS compared to PC-stable, although this difference is not significant at any of the tested settings. For example, at $\alpha = .05$ and $\lambda = .14$, MGM-PCS has an average precision of .739 (standard error of .0057) compared to PC-stable which achieves mean precision of .744 (standard error is .0055). In the limit of $\lambda \rightarrow 0$, the MGM graph will become fully connected so MGM-PCS becomes equivalent to PC-stable.

On the other extreme, the highest settings of lambda result in very sparse initial graphs which have good precision but poor recall. In general, we see that adding the MGM step increases precision of the PC-stable procedure, at a small cost to recall, depending on the sparsity parameter setting. We see a similar trend in the LD dataset as well (see **Appendix A** for equivalent figures for the LD data). In addition, all of our algorithms have both lower precision and recall on edges involving discrete variables which suggests that they are more difficult to learn. This observations differs from the LD setting where we actually achieve the best recall on these *dd* edges, although still diminished precision compared to *cc* and *cd*. Finally, these results show that CPSS is a good option for users that want to ensure very high precision in their network estimates, and is certainly preferable to using an overly sparse setting of lambda.

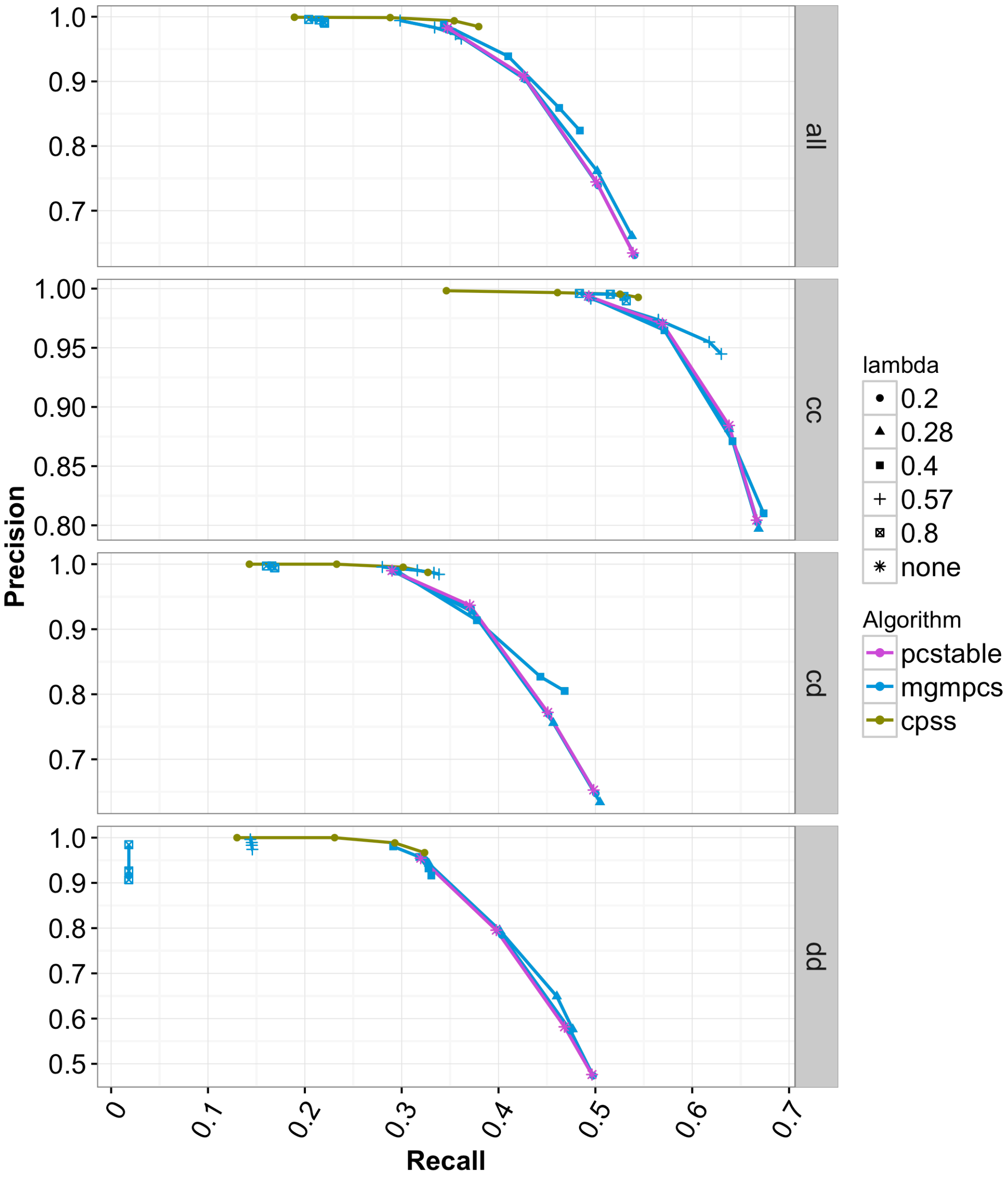


Figure 3.2 Precision-Recall curves of edge adjacency recovery on high-dimensional dataset for $.2 \leq \lambda \leq .8$ (represented by different shaped points) and $.001 \leq \alpha \leq .1$. For a given setting of λ, the different settings of α are connected by lines with colors corresponding to the algorithm. The cpss line shows the settings of error rate $q \in \{.001, .01, .05, .1\}$.

3.4.1.2 Direction Recovery

Next, we evaluated how well each algorithm was able to recover the directions in the directed edges of the true Markov equivalence class. For these tests, the positive class is all estimated directed edges, and the negative class is both undirected edges and the absence of an edge. So, an estimated edge is only considered a true positive if it correctly identifies both the existence and the orientation of the edge. **Figure 3.3a** shows these results across all of our algorithms. Starting from an MGM graph increases direction recovery performance in PC-stable. The main reason for this improvement appears to be the fact that PC-stable alone returns a large number of bidirected edges and only finds a small number of edges with a single direction. Bidirected edges are returned when the v-structure orientation rule in step 2 of PC-stable implies both directions for an edge. We treat these as undirected edges in our statistics. Starting from an MGM graph reduces the number of bidirected edges and increases the number of directed edge predictions. This is evident by the large increase in directed edge recall, but comes at the price of reduced precision for higher independence test thresholds, $\alpha \in \{.05,\ .1\}$.

Figure 3.3b gives us a detailed view of the direction recovery performance of CPC-stable, MGM-CPCS, and CPSS. As with adjacency recovery we see that as we increase lambda we achieve higher precision at the cost of recall. The reduced recall in $\lambda \in \{.28, .4\}$ is only slight combined with a significant increase in precision. We can also see that our heuristic for adapting CPSS to directed network recovery is perhaps too conservative as the recall is greatly reduced while precision is near perfect. Indeed with this set up CPSS predicts the directions of less than 10 edges on average, for the most lenient error rate, $q = .05$, so it does not seem to be a useful option for edge direction predictions.

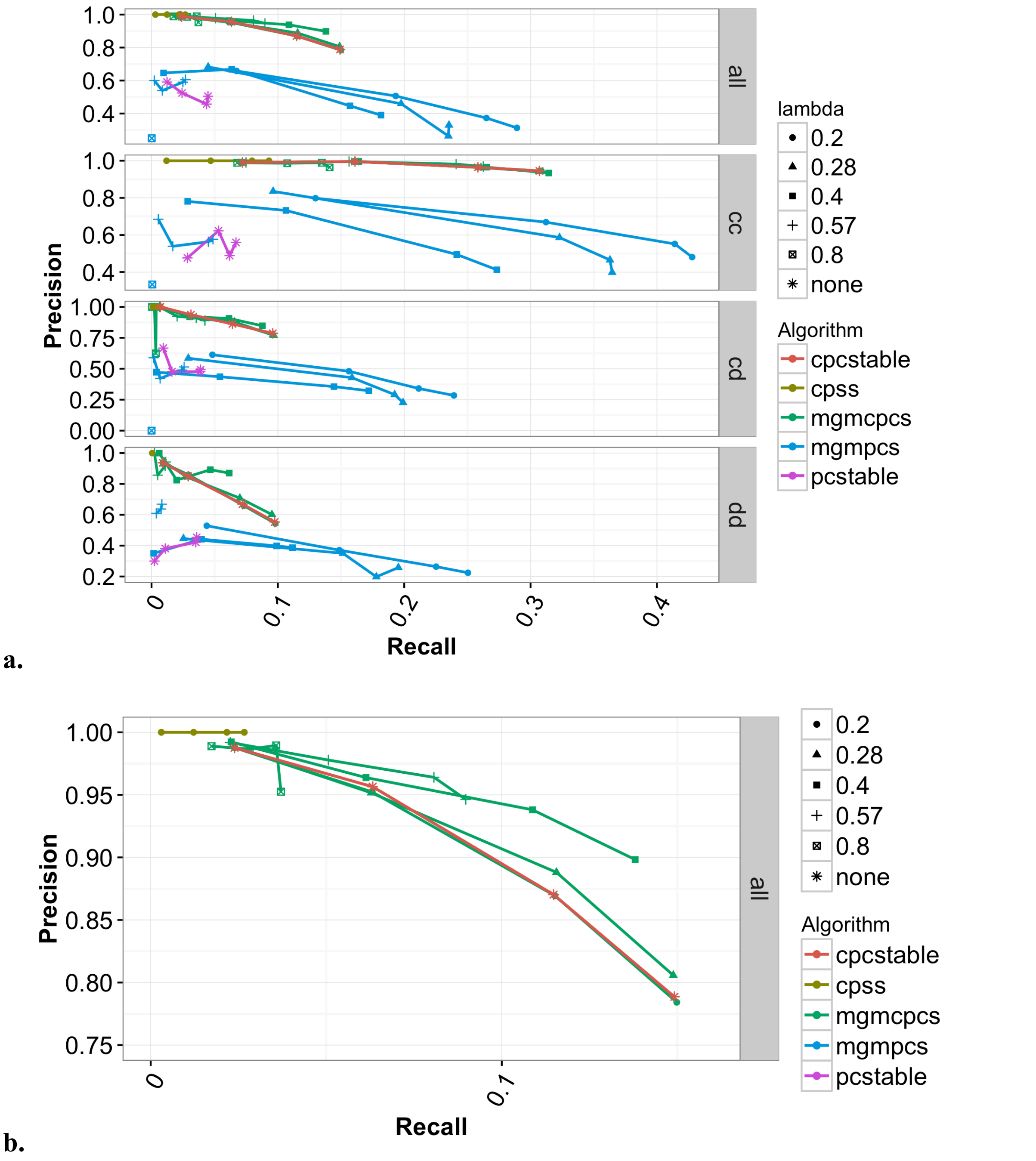


Figure 3.3 Precision-Recall curves of edge direction recovery on high-dimensional dataset for $.2 \leq \lambda \leq .8$ and $.001 \leq \alpha \leq .1$. **a** Full range of algorithms and edge types. **b** Detail view of CPC-Stable and MGM-CPC-Stable performance averaged over all edge types.

Overall, direction recovery is difficult in high dimensions. While MGM-PCS approaches direction recall of .3, this is paired with abysmal precision of less than .5. CPC-stable and MGM-CPCS give us high precision, but are able to recall less than 15% of true directed edges. Given that our heuristic to adapt CPSS to the problem of direction estimation that produces extremely high precision but very low recall, there is room for improvement in developing a less stringent heuristic.

3.4.1.3 Combined measures of network recovery

The Structural Hamming Distance (SHD) is a combined measure of adjacency and direction, that gives us an alternative network estimation metric that does not necessitate balancing precision versus recall. **Table 3.1** shows the "best case" performance of our algorithms, where the parameters settings are chosen to maximize the SHD both averaged over all edges, and broken down by each edge type. Since SHD is a distance measure, smaller values indicate better performance. By this measure, MGM-PCS and MGM-CPCS both significantly outperform their counterparts on the HD data. We see a similar trend in the LD data (**Appendix A**), where MGM-PCS performs significantly better than PC-stable, while MGM-CPCS has a slight but non-significant advantage over CPC-Stable.

Table 3.1 Parameter settings with the best SHD performance by edge type in high-dimensional data set

Algorithm	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$	*Type*	*SHD*
PC-Stable	0.01	none	all	600.95 (2.25)
	0.01	none	cc	130.00 (2.340)
	0.01	none	cd	308.40 (4.20)
	0.001	none	dd	160.45 (3.24)
MGM-PCS	0.01	0.14	all	567.75 (3.34)
	0.05	0.14	cc	108.45 (2.21)
	0.01	0.14	cd	294.70 (3.74)
	0.001	0.1	dd	157.30 (3.28)
CPC-Stable	0.01	none	all	588.10 (2.37)
	0.05	none	cc	111.60 (2.44)
	0.01	none	cd	307.05 (4.18)
	0.01	none	dd	160.80 (2.85)
MGM-CPCS	0.1	0.4	all	564.90 (4.46)
	0.1	0.57	cc	107.05 (2.32)
	0.1	0.4	cd	296.70 (4.17)
	0.1	0.4	dd	157.05 (3.25)

Since the best case performance will be difficult to achieve when the true graph is unknown, especially in this setting where a robust parameter setting scheme is not readily available, we also show SHD performance versus the number of predicted graph edges. These results, presented in figure **Figure 3.4** show that for parameter settings for MGM-CPCS that produce similar numbers of edge predictions to CPC-stable, the hybrid algorithm can improve SHD performance. Very sparse settings of λ result in networks with a large SHD because so many edges are missing compared to the true graph. These too-sparse settings of the MGM are evident from the number of predicted edges, however, so they should be easy for a user to identify.

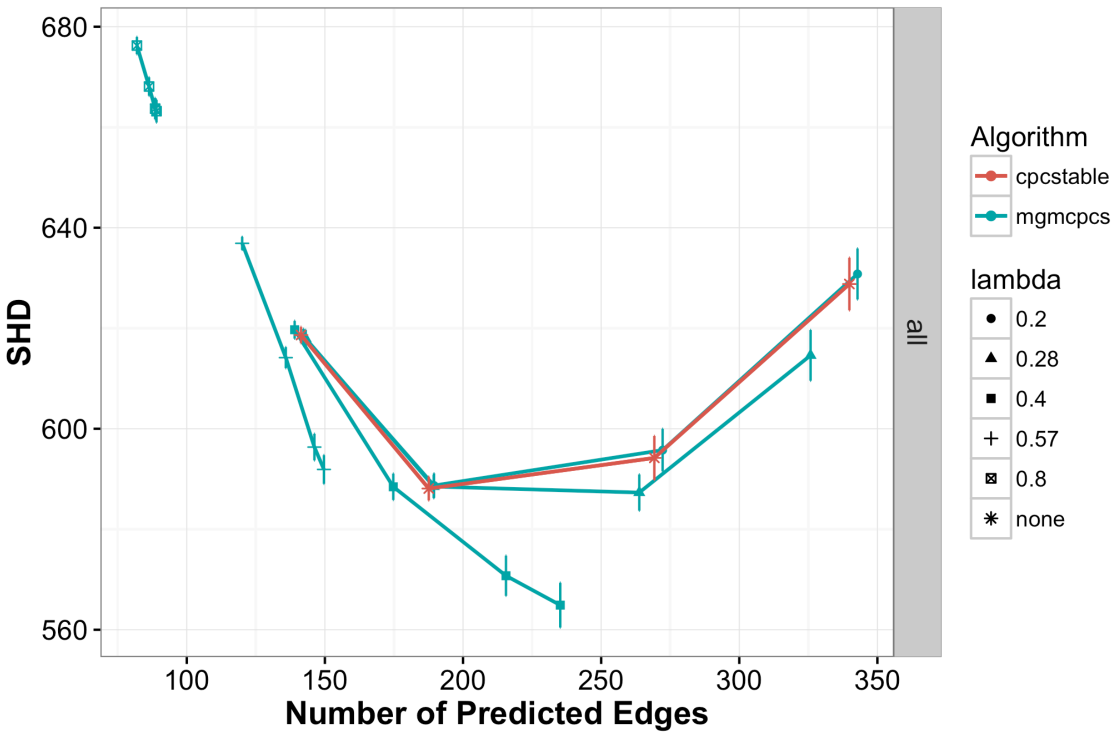


Figure 3.4 Structural Hamming Distance on high-dimensional dataset for CPC-stable and MGM-CPCS with for $.2 \leq \lambda \leq .8$ and $.001 \leq \alpha \leq .1$. The lower the SHD, the closer the predicted graph is to the true graph.

3.4.1.4 Run Time

We compared the running times of our algorithms at different parameter settings. **Figure 3.5** shows these results for the HD dataset. MGM-PCS and MGM-CPCS are significantly faster than PC-stable for sparser settings of λ, but significantly slower for low values of α and low values of λ. In the LD data (**Appendix A**), we see the increase in speed from the MGM step at almost all settings of α and λ. It is important to note that our MGM learning method is not parallelized, but the directed learning steps are, so a parallelized MGM learning algorithm could result in even larger speed improvements. The edge convergence approach we use to learning the MGM is essential to this performance improvement.

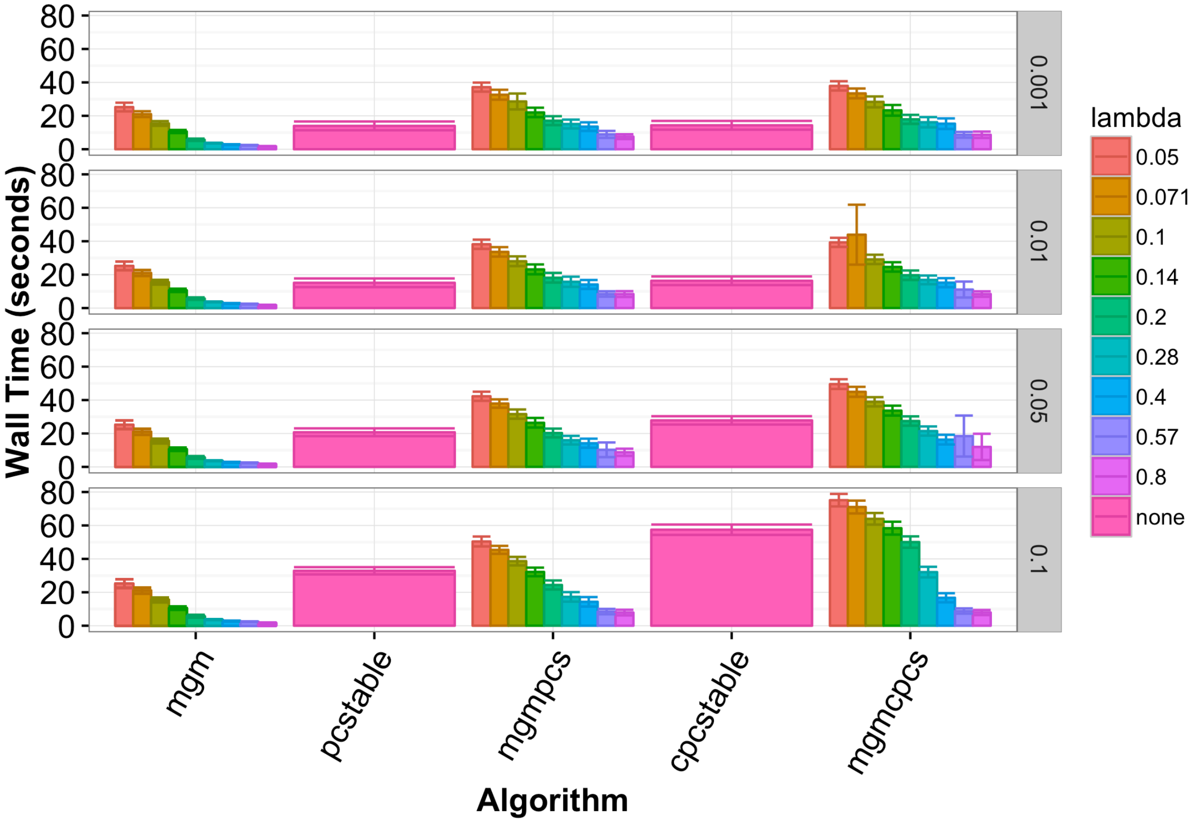


Figure 3.5 Average running times with 95% confidence interval error bars of search algorithms on high dimensional data. Each row of bars corresponds to a different setting of α and each color corresponds to a different setting of λ. Directed search steps were run in parallel on a 4 core laptop.

3.4.2 Application to Breast Cancer Data

We applied MGM-PCS to gene expression and clinical data from breast cancer patients curated by TCGA (Cancer Genome Atlas 2012). For the analysis we used the 500 genes with the highest variance across samples. We also included the clinical variables for hormone receptor status, node and tumor staging codes, and PAM50 subtype. PAM50 is a subtyping scheme that uses gene expression patterns from 50 genes to categorize tumors (Parker et al. 2009), thirteen of which were in the high variance set. We ran MGM-PCS with a sparsity of $\lambda = .2$ (selected based on stability of edges across subsamples, (Liu et al. 2010)), and $\alpha = .05$. The output network

(**Figure 3.6**) had 8 high variance genes connected to the PAM50 variable, 3 of which were among the 13 included in the analysis (Fisher's test, p=1.83*10^{-5}).

In addition, we find a number of predicted edges that are supported by biological knowledge. Each clinical variable corresponding to receptor status (ER, PR, HER2) was linked with the gene expression profile of that receptor: progesterone receptor with PGR1, HER2 with ERBB2, and estrogen receptor with ESR1. We find GATA3 is linked to ESR1, which relates to a recent study (Cimino-Mathews et al. 2013) that found GATA3 to be central in luminal (i.e. estrogen receptor positive). The lymph node stage variable, which indicates degree of lymph node metastasis, in our predicted network was only linked to the expression of E-cadherin gene (CDH1). Hypermethylation and decreased expression of CDH1 has been linked to infiltrating breast cancer (Caldeira et al. 2006).

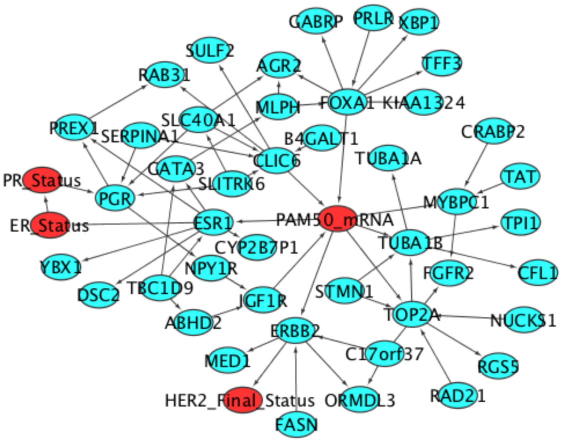


Figure 3.6 Predicted subnetwork for breast cancer dataset, with discrete clinical variables shown in red

3.5 CONCLUSIONS AND FUTURE WORK

We have shown that the combination of undirected graph search by optimizing conditional-Gaussian pseudo-likelihood over mixed data types, followed by directed graph search can recover causal information in complex systems containing both Gaussian and categorical variables. We have also shown that these methods can recover valuable information in real biomedical data. In many cases, our hybrid searches are faster and perform better than the directed search steps by themselves. In the worse case, our hybrid algorithms do no worse than the single algorithms searches and are slightly slower.

In addition to our newly proposed algorithms, this work provides an in-depth study of the challenges of causal learning on mixed data types. At both high and low dimensional settings, we are able to recover true edges from simulated data with high precision. Recall is more challenging, especially of edge direction and at high-dimensional settings. As expected, recovering edges and directions involving categorical variables was more difficult in high-dimensional settings, but this trend was surprisingly not obvious in the low-dimensional setting.

Directed MGMs are promising tools for exploratory biomedical research. As shown in this chapter and the next, they are able to recover both known and novel relationships between variables. We expect further work with these models to yield many more viable hybrid algorithms, as an undirected MGM can be adapted to serve as a starting point for a wide range of casual discovery algorithms.

4.0 APPLICATION TO METASTATIC MELAOMA

Here we present a case study for integrative data analysis that begins with an application of MGM-PCS, described in the previous chapter, to multimodal data from a cohort of metastatic melanoma patients, and follows our work to both validate and study the mechanism behind a biomarker identified by our algorithm. This study serves as an example of a successful application for our network learning algorithm for identifying a direct causal relationship between a single nucleotide polymorphism in the PARP1 gene, rs1805407, and response to chemotherapy. Through a combination of computational work performed by us and cell line experiments performed by Irina Abecassis and Maria Kapetanaki, we found evidence for a mechanism that begins the explain the causal link between these two variables. Because this mechanism does not involve any of the other variables that were measured in the melanoma cohort, it confirms our initial prediction of a direct connection between the variables given the available data.

4.1 CHAPTER SUMMARY

Personalized cancer therapy relies on the identification of patient subsets with differential responses to therapeutic interventions and stratifying them to maximize the therapeutic index or administer alternative regimens. We applied MGM-PCS to analyze mRNA and microRNA expression, DNA methylation, SNP, and clinical variables in a cohort of metastatic melanoma patients in order to identify direct causal interactions between variables. Our results show that ten gene expression, four methylation variables and SNP rs1805407 are directly linked to response to chemotherapy. SNP rs1805407 is located in PARP1, a DNA repair gene critical for chemotherapy response and for which FDA-approved inhibitors are clinically available (olaparib). We demonstrate that PARP inhibitors are synergistic to chemotherapy in cancer cells carrying the PARP1 variant, but they are additive or antagonistic to chemotherapy in wild-type cancer cells. Additionally, we found that TCGA melanoma and ovarian cancer patients that carry this SNP have increased expression of the PARP1-003 splice variant, a truncated form of PARP1. Based on these results we postulate that SNP rs1805407 is directly linked to increased sensitivity to PARP inhibitors, potentially through enhanced PARP1 trapping, in cancer cell lines from various histologies and most importantly ovarian cancer. Our results suggest that the combination of chemotherapy and PARP1 inhibition may benefit the carriers of the PARP1 SNP rs1805407. These findings demonstrate the utility of MGM-PCS and will inform personalized therapy to select patients more likely to respond to PARP inhibitors.

4.2 BACKGROUND

Cancer is among the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012 (World Cancer Report 2014 and (de Martel et al. 2012)). The number of new cases is expected to rise by about 70% over the next 2 decades. Advances in cancer management have improved the overall outlook of patients with metastatic malignancies but chemotherapy remains a mainstay of treatment for most common cancers. Virtually all patients develop resistance to chemotherapy after prolonged exposure given the first order kinetics of cytotoxics that generally cannot eradicate cancer. Understanding the mechanisms of this resistance presents new opportunities to improve the therapeutic index of cytotoxic agents and identify novel drug targets.

A large proportion of cytotoxic agents exert their effect through DNA damage. Thus, DNA repair pathways constitute cells' main resistance mechanisms and potential drug targets. Base excision repair, a predominant pathway for single strand break (SSB) damage repair, utilizes a family of related enzymes termed poly-(ADP-ribose) polymerases (PARP), which are activated by DNA damage (Luo and Kraus 2012). Given the critical role of PARP1 in base excision repair, PARP inhibition emerged as a therapeutic target and early studies demonstrated dramatic potentiation of chemotherapeutic agents in the presence of PARP inhibition (Bryant et al. 2005; Farmer et al. 2005). Recent evidence indicates that, in addition to the catalytic inhibition of PARP activity, PARP inhibitors (PARPi) induce cytotoxic PARP-DNA complexes through PARP "trapping" that augment the cytotoxicity of alkylating agents. It is therefore of utmost importance to identify molecular features that act not only as biomarkers for patient stratification but also offer insights into the mechanisms of resistance to chemotherapy. Metastatic melanoma remains an excellent model for chemotherapy resistance given its refractory nature, despite the fact that current management of metastatic melanoma is mostly based on non-chemotherapy based strategies (e.g., targeted and immune-based therapies).

In this study, we apply a novel graphical method we developed, MGM-PCS, to high-throughput data from a cohort of metastatic melanoma patients on chemotherapy. We identified various features that were directly linked to response to treatment, including a SNP in the PARP1 gene that is highly predictive of resistance to chemotherapy. We went on to characterize the impact of this PARP1 variant on PARPi sensitivity and demonstrated its utility as a predictive biomarker of PARPi sensitivity *in vitro.* Given the role of PARP1 in DNA repair, we propose this SNP as a biomarker for PARPi sensitivity to guide patient selection for treatment regimens incorporating PARPi's in combination with alkylating agents.

4.3 MATERIALS AND METHODS

4.3.1 Melanoma study design

Using a retrospective cohort study design, we evaluated 69 patients with metastatic melanoma who were treated with alkylator-based chemotherapy at the Melanoma Center of the University of Pittsburgh Cancer Institute (UPCI). Frozen tissues were available from metastatic lesions on 21 patients and formalin-fixed paraffin embedded tissues from 45 patients (total n=69). Only pre-treatment tumor specimens were included in this analysis. In addition, chemotherapy regimens studied were primarily single-agent dacarbazine (DTIC), single-agent temozolomide (TMZ) or DTIC-based combinations (including CVD, Cisplatin + Vinblastine + DTIC). Response to chemotherapy was defined as documented objective tumor regression upon treatment. Patients with disease progression after 2 cycles of chemotherapy or with stable disease lasting less than 4 months were considered non-responders.

4.3.2 Using Mixed Graphical Model learning (MGM-PCS) to integrate -omics and clinical data

We normalized the data for use with MGM-PCS in the following way. Each continuous variable was transformed so that its distribution across patients was normal with truncated tails using the non-paranormal method (Liu et al. 2009). In addition, the 10% of SNPs with the lowest variance were filtered out.

To filter this large dataset down to a size that was feasible to use with MGM-PCS, we filtered the variables based on their pair-wise correlation with the response to treatment variable. In order to accurately calculate correlation between this discrete variable and other discrete or continuous variables, we used a generalized correlation metric, described below. With this metric, we filtered the data down to the 1000 variables that were most correlated with response to treatment.

4.3.3 Filtering with generalized correlation

We use the following strategy to measure association between a continuous and categorical variable or two categorical variables. We would like to calculate the equivalent of Pearson's product moment coefficient for each possible pairing of these variables. The general formula for Pearson's correlation between two vectors of observations, *X* and *Y,* with means μ_X and μ_Y and standard deviations σ_X and σ_Y is $r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ where covariance is defined as $cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$. This is a standard calculation for pairs of continuous variables because mean and standard deviation are well defined. For pairs of binary variables, these values are also well defined, and this formulation is called the Matthews' Correlation Coefficient. For categorical variables we can calculate the covariance on a category by category basis. So for a categorical *X* continuous *Y*, we can focus on a, one of the categories of *X* when calculating a sample covariance: $cov(X_a, Y) = E\big[\big(X_a - \mu_{X_a}\big)(Y - \mu_Y)\big] = \frac{1}{N-1}\sum_{i=1}^{N}[(\mathbb{I}(X_i = a) - \hat{p}_a)(Y_i - \hat{\mu}_Y)]$ where $\mathbb{I}(X_i = a)$ is an indicator function that is 1 when $X_i = a$ and zero otherwise, and $\hat{p}_a = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(X_i = a)$ or the empirical probability of observing a in X. Since $\mathbb{I}(X_i = a)$ is equivalent to a Bernoulli random variable now it is easy to see that the sample standard deviation is $\hat{\sigma}_{X_a} = \sqrt{\frac{N}{N-1}\hat{p}_a(1 - \hat{p}_a)}$. Similarly, if both X and Y are categorical we now look at each possible pairing of categories separately so $cov(X_a, Y_b) = \frac{1}{N-1}\sum_{i=1}^{N}[(\mathbb{I}(X_i = a) - \hat{p}_a)(\mathbb{I}(Y_i = b) - \hat{q}_b)]$ where $\hat{q}_b$ is the empirical probability of observing *b* in *Y*. So, in a discrete-continuous pair, we now have a vector for the covariance and a vector for the standard deviations corresponding to the different levels of the categorical variable, we use the l_2 norm to calculate a single score from these vectors (where X is categorical): $r_{XY} = \frac{\|cov(X,Y)\|_2}{\|\sigma_X\|\sigma_Y}$. In the discrete-discrete case we have two matrices corresponding to the possible pairs of levels in the two variables, and we combine them with the Frobenius norm: $r_{XY} = \frac{\|cov(X,Y)\|_F}{\|\sigma_X\sigma_Y\|_F}$. Both of these cases result in non-negative values so to make the continuous-continuous values comparable with the others we take the absolute value so scores for all pairs of edges fall on the interval [0,1].

One motivation for this approach is that these sample covariances turn out to be proportional to the partial gradients of negative log pseudolikelihood in a factorized (i.e. zero edges) MGM as described above with respect to the edge parameters and variable levels (see (Lee and Hastie 2013) supplement). Namely: $\frac{\partial \tilde{l}}{\partial \beta_{ij}} = -2 * (N-1) * cov(X, Y)$, $\frac{\partial \tilde{l}}{\partial \rho_{ij}(a)} = -2 *$

$(N-1) * cov(X_a, Y)$ and $\frac{\partial \tilde{l}}{\partial \phi_{ij}(a,b)} = -2 * (N-1) * \ cov(X_a, Y_b)$ where X is the indexed by i and Y is indexed by j in the MGM and the pairs of X and Y are continuous-continuous, discrete-continuous, and discrete-discrete respectively.

4.3.4 Patient data from The Cancer Genome Atlas

We collected publically available RNAseq and genotype data generated by the The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/) for 418 ovarian cancer tumors and 293 melanoma tumors. RSEM (Li and Dewey 2011) quantified transcript abundances from TCGA were used for genome-wide mRNA expression profiling. We used Kallisto (Bray et al. 2015) to quantify PARP1 splice variant abundance. Genotype data came from the Affymetrix Genome-Wide Human SNP Array 6.0 platform.

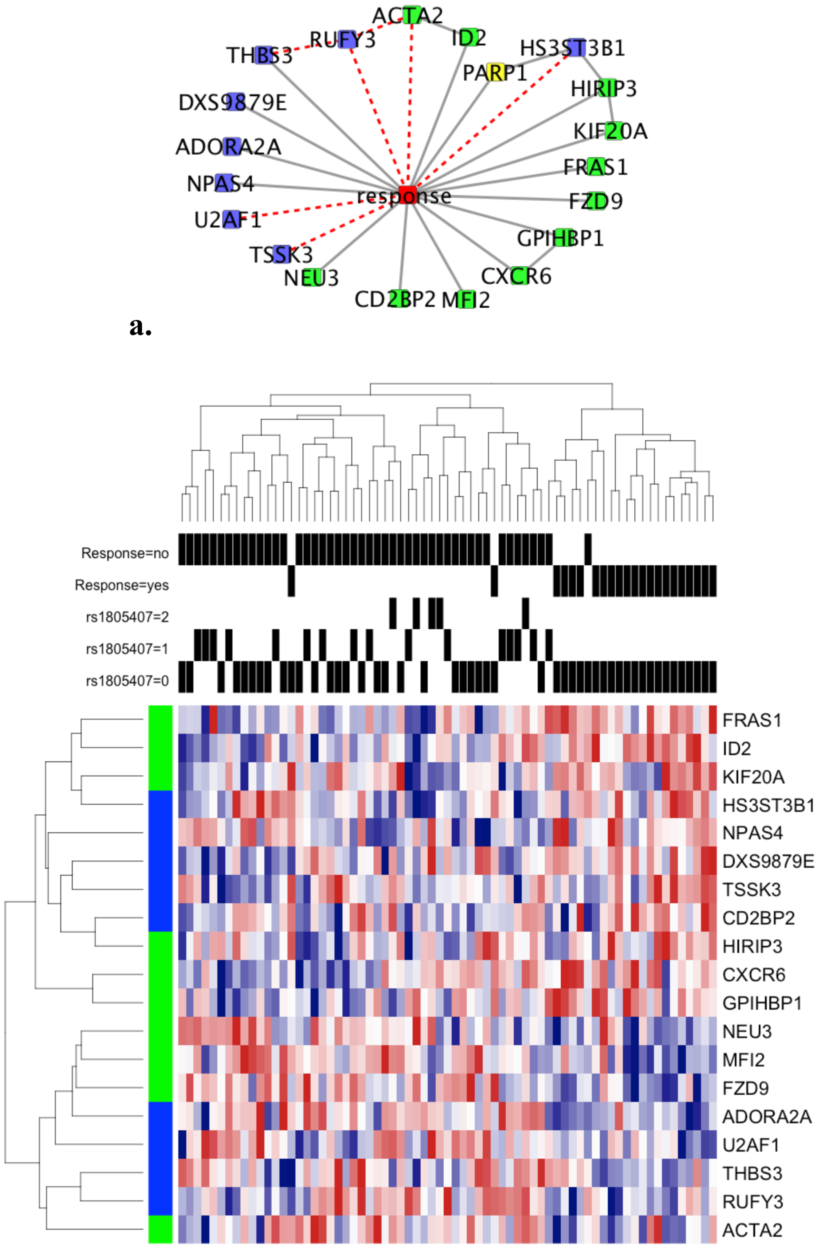
4.3.5 SNP imputation on TCGA samples and NCI-60 cell lines

We obtained NCI-60 data from Cell Miner in June 2013 (http://discover.nci.nih.gov/cellminer/). For those cell lines or TCGA samples for which the identity of SNP rs1805407 was not available we used imputation to infer its identity. Using SNAP (Johnson et al. 2008) we found 51 SNPs to be in perfect linkage disequilibrium (LD) with rs1805407 (R^2 = 1). Of these, 9 variants were covered by the Affymetrix SNP Array 6.0 used by the TCGA. To determine the rs1805407 genotype in TCGA samples we used birdseed calls (Korn et al. 2008) available from the genotype data. Only samples with a birdseed confidence less than 0.1 or where all 9 SNPs in perfect LD agreed with the birdseed call were used.

4.4 RESULTS

4.4.1 Identifying predictive markers of treatment for metastatic melanoma patients

Our melanoma dataset consisted of gene expression, microRNA expression, DNA methylation, and data from the selected SNP panel. We used MGM-PCS to learn a network over the top 1000 features most correlated (pairwise) with "*response*" clinical variable, a binary variable indicating response/no response to TMZ treatment dichotomized at presence of a response or stability of disease at 4 months of therapy. The 1000 most correlated features in the input dataset included 557 mRNA expression probes, 425 methylation probes, 14 miRNA probes and 4 SNPs. BRAF mutation status was also included in the input variables to see if it had an effect on any of the features linked to *response*, although its direct correlation with it in this dataset was poor (R^2=0.025). The largest interconnected output network included 20 features directly connected to the *response* variable in our initial (undirected) learned network (**Figure 4.1a**). We emphasize that these features were connected to the *response* variable not only because they have high pairwise correlation with it (**Figure 4.1b**), but also because they are dependent on response even when conditioned on all other variables in the filtered dataset. In this sense, they represent direct (causal) interactions and not simple biomarkers. From the 20 features initially connected to response in the undirected MGM model, 15 were left connected after the causal filtering (**Figure 4.1a**, black lines). A methylation feature for DXS9879E (LAGE3) is one of them, and it has been linked to survival in non-small cell lung cancer (Lokk et al. 2012). Notably, this important feature is not present in the top 20 (pairwise) correlated features with *response*, as other (indirect) interactions yielded higher pairwise correlation values. ID2 expression is also directly linked to *response*, and is known to induce growth and proliferation in squamous cell carcinoma (Wang et al. 2012a).



b.

Figure 4.1 a. Conditional Gaussian sub-network around response to treatment with edge filter. Blue nodes represent methylation probes, green nodes represent mRNA expression probes and yellow nodes represent SNPs. Dashed red lines indicated edges removed by filtering step. **b.** Heatmap of directly connected links to response to TMZ. Black bars show rs1805407 status and response to treatment. Rows marked with green are mRNA expression profiles and those marked with blue are methylation profiles.

4.4.2 SNP rs1805407 in PARP1 is strongly associated with worse outcome in melanoma patients

SNP rs1805407 (PARP1) is the only SNP in our dataset that is directly linked to *response*. We found this SNP to be an excellent predictor of worse outcome since all 21 patients that had the SNP (C/T or C/C) showed no response to TMZ treatment while of the remaining 48 patients with rs1805407 T/T, 22 responded to treatment (*p*-value=4.6e-5). Because of the strong direct (causal) association, the role that PARP1 plays in repairing TMZ mediated DNA damage, and the availability of PARP inhibitors, we decided to investigate this finding further. rs1805407 is located on the 2nd intron of PARP1, ~4 Kbp downstream of the PARP1 transcription start site and 35 bp downstream of the 3' splice site of exon 2. We used SNAP (Johnson et al. 2008) to find other SNPs in strong linkage disequilibrium (LD) with rs1805407. The CEU population panel of 1000 Genomes pilot 1 contained 51 variants in perfect LD ($R^2 = 1$) (**Appendix B, Table B.1**) with our selected variant. Of these, two are upstream of the transcription start site (TSS). Specifically, rs6665208 is 3,573 bases upstream of the PARP1 TSS and overlaps with ENCODE ChIP peaks for MAFF and MAFK, which are both related to blood cancers (Balkhi et al. 2006). rs2077197 is 238 bases upstream and overlaps peaks for AP-2α, CTCF, HA-E2F1, ZBTB7A, Pol2, CEBPB and YY1. Many of these factors are related to cancer. For example, AP-2α and ZBTB7A are known tumor suppressors (Liu et al. 2014; Su et al. 2014). CEBPB plays a role in senescence of prostate cancer cells (Barakat et al. 2015) and in multi-drug resistance (Riganti et al. 2015). E2F1 is induced by DNA damage (Lin et al. 2001). CTCF is an insulator protein and YY1 participates in long-range chromosomal interactions.

Another SNP (rs1805405) is also in perfect LD with rs1805407 ($R^2 = 1$) and is annotated as a 'Splice region variant' because it is located 5 bp upstream from the splice site between intron 2 and exon 3. Finally, we found a strong dependence between rs1805407 and two other SNPs that have been previously associated with melanoma susceptibility: rs3219090 (D' = 1, R^2 = 0.43 (Macgregor et al. 2011) and rs2249844 (D' = 1, R^2 = 0.46) (Davies et al. 2014). Due to high LD values with rs1805407 none of them was included in our original SNP panel.

4.4.3 SNP rs1805407 is related to decreased cytotoxicity of alkylating agents in cell lines with the variant

Given that rs1805407 is associated with worse outcome of melanoma patients treated with TMZ and that PARP1 has a critical role in repair of DNA lesions caused by TMZ, one plausible hypothesis is that rs1805407 is either associated with increased PARP1 expression and/or activity or decreased PARP1 trapping after treatment with alkylating agents (Murai et al. 2012; Murai et al. 2014a; Murai et al. 2014b). We looked for more insights into the role of rs1805407 in cell response to various drugs. The NCI-60 Cell Miner database (Reinhold et al. 2012) contains the response of 60 cell lines to ~50,000 compounds. We evaluated whether drugs affect differentially the cell lines that have at least one copy of rs1805407 (C/T or C/C) *vs* WT (T/T). The Affymetrix 500k SNP arrays used by Cell Miner did not include rs1805407, so we used the *k*-nearest neighbors method with three of the 51 perfectly correlated variants with probes in the array (rs1073991, rs10799349 and rs3219027) to infer the rs1805407 genotype in each cell line. Analysis of the IC_{50} values in cell lines predicted to have at least one allele of rs1805407 (n = 23; C/T or C/C) *vs* wild type (n=37; T/T) showed statistically significant resistance or sensitivity to four compounds, three of which are alkylating agents with action similar to TMZ (TMZ is not included in the NCI-60 dataset) (**Table 4.1**). Cell lines containing the PARP1 SNP showed only a slight increase in sensitivity to the PARPi olaparib used as a single agent (**Table 4.1**).

Carmustine and Cyclophosphamide are classical DNA damaging alkylating agents. Parthenolide, a compound that induces apoptosis in acute myelogenous leukemia (AML) and progenitor cells (Guzman et al. 2005). Increased sensitivity was observed for Irofluven, an alkylating agent that inhibits DNA replication (Wang et al. 2007). For comparison purposes, we also added the PARP1 inhibitor Olaparib (which is not statistically significant when used as single agent). These results are compatible with the hypothesis that SNP rs1805407 (or one of the 51 SNPs in perfect LD with it) may cause increased PARP1 expression and/or activity thus helping to repair the damage caused by TMZ; or decreased PARP1 trapping potentially eliminating the additional cytotoxic effects of PARP-DNA complexes induced by PARP inhibition.

Table 4.1 Drug compounds with differential IC_{50} values on WT vs SNP cell lines for rs1805407. GI50 values derived from NCI60. Statistical significance was assessed with Wilcoxon rank sum test.

NSC	Name	FDA status	u	p
26271	Cyclophosphamide	FDA approved	175	0.01
157035	Parthenolide	FDA approved	148	0.02
683863	Irofulven (Hydroxymethylacylfulvene)	FDA approved	267.5	0.02
409962	Carmustine	FDA approved	294	0.03
747856	Olaparib	FDA approved	342.5	0.64

4.4.4 SNP rs1805407 is related to PARP inhibitor potentiation of alkylating agent cytotoxicity

Experimental validation of the association of a SNP to increased PARP1 activity in patient-derived tissues is not straightforward because PARP1 is an inducible enzyme and its activity may depend on the timing of the biopsy with respect to prior therapies. Alternatively, one can use cell lines to test whether PARP1 inhibition affects the response to alkylating agents in an SNP-dependent way. This can be done by blocking PARP1 after treatment with an alkylating agent in cells with or without the SNP.

Literature search (performed in conjunction with Irina Abecassis) identified 13 cell lines (**Table B.2**) from various tumor types, which had reported activity of alkylating agents alone or in combination with a PARPi (CEP-6800, AG14361, NU1085, NU1025 or ABT-888) (Delaney et al. 2000a; Miknyoczki and Jones-Bolin 2003; Tentori et al. 2003; Calabrese et al. 2004a; Wang et al. 2012b; Davidson et al. 2013). We used qRT-PCR (experiments preformed by Irina Abecassis) to confirm the rs1805407 genotypes in these cell lines. We classified cell lines as "resistant" to the combination of TMZ with PARPi if the reported experiments did not show potentiation of cytotoxicity of TMZ when a PARPi was added; likewise, cell lines that showed significant potentiation of TMZ cytotoxicity with PARPi were classified as "sensitive". Four of Seven cell lines that had at least one C in position rs18050407 were sensitive, and all six cell lines that were WT (T/T) were resistant (*p*=0.01, G test).

Although these are intriguing observations, the information used in the analysis was collected from different labs that used different PARPi to determine sensitivity or resistance. Therefore, we performed similar experiments on nine cell lines from various histologies (melanoma, lung, colon, ovarian, and breast cancer): five WT (T/T) and four C/T for SNP rs1805407 (**Table B.3)**. MMS was used as alkylating agent and ABT-888 was used to inhibit PARP activity. All four SNP cell lines were found to be significantly more sensitive to the combination treatment in agreement with our hypothesis; while in all five WT cell lines the combination treatment had no potentiation effect (**Figure 4.2a**). Notably, for the WT cell line SW620 ABT-888 significantly increased the IC_{50} of MMS suggesting potential antagonism (experiments performed by Irina Abecassis).

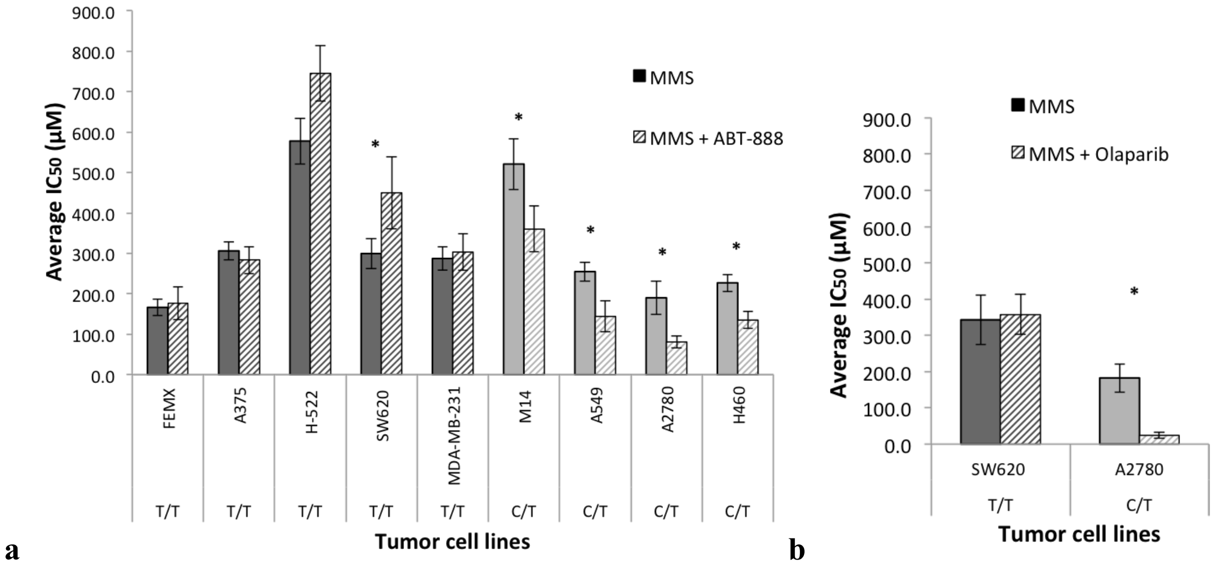


Figure 4.2 PARP1/SNP genotype is predictive of MMS+PARPi combination treatment efficacy. Plot of the IC_{50} values. **a.** ABT-888; and **b.** olaparib treatment. *MMS*: alkylating agent used; *ABT-888* or *olaparib*: PARP1 inhibitor used; *left bars*: MMS only; *right bars*: MMS+PARPi. *Dark grey bars:* wild type (T/T) for rs1805407; *light grey bars:* heterozygotes (C/T). *Star* indicates that combination treatment (MMS+PARPi) has significantly different effect than alkylating agent alone (p<0.05, Student's t-test, paired two-tailed).

We extended those results to olaparib treatment of A2780 (SNP) and SW620 (WT) cell lines and observed similarly significant potentiation of MMS cytotoxicity with the addition of olaparib in A2780 and no potentiation in the SW620 cells (**Figure 4.2b**). We note that the potentiation factor in A2780 cells increased from 2.35 with ABT-888 (10 nM) to 4.65 with olaparib (5 nM) (**Table B.3**).

4.4.5 PARP inhibitors and alkylating agents exhibit synergy in relation to SNP rs1805407 and antagonism in relation to the wild-type genotype

We investigated the potential combinatorial effects of alkylating agent (MMS) and PARPi (ABT-888) on cell lines with different rs1805407 genotypes. Exponentially dividing cells were exposed for 72h to increasing concentrations of ABT-888 (0-500 μM) or MMS (0-1 mM) alone (single drug treatment) or combined at a fixed ratio based on their corresponding IC_{50} value (drug combination treatment). Cell survival was assessed by MTT assay. We then determined the Chou-Talalay combination index (C.I.) (Chou and Talalay 1984) of ABT-888 with MMS in four established cell lines: two with the variant C/T (A2780-ovarian cancer and M14-melanoma) and two WT T/T (SW620-colon and H522-lung cancer). Although C.I. is not a statistical measure, it nevertheless provides insight into whether the effect of both agents is additive (C.I.=1), synergistic (C.I.<1), or antagonistic (C.I.>1). Both cell lines with the SNP variant exhibited synergy (strong and moderate effect in A2780 and M14, respectively; **Table 4.2**). In WT cell lines the effect is additive at best (H522). Interestingly, the SW620 cell line had decreased MMS cytotoxicity after the addition of ABT-888. This might indicate antagonism as the C.I. was substantially higher than 1.

Table 4.2 ABT-888/MMS combination indices (C.I.-ED50) in WT vs PARP SNP rs1805407 carrier cell lines. C.I.-ED50 values (mean ± SD) are indicated. For each cell line, equipotency ratios were calculated from the IC50 of MMS and ABT-888 used as single agent by MTT assay as outlined in the Methods section. Data from n≥3 independent experiments was used to calculate the Combination Index (C.I.) with the Compusyn program according to the method of Chou and Talalay. A C.I. < 1, = 1, and >1 is indicative of synergism, additivity and antagonism, respectively. ED50 is defined as the median effective dose.

Cell line	PARP1/SNP genotype	Combination ratio	C.I.-ED50	Interpretation
A2780	C/T	3.8:1	0.18 (± 0.04)	strong synergy
M14	C/T	2.3:1	0.74 (± 0.08)	moderate synergy
H522	T/T	2.0:1	1.26 (± 0.26)	slight antagonism/additivity
SW620	T/T	1.2:1	1.41 (± 0.08)	moderate antagonism

4.4.6 SNP rs1805407 is linked to higher expression of PARP1-003 splicing variant

So far, we have established that PARP1 SNP rs1805407 is directly linked to worse response to alkylating therapy in metastatic melanoma patients and in combination treatment of cell lines. Furthermore, we have shown that there is a synergistic effect of alkylating agent MMS and ABT-888 in cell lines with the SNP. However, we do not have a plausible molecular mechanism for this phenomenon. Analysis of our melanoma cohort and the TCGA melanoma data showed that PARP1 expression does not increase significantly in metastatic versus non-metastatic patients (data not shown). According to Ensembl, however, there are ten PARP-1 alternatively spliced transcript isoforms, four of which are protein coding: the predominant form, PARP1-001, and PARP1-003, PARP1-005 and PARP1-201. In the predominant form of PARP1 transcript, SNP rs1805407 is located 35 bases downstream of the end of exon 2 and it is in perfect LD with SNP rs1805405, which is s located 5 bp upstream from the intron 2/exon 3 splice junction and is annotated as 'Splice region variant'. PARP1-003 (ENST00000366790) misses the splice site at the end of exon 2/exon-3 (chr1:226,590,083; hg19) and continues translating another 59 amino acids before it reaches a stop codon (**Figure 4.3**). Thus the resulting protein, PARP1-003, is short (155 amino acids instead of 1,014) and contains only the first of the two zinc finger domains (ZF1) and none of the two catalytic PARP1 domains (Eustermann et al. 2011). It is known that ZF1 plays a role both in recognition and DNA binding and PARP catalytic activity (Ikejima et al. 1990; Mortusewicz et al. 2007; Altmeyer et al. 2009), but its overexpression leads to inhibition of alkylation-induced DNA repair in mammalian cells (Molinete et al. 1993).

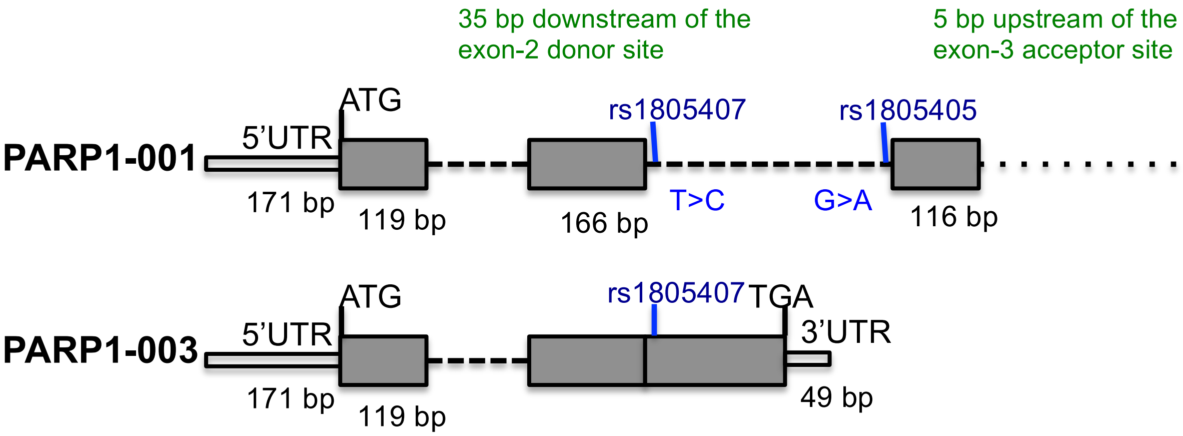


Figure 4.3 Genomic structure of the fist exons of PARP1 main transcript (PARP1-001) and its truncated alternatively spliced form (PARP1-003). This variant is found to be increased in patients with at least one rs1805407 minor allele.

To further study the potential role of rs1805407 in the function of PARP1, we calculated the abundances of the various PARP1 isoforms in 418 TCGA ovarian cancer patients (Level 2 data) and compared these abundances between groups of patients with and without rs1805407. We found a strong association between rs1805407 and the abundance of the PARP1-003 isoform. Carriers of at least one variant allele (N = 132) have small but statistically significantly higher ratio of PARP1-003 abundance to PARP1-001 abundance compared to non-carriers (N=286) (p = 4e-08, Mann Whitney U test). This ratio ranges from 0.11% for non-carriers to 0.32% for SNP homozygotes. Similar, but less significant, results were observed in the 293 TCGA metastatic melanoma patients (data not shown).

Finally, we performed quantitative PCR (qRT-PCR) on three cell lines (T/T: H522, SW620; C/T: A2780) with primers specific for the full length PARP1 (PARP1-001) and PARP1-003 (experiments performed by Maria Kapetanaki). The qRT-PCR was performed on RNA collected before and after treatment with either DMSO as control or ABT-888. We found that the expression of both isoforms in A2780 cells is significantly higher than in WT cells. Also, ABT-888 significantly inhibits PARP1-003 mRNA expression, while it has no effect on full length PARP1. However, PARP1-003 downregulation in WT cells is more profound than in SNP cells (74% and 69% for H522 and SW620, respectively compared to 38% for A2780). This result shows a trend that is consistent with the theory that SNP carrier cells have increased expression of the short, ZF1 containing PARP1-003 isoform, which makes them more sensitive to a combination therapy as its overexpression leads to inhibition of alkylation-induced DNA repair (Molinete et al. 1993) and catalytic inhibition of PARP leads to PARP trapping (Hopkins et al. 2015).

4.5 DISCUSSION

Application of different therapies in well-defined subgroups of patients is the holy grail of cancer precision medicine. Using MGM-PCS, a new method for learning probabilistic graphical models over multi-modal biomedical and clinical data (Sedgewick et al, in preparation), we discovered a novel biomarker (PARP1 SNP rs1805407) that can be used to identify patients who respond poorly to chemotherapy and potentially more favorably to PARP inhibition. SNP rs1805407 has relatively high prevalence (24-32% and 65.5% in European and African populations, respectively). Adequately powered future validation studies will test the clinical application of this SNP as a response biomarker.

PARP1 acts as a “molecular sensor” to identify DNA single strand breaks. It is then recruited and activated as a homodimer in a fast reaction which is amplified 10 to 500-fold with formation of poly-(ADP-ribose) (PAR) polymers within 15-30 seconds. Upon binding to a damaged strand via its zinc finger DNA-binding domains, PARP-1 undergoes a conformational change inducing the C-terminal catalytic domain to transfer ADP-ribose moieties to protein acceptors, including the central auto-modification domain of PARP1 itself. AutoPARylation of PARP1 and PARylation of chromatin proteins promote the recruitment of DNA repair factors (Masson et al. 1998; El-Khamisy et al. 2003; Schreiber et al. 2006). Extensive autoPARylation of PARP1 results in dissociation from DNA, which is required for DNA repair completion (Satoh and Lindahl 1992). The impact of PARP inhibition is more profound than simple inhibition of catalysis. For instance, wild-type cells are more sensitive to a PARPi combined with the alkylating agent MMS than Parp1−/− mouse cells (Horton et al. 2005; Heacock et al. 2010; Kedar et al. 2012). Furthermore, PARP inhibition delays DNA repair to a greater extent than PARP depletion (Strom et al. 2011). To explain these results, a PARP1-trapping model has been proposed (Helleday 2011; Kedar et al. 2012). This model is based on the idea that PARP1 is trapped on DNA by PARPi's since the automodification and PAR synthesis electrostatically destabilizes the PAPR1-DNA complex and lead to rapid dissociation. PARPi's therefore stabilize PARP-DNA complexes, which are themselves cytotoxic and may underlie the differential efficacy of clinically relevant PARPi's, which differ markedly in their PARP trapping potency (Murai et al. 2014a).

Based on our findings we postulated that SNP rs1805407 affects PARP1 activity during treatment with alkylating agents. We found that SNP cell lines were sensitive to combination therapy, while WT in general were not. We repeated these experiments with the FDA approved PARPi olaparib in A2780 SNP carrier cells and the results were consistent and perhaps more profound than with ABT-888. Furthermore, we showed that in SNP cell lines the combination treatment was synergistic, while in WT cell lines was at best additive or even antagonistic. We also note that this effect is independent of BRCA1 mutation since none of the tested cell lines carried this mutation. These are very important results as SNP rs1805407 can be potentially used in the future to decide whether a patient will receive a combination therapy or not.

Next, we investigated potential molecular mechanisms that can explain these data. SNP rs1805407, located in the beginning of intron-2, is in perfect LD with rs1805405, which is located just 5 bp upstream of the intron-2/exon-3 splice junction. We investigated whether rs1805407 is signifying an increase in expression of PARP1-003, an alternatively spliced isoform that codes for a shortened protein that includes only exons 1 and 2, thus coding only for the first zinc finger (ZF1) of PARP1. In PARP1 protein, ZF1 is responsible for recognition of DNA lesions and for initiation of DNA repair (Ikejima et al. 1990; Mortusewicz et al. 2007; Altmeyer et al. 2009). However, its overexpression can lead to inhibition of DNA repair (Molinete et al. 1993). Also, the short isoform uniquely maintains the auto-modification domain, which is most critical to PARP trapping

By analyzing the TCGA RNA-seq melanoma data we discovered that patients with one or two alleles of rs1805407 have significantly increased relative abundance of isoform PARP1-003 vs PARP1-001 (the full length mRNA), but the difference was small in absolute numbers. We should note, however, that. TCGA data are generated from pre-treatment samples, generally from primary tumors, and in melanoma predominantly from metastatic sites. PARP1 expression is accelerated at the time of DNA damage. The presence of SNP rs1805405 may affect the splicing rates and relative abundances of PARP1 isoforms, thus altering their relative protein levels. This is relevant in patients, especially knowing that TMZ is utilized in regimens that span either 5 days or 21 days out of 28-day cycles. Further experiments in a clinical setting are required to determine whether this hypothesis is true.

We performed qRT-PCR to measure expression of PARP1-001 and PARP1-003 before after ABT-888 treatment. While qRT-PCR results can not measure relative abundance of PARP1-001 and PARP1-003, they showed that SNP carrier A2780 cells have higher baseline expression of PARP1-003 than WT cells. Our analysis identified two SNPs in the PARP1proximal promoter, which are in perfect LD with rs1805407, and overlap peaks of many cancer-related transcription factors. These may account for the differences in baseline abundance of both isoforms in SNP vs WT cells.

qRT-PCR also showed that ABT-888 downregulates PARP1-003 mRNA levels significantly in WT cells, while reduction in A2780 cells is modest. When a PARPi is utilized, then full length protein PARP can bind to the DNA lesion, but not initiate catalysis. Therefore, it becomes non-functional and trapped. This occurs on WT and SNP carriers at similar rates. However, SNP cells express significantly more PARP1-003 than WT cells, and PARP trapping will occur at a higher rate in the presence of more PARP1-003, resulting in increased cytotoxicity and synergism of PARPi with chemotherapy in cells with the variant SNP. In WT cells, the levels of PARP1-003 are likely not high enough to have a noticeable effect on cytotoxicity as our results show. Our findings were validated in cancer cell lines from various histologic subtypes indicating their relevance to the underlying biology of PARP inhibition; and are most remarkable in ovarian cancer for which the PARP inhibitor olaparib is FDA-approved.

Taken together, we postulate that SNP rs1805405 causes increased expression of both PARP1-001 and PARP1-003 in patients with one or two variant alleles. Higher baseline PARP1 expression may explain the resistant phenotype in patients in the absence of a PARPi, since higher expression of PARP1 is correlated with worse outcome (Goncalves et al. 2011). PARP trapping is not relevant in this case, since PARP1-001 is expressed at much higher level than

PARP1-003 (TCGA pre-treatment results), and a small fraction of PARP1-003 does not mitigate the resistant phenotype observed. However, PARP1-003 expression becomes important when PARPi is used, as the PARP1-001 will be trapped on the DNA and PARP1-003 may augment this trapping. In summary, this study identified PARP1 SNP rs1805405 as a key biomarker for stratification of cancer patients in combination therapy (chemotherapy + PARP inhibition). We did so by applying a new computational approach for data integration and analysis to a cohort of metastatic melanoma patients. We validated, in vitro, the biologic impact of rs1805407 on the outcome of PARP1 inhibition in combination with alkylating agents and generated testable hypotheses that will further solidify its role in patient selection. The potential limitations of the paper and future directions are twofold. One is that the current manuscript contains mostly *in vitro* confirmation of our findings from our melanoma patient cohort and TCGA data for melanoma and ovarian cancer. We plan to apply this methodology on prospective clinical trials that have been conducted with PARP inhibitors. Second, the details of the potential mechanism of action of this SNP need further investigation.

The accurate prediction of therapy outcomes based on the molecular characteristics of tumors may alter the current landscape of cancer therapy given that immunotherapy results in substantial objective responses only in subsets of patients. One can therefore expect that the accurate patient stratification into therapy-response categories may allow the overall percentage of disease control to soar utilizing the current therapeutic armamentarium. In addition, insights into the molecular mechanism that characterizes a therapy-resistant phenotype may usher in new strategies to overcome therapy resistance.

5.0 PATHWAY BASED DATA INTEGRATION

Up until this point, this dissertation has focused on *de novo* network reconstruction methods that attempt to learn a network from a given dataset without any prior knowledge of the relationships between the variables. Essentially, we have been fitting a network structure to the data. We have drawn from the large amount of available biological knowledge to validate the network predictions generated by these methods, but not to learn the networks themselves. In this chapter we will present new work on a well established method that comes from the opposite direction in that it relies on a network structure curated from biological data and fits the data to this network. The algorithm we work with here, PARADIGM(Vaske et al. 2010; Sedgewick et al. 2013), also fits with the overall theme of this dissertation in that it is designed to integrate multiple data sources in order to accurately model the protein and pathway activity in the cell. PARADIGM is a popular algorithm for studying cancer data and is well established for its ability to integrate genome-wide DNA copy number and mRNA data. In addition since PARADIGM constructs its network from user defined “dogmas” (as in the central dogma of molecular biology) and network knowledge files, it is in principal easy to extend to more data types. In this chapter we will describe our work on adding another data type commonly used in cancer research, microRNA (miRNA) to the PARADIGM model. Our new model is able to recover miRNA markers in breast cancer tumors that are well established in the literature, which suggests that it will also be useful for detecting new miRNA markers to assist with diagnosis and treatment of cancer.

5.1 CHAPTER SUMMARY

MicroRNAs play an important role in regulation of gene expression, and are known biomarkers for breast cancer as well as other malignancies. PARADIGM is a pathway based algorithm that allows for integration of multiple genomic data types with a curated pathway database to make pathway activity predictions. We added a model of gene silencing due to miRNA to the PARADIGM algorithm in order to study miRNA expression in a pathway context. We curated a set of 7751 miRNA-mRNA interactions from the union of 3 target prediction algorithms. These interactions involved 66 miRNA and 2814 mRNA transcripts. We ran our model on copy number, RNAseq and miRNAseq data from 697 patients in the TCGA breast cancer cohort, and studied changes in the learned interactions between active miRNAs and their targets between different subtypes. The miRNA-target pairs with the largest correlation changes between Basal and Luminal A subtypes were enriched for known oncogenes, and for miRNAs and genes related to the activity of miRNAs in cancer. In addition these targets are involved in a number of relevant signaling pathways including PI3K-AKT, JAK-STAT, RAP1 and RAS. Most of these highly differential links involved the miR-16 family of miRNAs which are known tumor suppressors. Two miRNA-mRNA target pairs showed the largest changes in link strength of any pathway links between Basal and Luminal A groups. The miRNAs in these pairs, miR-195 and miR-221, are both previously documented markers in breast cancer. By looking at changes in miRNA-target links between tumor subtypes, our extension to PARADIGM allowed us to identify both miRNAs and target genes involved in pathways relevant to breast cancer.

5.2 BACKGROUND

MicroRNA, or miRNA, are short (18-25 nucleotide) non-coding RNA molecules that target mRNA transcripts and silence genes via a variety of mechanisms. Gene silencing due to miRNA targeting plays a part in many biological processes, and miRNA target sequences have been predicted in as many as 30% of genes (Lewis et al. 2005). Dysregulation of miRNAs has been linked to a variety of human diseases including pulmonary fibrosis (Pandit et al. 2010), and atherosclerosis (Toba et al. 2014). miRNAs have been studied extensively in cancer, and many reviews of their role are available (Melo and Esteller 2011; Jansson and Lund 2012; Malumbres 2013; Ohtsuka et al. 2015). In order to post-transcriptionally silence genes, miRNAs associate with several proteins to form an RNA Induced Silencing Complex (RISC) that carries out the biological process that leads to silencing. While the membership of RISC can vary depending on the organism and context, the minimal component proteins in humans are the RNAase Dicer, which processes the miRNA transcripts into a mature form, the Argonaute family of proteins, the catalytic component of RISC, and TRBP, which recruits the Argonaute proteins to Dicer and the bound miRNA molecule (Chendrimada et al. 2005).

A key challenge for working with miRNA that is applicable to this study is how to identify the mRNA targeted by each miRNA. Due to the difficulty of experimental verification of miRNA-mRNA targeting, there are relatively few validated targets. Instead, a variety of methods exist to predict targeting based on factors such as sequence, binding energy and conservation. Often the amount of overlap between these various methods is low relative to the number of predicted targets, so a common approach is to use the union of several of these methods to find a set of high-confidence target predictions (Huang et al. 2011; Coronnello and Benos 2013). We follow the approach of mirConnX (Huang et al. 2011), and use the union of 3 popular methods: TargetScan (Friedman et al. 2009), miRanda (Enright et al. 2004), and PicTar (Krek et al. 2005).

A number of previous studies have attempted to combine miRNA target predictions with either pathway data (Lu et al. 2012), mRNA expression (Zhang et al. 2014), or both (Huang et al. 2011). MirSystem (Lu et al. 2012) links miRNA to pathway knowledge via their mRNA targets, and performs enrichment tests to determine which pathways are likely to be regulated by a given group of miRNAs. Zhang *et al* (Zhang et al. 2014) use causal learning methods combined with matched miRNA and mRNA data to predict miRNA activity in a condition specific manner. MirConnX (Huang et al. 2011) combines matched miRNA and mRNA data with target predictions and transcription factor regulation data to find condition specific regulatory networks. PARADIGM offers several advantages over these methods. First, while a number of these methods offer condition specific models, PARADIGM is able model patient-specific pathway activities, which allow for more flexible downstream analyses. In addition these methods study paired miRNA and mRNA data by looking at pairwise correlations between the miRNA-target pairs, while PARADIGM allows us to study these interactions using predictions of active miRNA silencing complexes. Thus, if proteins essential to the silencing pathway such as Argonaute or DICER are not active in the sample, PARADIGM will predict less miRNA regulation in that sample.

PARADIGM builds a factor graph out of a curated database of pathways in order to infer the unobserved levels of activity of individual proteins, protein complexes and families from observed DNA and mRNA data. The observed data is discretized to three levels corresponding to high, low and normal. For every protein in the PARADIGM pathway, a model of the central dogma of molecular biology (see **Figure 5.1**) is included in the factor graph. Each step in the dogma has an unobserved node in the graph: *DNA*, *RNA*, *protein*, and *active* (for activated protein). Each of these latent nodes is linked to observed data, if available, and to the *active* nodes of other genes that are annotated as regulators in the pathway database. The states of the latent nodes are then inferred from the data using loopy-belief propagation to perform Expectation-Maximization.

Although the relationships between the variables in PARADIGM are set, the parameters of the factors, which model the relationships between the nodes they connect, are learned by the algorithm. In our previous work with PARADIGM (Sedgewick et al. 2013), we added a model of regulation that allows the algorithm to learn parameters that describe the regulatory relationship between active proteins and the transcription, translation, or activation of the proteins that they regulate. So, although it is not possible to learn new edges with PARADIGM, by looking at the regulation parameters learned from the observed data, we can measure how strong edge is in a given set of samples.

5.3 METHODS

5.3.1 Pathway Model

miRNA is included in the PARADIGM model using the same dogma that coding RNAs use. The only dogma node that doesn't apply to miRNA is the *protein* node, and since there are not translational or activation regulators for the miRNA in our pathway, the *active* node will have the same state as the *RNA* node for a miRNA with high probability.

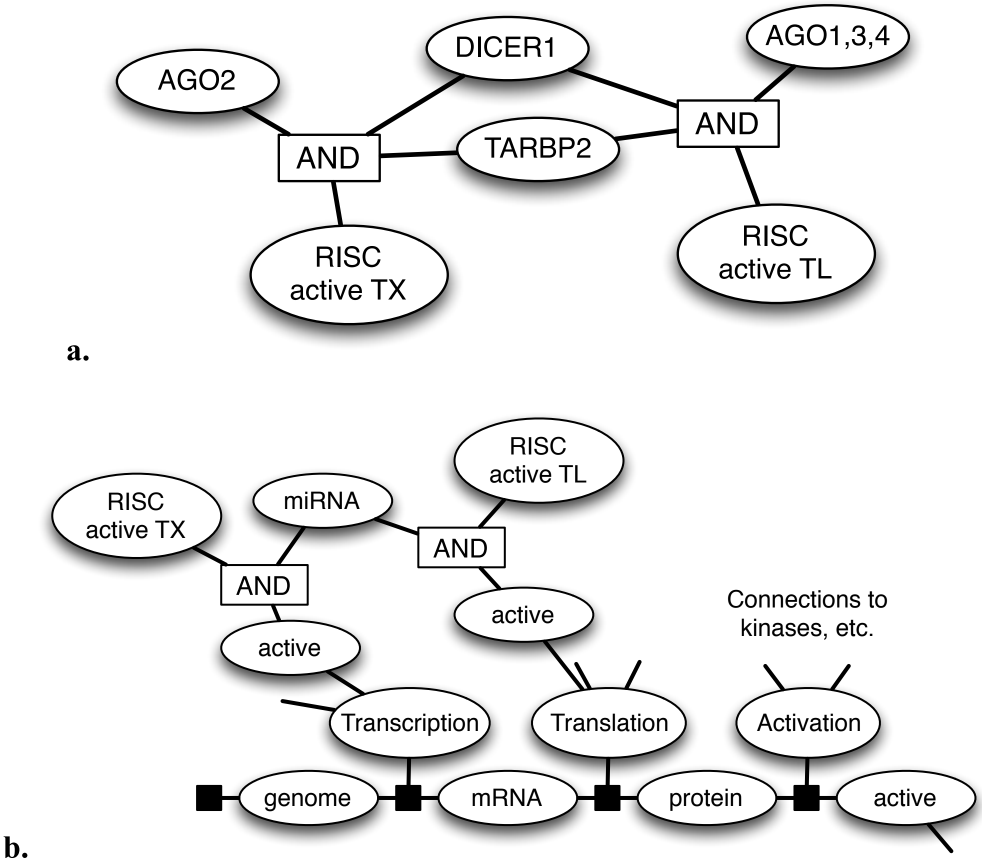


Figure 5.1 a Minimal RNA Induced Silencing Complex model, separated for transcriptional (TX) and translational (TL) regulation. **b** Model of how the active RISC-miRNA complexes interact with a protein dogma model.

Our RISC model uses the built-in complex model in PARADIGM, which is a "noisy AND" function. In other words, the predicted activity state of the complex is the minimum of the states of all the components of the complex with high probability, or another state with small error probabilities. **Figure 5.1a** shows our RISC model, which is separated by the putative regulation mechanisms of the different proteins in the Argonaute family. Argonaute 2 (AGO2) is part of a complex that regulates transcription because of its endoribunuclease activity that allows it to cleave mRNA molecules thereby silencing them (Kobayashi and Tomari 2016). Although this process happens post-transcription, kinetic studies of cleavage by AGO2 suggest that it happens rapidly enough that it will affect observed mRNA transcript levels (Ameres et al. 2007).

We treat the rest of the Argonaute family as translational regulators because their alternative silencing mechanisms are less likely to affect the observed mRNA transcript levels. These mechanisms include translation regulation activity such as direct translational repression via recruitment of additional factors and deadenylation of the poly(A) tail of the mRNA molecule, which in turn inhibits translation (Kobayashi and Tomari 2016). These different regulation models interact with the regulation nodes of a predicted target protein as shown in **Figure 5.1b**.

We compare 2 models in this study: the full model with both transcriptional and translational repression by RISC as presented in **Figure 5.1b**, and a simpler model that only adds the transcriptional regulation component corresponding to mRNA cleavage by AGO2. We work with the simpler model in case the full model gives too much weight to miRNA silencing. Given that without miRNA silencing there are 11,120 regulatory interactions in the pathway, none of which are translational regulators, it seems possible that adding 7,751 miRNA interactions as both transcriptional and translational repressors may drown out the signal of the rest of the pathway.

5.3.2 miRNA Target Predictions

We use intersection of miRNA-mRNA target predictions from 3 miRNA target prediction algorithms: TargetScan (Enright et al. 2004), miRANDA (Friedman et al. 2009), and Pictar (Krek et al. 2005). This database of targets comes from mirConnX (Huang et al. 2011). This procedure generated 7751 miRNA-mRNA interactions involving 66 miRNA and 2814 mRNA transcripts.

5.3.3 TCGA Breast Cancer

We used matched RNAseq, miRNAseq and DNA copy number data for 697 patients from the TCGA Breast Cancer Cohort. For the DNA copy number data we used GISTIC 2.0 predictions (Center 2016). To normalize the RNAseq data, we removed transcripts with zero reads in more than 50% of samples, log-scaled TPM values and median normalized each transcript across all samples. For miRNA normalization we filtered miRNAs with zeros reads in more than 75% of samples then log scaled the raw counts and median normalized each miRNA across all samples.

For validation of our PARADIGM model, we also use Reverse Phase Protein Array (RPPA), hormone receptor status from immunohistochemistry, survival and PAM50 subtype predictions for these patients.

5.3.4 PARADIGM Application Tests

To score how well a learned PARADIGM model matches the underlying biology of a sample, we use a battery of previously developed application tests. One test compares the pathway activities of the Estrogen Receptor gene ESR1 as well as the dimer ERα to the ER status determined by immunohistochemistry (IHC). Similarly, HER2 status from IHC is compared to the predicted activity of the corresponding gene, ERBB2. For another validation test, we compare protein data from RPPA to the PARADIGM predicted protein activities. A new test added in this study looks at the correlation of pathway activities of PLK1 and one of its transcriptional regulators, FOXM1. PLK1 is a hub in the PARADIGM network and involved in several feedback loops, one of which involves FOXM1, and we have found that a strong correlation between these closely linked genes is indicative of a good model fit.

Another dimension of testing is survival prediction. For this test, we filter out censored patients (i.e. alive at last checkup) and split the cohort of patients in to quartiles based on survival times. The top 25% of patients with the longest survival times are treated as the positive class, and the bottom 25% are used as the negative class. We performed this experiment on both the whole breast cancer cohort (15 patients of each class) and only the estrogen receptor positive (ER+) patients (9 patients of each class. We fit a linear SVM (from scikit-learn: http://scikit-learn.org/stable/index.html) on these labels (positive samples labeled as +1, negative as -1) using the Integrated Pathway Activities (IPA) produced by PARADIGM. We measure the classification accuracy using leave one out cross-validation.

5.3.5 Functional Enrichment

To asses the biological relevance of groups of genes we use standard gene set enrichment analysis methodology to test which pathways or gene annotations are over-represented in our set of genes. Specifically we use the ‘kegga’ and ‘goana’ functions built into the R package, limma (Ritchie et al. 2015).

5.4 RESULTS AND DISCUSSION

5.4.1 Distribution of miRNA-target Links

We compared the distributions of the PARADIGM link parameter correlations for models learned with both the transcription only model (**Figure 5.1a**), and the full model (**Figure 5.1b**).

Correlations are calculated from the parameters (which are essentially conditional probability tables, CPTs, see (Sedgewick et al. 2013) for more detail) that connect the "active" node of the regulating protein to one of the "regulation" nodes (either transcriptional, translation, or protein activation as shown in **Figure 5.1b**). The distributions of the different types of regulatory links are essentially the same between the models other than the additional miRNA translation regulation links added in the full model. This indicates that the translation links don't seem to have a noticeable affect on other links in the pathway.

All links are started from the same set of initial parameters which are either positive or negatively correlated based on whether the link is annotated as activation or inhibition in the pathway (specifically, a probability of .8 on the diagonal of the CPT and .1 off the diagonal). From **Figure 5.2** we can see that the miRNA translation regulation links do not seem to stray very far from the initial inhibition parameter setting, while the other link types spread nicely to match the data, with links not supported by the data drifting towards 0 and links that are consistent with the data spreading towards the positive and negative extremes. Possible explanations for this lack of learning include the fact

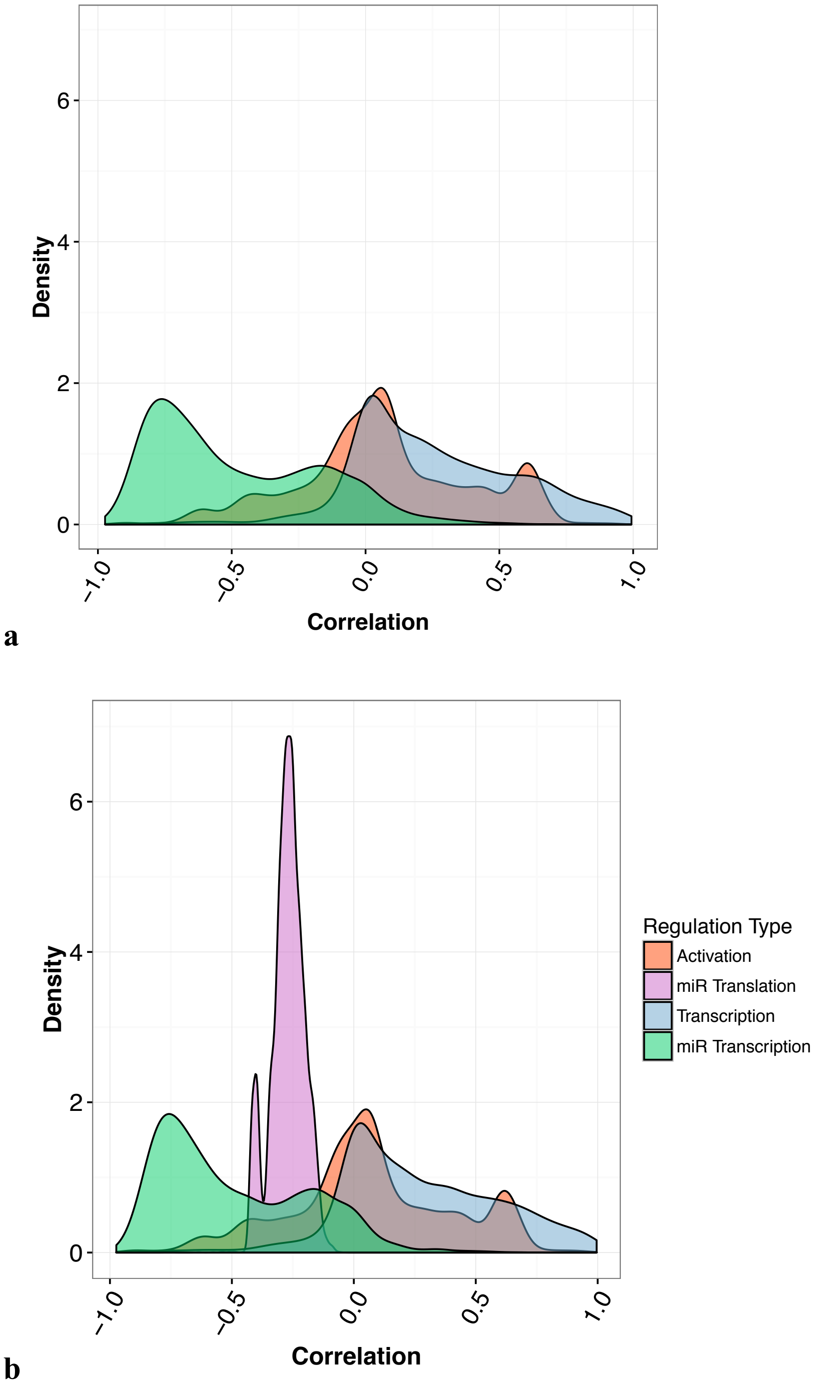


Figure 5.2 Density plots of Pearson's correlation of regulation links from parameters learned with PARADIGM using **a** the transcription only RISC model and **b** the full transcription and translation RISC model.

that these links are the only translational regulation links in the pathway, so they are not balanced by any activating translational regulation links, and secondly, the translation regulation nodes may be receiving less signal from the data because of their spot on the pathway between the *mRNA* node which is directly linked to the observed data and the *protein* node which is not. Activation regulation links are similarly positioned some distance from the data nodes in the pathway, but they seem to better able to conform to the data, perhaps because of the connections of the *active* nodes to the regulation nodes of other proteins and complexes.

5.4.2 Survival Prediction

To see how well the Integrated Pathway Activities (IPAs) predicted by the different PARADIGM models represent the underlying biology of the tumors, we studied how well they are able to predict patient survival. We treat this task as a classification problem where the two classes are patients in the top quartile or the bottom quartile of survivals. Due to incomplete survival data for many patients, this left us with a set of 30 patients, 15 high survival and 15 low survival. The miRNA transcription regulation model performed the best, an SVM trained on IPAs from this model achieved a leave-one-out cross validation accuracy of 60% while the full model achieved poor accuracy of 43%, as did a model learned without any miRNA data, 37% accuracy. The performance of the simpler model is comparable to doing the classification with RNAseq data (59% accuracy) or RNAseq and miRNAseq data together (62% accuracy).

5.4.3 Correlation of IPAs with Protein and IHC Data

Another method for validating our models is to compare to other data types. We compare the IPAs from each model for ESR1 and the ERα homodimer to compare to estrogen receptor status as measured by IHC, and for ERBB2 to compare to IHC measured HER2 status. The IHC experiment gives a call of positive or negative for each hormone receptor, so we performed a two sample ranksum test of the IPAs for the positive versus negative groups of the corresponding hormone receptor. All three models performed well on these tests. The full miRNA model had highly significant p-values from the tests: 2.9e-48 for ESR1, 2.9e-47 for the ERα homodimer, and 1.2e-9 for ERBB2. The transcription-only miRNA model has slightly lower p-values: 1.4e-49 for ESR1, 7.9e-48 for ERα homodimer, and 2.8e-11 for ERBB2. The original PARADIGM model without any miRNA data had the lowest p-values for ESR1 (8.5e-50) and ERBB2 (9.1e-12), but the highest for ERα homodimer (9e-46).

We compared the IPAs from each model to protein concentrations from 173 proteins in the TCGA breast cancer samples measured with RPPA. For each protein we measured the correlation of the RPPA data with the IPAs across all patients with RPPA measurements. We then averaged these correlations. There was no separation between the 3 models for this test. All three models had an average of Spearman correlations of .24 between the RPPA and IPAs with standard errors between .017 and .018. This result is somewhat surprising since 45 of the 173 proteins found in the RPPA data are targets of the miRNAs that we added to PARADIGM. It is possible, however, that these proteins are strongly regulated in the original PARADIGM network so that adding miRNA did not alter their IPAs noticeably.

5.4.4 Top miRNA-target Links

Although the battery of tests in the previous section does not clearly separate either of the miRNA regulation models, we now choose to focus on the links learned by the transcription-only model because it performed better in all tests and because **Figure 5.2** shows that the transcription regulation links are likely to have more informative parameters. In this section we investigate how our miRNA-target links change between breast cancer subtypes, specifically we compare the 97 patients with the aggressive basal tumors to 288 patients with more treatable luminal A tumors.

We sorted miRNA-target links by the largest change in correlation between the basal and luminal A subgroups. Of the top 10 links with large correlation changes between the groups, 9 of them involve miR-16. This is likely due to the very low IPA of miR-16 in basal tumors (median -4.0) compared to luminal A tumors (median 0) (Wilcoxon $p < 2e-16$). miR-16 is a known tumor suppressor that has been characterized in a variety of cancers including lymphoma, leukemia and breast cancer (Aqeilan et al. 2010; Rivas et al. 2012). The targets of the top 200 links by correlation change are significantly enriched (false discovery rate < .05) for a number of pathways relevant to cancer, shown in **Table 5.1**. While the majority of these pathways are cancer related, the fact the we recovered the "MicroRNAs in cancer" pathway as an enrichment for targets, not just the miRNAs gives us a good verification that we have found relevant links.

Table 5.1 KEGG enrichment for the gene targets of the top 200 miRNA-target links by correlation change between basal and luminal A breast cancer subgroups.

Pathway	Pathway Size	Number Found	FDR
Jak-STAT signaling pathway	32	8	3.499e-06
Rap1 signaling pathway	70	10	1.506e-05
PI3K-Akt signaling pathway	95	9	2.470e-03
MicroRNAs in cancer	78	8	4.457e-03
Melanoma	23	5	4.649e-03
Pathways in cancer	132	10	5.570e-03
Focal adhesion	65	7	1.121e-02
Regulation of actin cytoskeleton	67	7	1.362e-02
Endocytosis	70	7	1.806e-02
Ras signaling pathway	73	7	2.360e-02
HTLV-I infection	73	7	2.360e-02
Proteoglycans in cancer	73	7	2.360e-02
Wnt signaling pathway	55	6	3.732e-02
Transcriptional misregulation in cancer	57	6	4.546e-02

In addition to correlation, we commonly use a G-test to measure the statistical dependence of the variables, which we refer to as link “strength”. The G-test allows us to uncover links that are highly dependent, but do not necessarily have a linear relationship that can be captured by Pearson’s correlation. Looking at the rank difference of G-test p-values between the basal and luminal groups, reveals that two miRNA regulation links have the largest change out of all links in the pathway. miR-221-ARF4 shows a strong connection in the luminal A subgroup (FDR = 9.9e-7), but a relatively weaker relationship in basal tumors (FDR = 1.5e-3). Both nodes in this link have been previously linked to breast cancer: overexpression of miR-221 is linked to aggressive, basal tumors through promotion of epithelial-to-mesenchymal transition (Shah and Calin 2011) and ARF4 expression is linked to cell migration and metastasis in breast cancer (Jang et al. 2012). Similarly, miR-195-BDNF has a strong silencing relationship in luminal A tumors (FDR = 5.6e-21) that is weaker in basal patients (FDR = 3.2e-3). miR-195 has been identified as a potential circulating biomarker to diagnose breast cancer (Heneghan et al. 2010) and BDNF is a growth factor that has been shown to promote tumor growth and proliferation in colon cancer (Yang et al. 2013).

5.5 CONCLUSIONS

By adding miRNAs, miRNA target predictions, and a model of the RNA induced silencing complex to PARADIGM, we were able to create a model that can interrogate miRNA induced gene silencing in a pathway context. Based on our comparison between a transcription regulation only model to a RISC model that regulates genes at both the transcriptional and translational level, we find that our model is better able to learn miRNA-target links at the transcriptional regulation level. By comparing differential miRNA silencing in tumor subgroups we identified miR-221, miR-195 and miR-16 as important regulators in breast cancer. All of these miRNAs had been previously studied in breast cancer, and the genes they targeted proved to be enriched for cancer related pathways as well. Thus the predictions made by our model had strong support in the literature.

6.0 CONCLUSIONS AND FUTURE WORK

In this dissertation, we have presented a variety of network based data integration methods. On one end of the spectrum we have developed algorithms that attempt to reconstruct causal networks *de novo*, on the other end, we extended an algorithm for pathway based inference of protein activity. We have shown that these types of integrative methods are powerful tools that we believe are essential for modern biomedical research. Using our methods we were able to identify a prognostic biomarker for response to treatment temozolomide in metastatic melanoma. This marker also suggests that a combination treatment strategy that may help treat non-responders more effectively. In addition, by adding miRNA to PARADIGM, we were able to effectively study changes in gene silencing that differentiate more or less aggressive breast cancer subtypes. As collecting multimodal data from patient samples becomes the standard of care in cancer and other diseases, we expect integrative analyses like ours to continue to assist in the understanding, diagnosis and treatment of human disease.

There are many opportunities for future development of the methods described in this thesis. For the MGM described in chapter one, a pressing challenge is to make the learning procedure efficient enough to handle genome scale data. This could be achieved by using a different optimization algorithm, or by switching to a learning method that uses separate regressions rather than optimizing the pseudolikelihood. A model based on separate regression would have more flexibility in distributional assumptions and be easy to parallelize. In both

MGM and the hybrid MGM-causal search methods, more work needs to be done towards encoding prior knowledge in the model. It is simple to force a model to always include or exclude edges based on knowledge, but a more nuanced approach is necessary for biological pathway knowledge where an edge may only exist in certain cell types under certain conditions. Another exciting direction for causal search over mixed data is to develop an efficient scoring method to use with score-based causal search methods. Recent work with score based methods on continuous data has been promising for application to very large datasets, so a mixed score may allow us to take advantage of these new methods. Finally, there are many more data types that could be added to the PARADIGM model including DNA methylation, ribosome profiling data, and protein measurements. As with miRNA, the challenges for any addition of new data types to the pathway model are to balance the added model complexity with potential gains in inference accuracy.

APPENDIX A

PERFORMANCE OF DIRECTED SEARCH ON LOW-DIMENSIONAL DATA

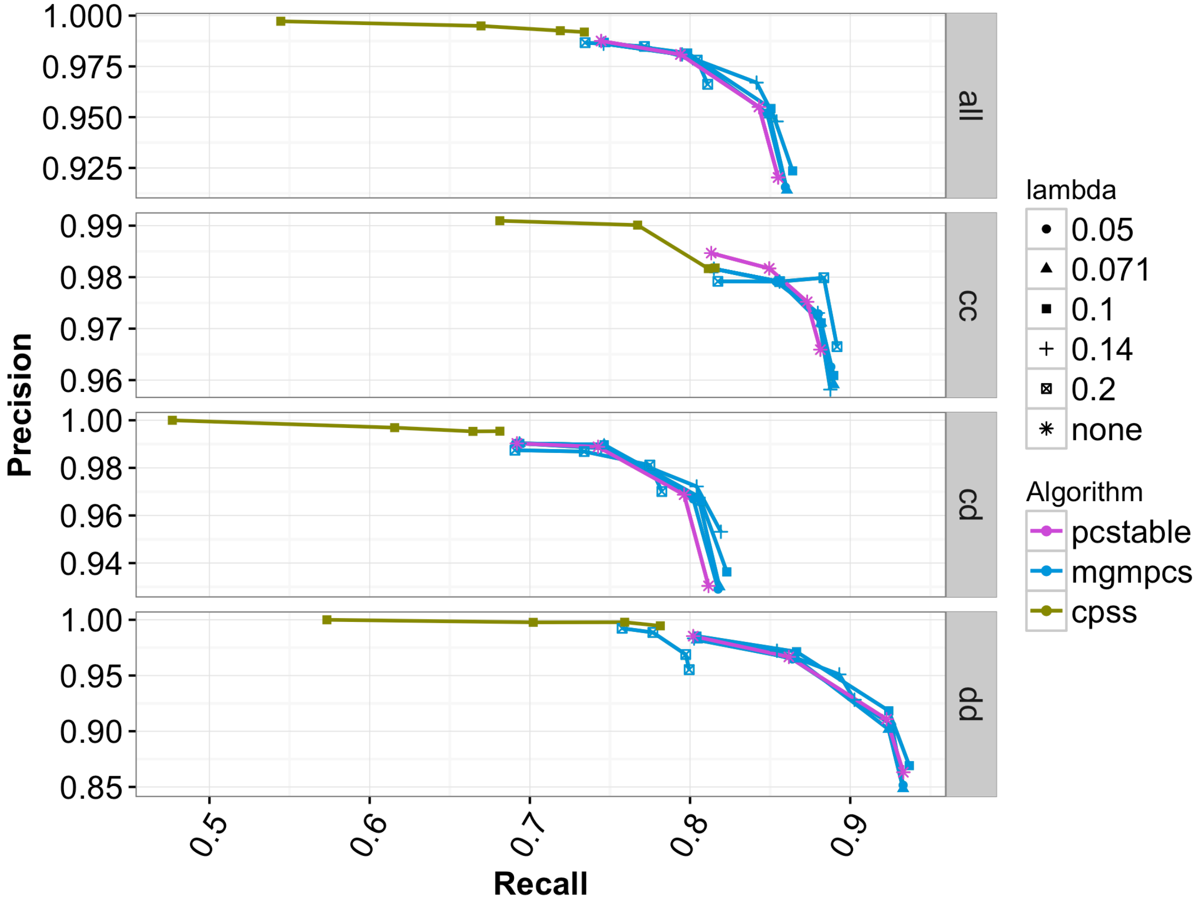


Figure A.2 Precision-Recall curves of edge adjacency recovery for $.05 \leq \lambda \leq\ .2$ and $.001 \leq \alpha \leq\ .1$. and the full range of algorithms and edge types.

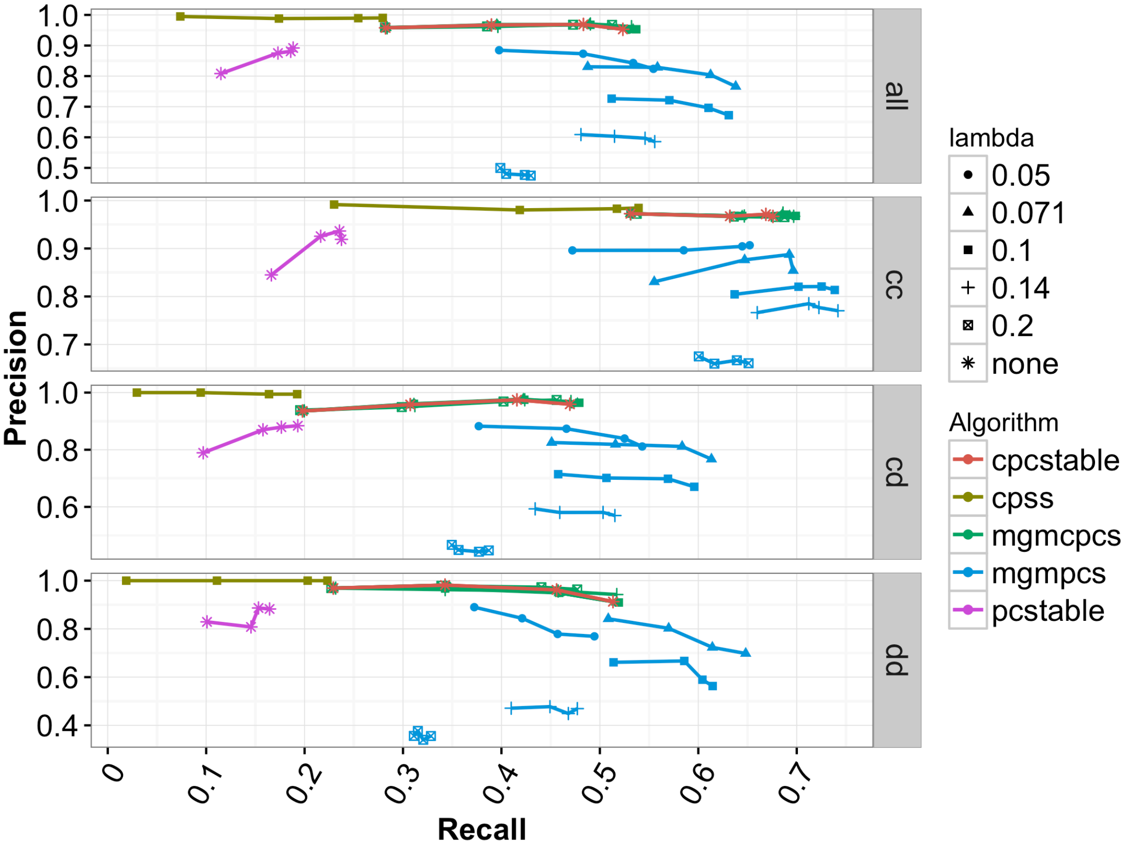


Figure A.2 Precision-Recall curves of edge direction recovery for $.05 \leq \lambda \leq\ .2$ and $.001 \leq \alpha \leq\ .1$. and the full range of algorithms and edge types.

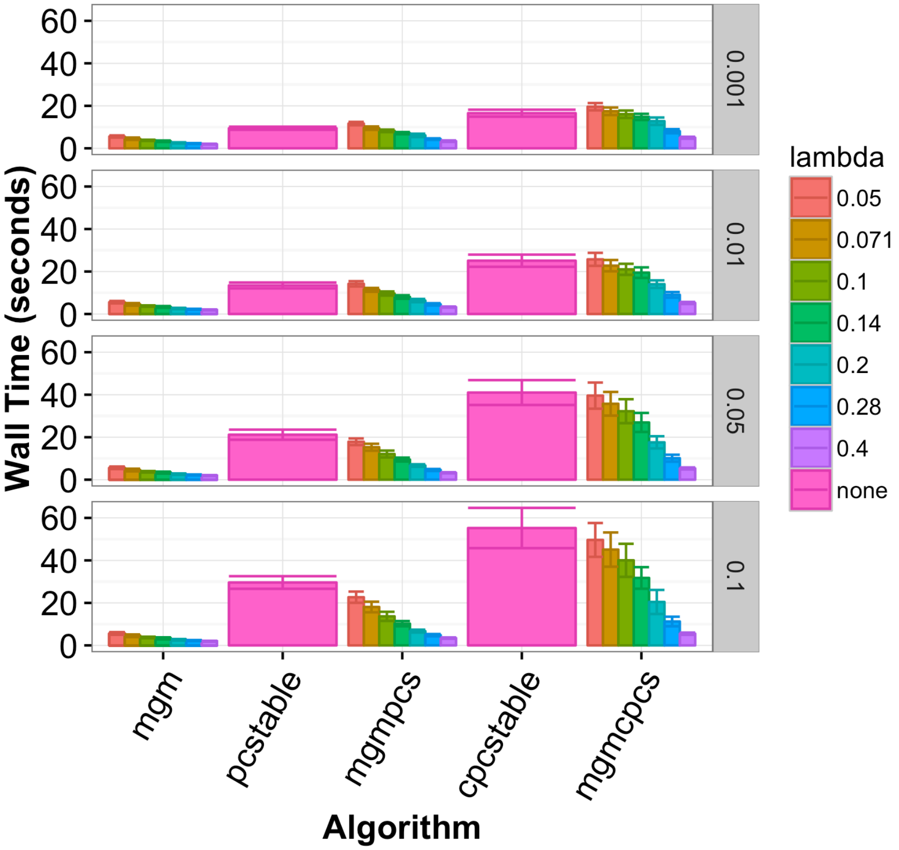


Figure A.3 Running times of search algorithms on low dimensional data. Directed search steps were run in parallel on a 4 core laptop.

Table A.1 Parameter settings with the best SHD performance by edge type in low-dimensional data set

Algorithm	α	λ	*Type*	*SHD*
PC-Stable	0.05	none	all	95.35 (2.4786)
	0.05	none	cc	20.70 (1.6481)
	0.05	none	cd	52.45 (1.5139)
	0.05	none	dd	22.20 (1.2599)
MGM-PCS	0.05	0.071	all	63.25 (2.8385)
	0.1	0.1	cc	11.40 (1.3638)
	0.05	0.071	cd	35.50 (1.9310)
	0.01	0.071	dd	14.75 (1.4343)
CPC-Stable	0.1	none	all	67.05 (2.7772)
	0.05	none	cc	11.20 (1.4190)
	0.1	none	cd	39.45 (2.1502)
	0.05	none	dd	15.80 (1.3111)
MGM-CPCS	0.1	0.14	all	63.35 (2.7513)
	0.1	0.2	cc	10.75 (1.3116)
	0.1	0.14	cd	37.70 (2.0877)
	0.1	0.14	dd	14.70 (1.2011)

APPENDIX B

SUPPLEMENTARY TABLES FOR PARP1 CASE STUDY

Table B.1 PARP1 SNPs in LD with rs1805407.

SNP	Distance	R^2	D’	Chr	Coord_hg18	GeneVariant
rs3219031	437	1	1	chr1	224656019	INTRONIC
rs3219027	580	1	1	chr1	224657036	INTRONIC
rs6701634	1792	1	1	chr1	224654664	INTRONIC
rs3754370	2445	1	1	chr1	224658901	INTRONIC
rs3768347	2912	1	1	chr1	224659368	INTRONIC
rs3768346	3021	1	1	chr1	224659477	INTRONIC
rs7522351	3435	1	1	chr1	224659891	INTRONIC
rs7525191	3438	1	1	chr1	224659894	INTRONIC
rs4653732	4273	1	1	chr1	224660729	INTRONIC
rs10799349	4317	1	1	chr1	224652139	INTRONIC
rs7542788	4530	1	1	chr1	224651926	INTRONIC
rs7548007	4553	1	1	chr1	224651903	INTRONIC
rs4653733	4780	1	1	chr1	224661236	INTRONIC
rs60698376	5024	1	1	chr1	224661480	N/A
rs4653731	5861	1	1	chr1	224650595	INTRONIC
rs2077197	6206	1	1	chr1	224662662	UPSTREAM
rs12240196	6350	1	1	chr1	224650106	INTRONIC
rs59672299	7760	1	1	chr1	224664216	N/A
rs1073991	8759	1	1	chr1	224647697	INTRONIC
rs2136876	8880	1	1	chr1	224647576	INTRONIC
rs1000033	9446	1	1	chr1	224647010	INTRONIC
rs6665208	9541	1	1	chr1	224665997	UPSTREAM
rs1002153	9646	1	1	chr1	224646810	INTRONIC
rs2280712	9740	1	1	chr1	224646716	INTRONIC
rs1805405	9812	1	1	chr1	224646644	SPLICE_SITE, INTRONIC

rs6679573	11114	1	1	chr1	224667570	INTERGENIC
rs10915987	11848	1	1	chr1	224668304	INTERGENIC
rs3219043	12239	1	1	chr1	224644217	INTRONIC
rs77173384	12382	1	1	chr1	224668838	N/A
rs28407557	12564	1	1	chr1	224669020	INTERGENIC
rs4653445	12927	1	1	chr1	224643529	INTRONIC
rs2293464	13537	1	1	chr1	224642919	INTRONIC
rs12068460	13912	1	1	chr1	224670368	INTERGENIC
rs3219053	15279	1	1	chr1	224641177	INTRONIC
rs1805408	16431	1	1	chr1	224640025	INTRONIC
rs3219058	17039	1	1	chr1	224639417	INTRONIC
rs6681537	19603	1	1	chr1	224676059	INTERGENIC
rs3219073	20458	1	1	chr1	224635998	INTRONIC
rs2271343	22270	1	1	chr1	224634186	INTRONIC
rs732284	22825	1	1	chr1	224633631	INTRONIC
rs3219115	32892	1	1	chr1	224623564	INTRONIC
rs752308	38327	1	1	chr1	224618129	INTRONIC
rs747658	38655	1	1	chr1	224617801	INTRONIC
rs747659	39092	1	1	chr1	224617364	INTRONIC
rs6664761	39642	1	1	chr1	224616814	INTRONIC
rs2282400	42834	1	1	chr1	224613622	DOWNSTREAM
rs6675427	45851	1	1	chr1	224610605	DOWNSTREAM
rs6675327	45924	1	1	chr1	224610532	DOWNSTREAM
rs6661762	46142	1	1	chr1	224610314	DOWNSTREAM
rs1991865	48782	1	1	chr1	224607674	INTERGENIC
rs12092726	50806	1	1	chr1	224605650	INTERGENIC
rs3219023	1223	0.947	1	chr1	224657679	INTRONIC
rs7531668	6186	0.945	1	chr1	224662642	UPSTREAM
rs12025487	15060	0.945	1	chr1	224671516	INTERGENIC
rs1109032	28430	0.945	1	chr1	224628026	INTRONIC
rs3754375	28768	0.945	1	chr1	224627688	INTRONIC
rs4653735	10521	0.891	1	chr1	224666977	UPSTREAM
rs878367	52311	0.891	1	chr1	224604145	INTERGENIC
rs7527192	6246	0.838	1	chr1	224662702	UPSTREAM

Table B.2 Potentiation of response to chemotherapy or radiation combined with PARP inhibition (from literature). PARP1 SNP rs1805407 genotyping analysis of a panel of human cancer cell lines. All six of the cell lines reported in the literature to be "resistant" to chemotherapy + PARPi combination treatment were WT for the rs1805407 locus. Four out of the seven cell lines reported to be "sensitive" had at least one copy of C in this locus. Cell line was considered "sensitive" when chemopotentiation ratio was ≥ 2. *S:* sensitive; *R:* resistant.

Cell line	**Tumor type**	**Response**	**Tumor type**	**rs1805407 Genotype**	**Therapy type**	**PARPi agent**	**REFS**
LoVo	Colon	R	Colon	T/T	TMZ	NU1025/NU1085, AG14361	(Delaney et al. 2000b; Calabrese et al. 2004b)
SW620	Colorectal	R	Colorectal	T/T	Irinotecan	ABT-888	(Davidson et al. 2013)
H522	Lung	R	Lung	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
HT-29	Colon	R	Colon	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
SKOV-3	Ovarian	R	Ovarian	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
LS174T	Colon	S	Colon	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
HCT-116	Colon	S	Colon	T/T	Irinotecan	ABT-888	(Davidson et al. 2013)
MDA-MB-231	Breast	R	Breast	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
MCF-7	Breast	S	Breast	T/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)
Calu-6	Lung	S	Lung	C/T	TMZ	CEP-6800	(Miknyoczki et al. 2003)
M14	Melanoma	S	Melanoma	C/T	TMZ	3-aminobenzamide	(Tentori et al. 2003)
A549	Lung	S	Lung	C/T	TMZ	NU1025/NU1085, AG1436	(Delaney et al. 2000b; Calabrese et al. 2004b)
A2780	Ovarian	S	Ovarian	C/T	TMZ	NU1025/NU1085	(Delaney et al. 2000b)

Table B.3 Results from MMS treatment of cell lines with and without PARP1 inhibitor (ABT-888 or olaparib). The data from the MTT assays were expressed as mean ± standard deviation (SD). The ratio between the IC_{50} means of MMS treatment alone and in combination with ABT-888 or olaparib was calculated for each cell line. A Potentiation factor (ratio) ≤ 1 indicates no chemo-potentiation.

ABT-888 (10 nM):

Cell line	Tissue origin	PARP1/SNP genotype	MMS IC_{50} (µM)	MMS + ABT-888 IC_{50} (µM)	Potentiation factor	*p*-value:
FEMX	melanoma	T/T	166.3 (± 20.2)	176.0 (± 40.4)	0.945	0.626
A375	melanoma	T/T	306.0 (± 22.1)	283.3 (± 33.5)	1.080	0.172
H-522	lung	T/T	577.7 (± 56.8)	745.3 (± 68.6)	0.775	0.147
SW620	colon	T/T	299.4 (± 37.0)	449.4 (± 89.1)	0.666	0.047
MDA-MB-231	breast	T/T	287.2 (± 28.7)	303.2 (± 45.1)	0.947	0.530
M14	melanoma	C/T	520.8 (± 63.4)	359.8 (± 56.7)	1.447	0.005
A549	lung	C/T	254.8 (± 23.9)	143.9 (± 37.8)	1.771	0.002
A2780	ovarian	C/T	190.0 (± 41.0)	80.8 (± 14.7)	2.351	0.003
H460	lung	C/T	227.0 (± 21.4)	134.9 (± 20.5)	1.682	0.002

Olaparib (5 nM):

Cell line	Tissue origin	PARP1/SNP genotype	MMS IC_{50} (µM)	MMS + olaparib IC_{50} (µM)	Potentiation factor	*p*-value
SW620	colon	T/T	342.7 (±68.7)	357 (±55.6)	0.960	0.720
A2780	ovarian	C/T	182.4 (± 24.0)	39.2 (± 8.8)	4.651	0.017

BIBLIOGRAPHY

Abecassis I, Sedgewick AJ, Romkes M, Buch S, Nukui T, Kapetanaki MG, Kirkwood J, Benos PV, Tawbi H. 2015. PARP1 Variant Confers Sensitivity to PARP1 Inhibitors in Cancer Cells Suggesting an Improved Therapeutic Strategy. *Submitted*.

Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **26**(3): 392-398.

Akaike H. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199-213. Springer.

Altmeyer M, Messner S, Hassa PO, Fey M, Hottiger MO. 2009. Molecular mechanism of poly(ADP-ribosyl)ation by PARP1 and identification of lysine residues as ADP-ribose acceptor sites. *Nucleic Acids Res* **37**(11): 3723-3738.

Ameres SL, Martinez J, Schroeder R. 2007. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell* **130**(1): 101-112.

Aqeilan RI, Calin GA, Croce CM. 2010. miR-15a and miR-16-1 in cancer: discovery, function and future perspectives. *Cell Death Differ* **17**: 215-220.

Balkhi MY, Trivedi AK, Geletu M, Christopeit M, Bohlander SK, Behre HM, Behre G. 2006. Proteomics of acute myeloid leukaemia: Cytogenetic risk groups differ specifically in their proteome, interactome and post-translational protein modifications. *Oncogene* **25**(53): 7041-7058.

Barakat DJ, Zhang J, Barberi T, Denmeade SR, Friedman AD, Paz-Priel I. 2015. CCAAT/Enhancer binding protein beta controls androgen-deprivation-induced senescence in prostate cancer cells. *Oncogene* **34**(48): 5912-5922.

Becker SR, Candès EJ, Grant MC. 2011. Templates for convex cone problems with applications to sparse signal recovery. *Math Prog Comp* **3**(3): 165-218.

Besag J. 1975. Statistical analysis of non-lattice data. *The statistician*: 179-195.

Bollobás B, Borgs C, Chayes J, Riordan O. 2003. Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 132-139. Society for Industrial and Applied Mathematics.

Bøttcher SG. 2001. Learning Bayesian networks with mixed variables. In *Eighth International Workshop on Artificial Intelligence and Statistics*, pp. 149-156, Key West, Florida.

Bray N, Pimentel H, Melsted P, Pachter L. 2015. Near-optimal RNA-Seq quantification. *arXiv preprint arXiv:150502710*.

Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**(7051): 701-703.

Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, Kyle S, Meuth M, Curtin NJ, Helleday T. 2005. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**(7035): 913-917.

Calabrese CR, Almassy R, Barton S, Batey Ma, Calvert aH, Canan-Koch S, Durkacz BW, Hostomsky Z, Kumpf Ra, Kyle S et al. 2004a. Anticancer Chemosensitization and Radiosensitization by the Novel Poly(ADP-ribose) Polymerase-1 Inhibitor AG14361. *JNCI Journal of the National Cancer Institute* **96**: 56-67.

Calabrese CR, Almassy R, Barton S, Batey MA, Calvert AH, Canan-Koch S, Durkacz BW, Hostomsky Z, Kumpf RA, Kyle S et al. 2004b. Anticancer chemosensitization and radiosensitization by the novel poly(ADP-ribose) polymerase-1 inhibitor AG14361. *Journal of the National Cancer Institute* **96**(1): 56-67.

Caldeira JR, Prando EC, Quevedo FC, Neto FA, Rainho CA, Rogatto SR. 2006. CDH1 promoter hypermethylation and E-cadherin protein expression in infiltrating breast cancer. *BMC Cancer* **6**: 48.

Cancer Genome Atlas N. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418): 61-70.

Center BITGA. 2016. SNP6 Copy number analysis (GISTIC2). Broad Institute of MIT and Harvard.

Chen S, Witten D, Shojaie A. 2014. Selection and estimation for mixed graphical models. In *arXiv preprint arXiv:13110085v2 [statME]*.

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R. 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**(7051): 740-744.

Cheng J, Levina E, Zhu J. 2013. High-dimensional Mixed Graphical Models. *arXiv preprint arXiv:13042810*.

Chou TC, Talalay P. 1984. Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Advances in enzyme regulation* **22**: 27-55.

Cimino-Mathews A, Subhawong AP, Illei PB, Sharma R, Halushka MK, Vang R, Fetting JH, Park BH, Argani P. 2013. GATA3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Human pathology* **44**(7): 1341-1349.

Colombo D, Maathuis MH. 2014. Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research* **15**: 3741-3782.

Consortium GO. 2015. Gene ontology consortium: going forward. *Nucleic acids research* **43**(D1): D1049-D1056.

Coronnello C, Benos PV. 2013. ComiR: combinatorial microRNA target prediction tool. *Nucleic acids research*.

Davidson D, Wang Y, Aloyz R, Panasci L. 2013. The PARP inhibitor ABT-888 synergizes irinotecan treatment of colon cancer cell lines. *Investigational new drugs* **31**: 1-14.

Davies JR, Jewell R, Affleck P, Anic GM, Randerson-Moor J, Ozola A, Egan KM, Elliott F, Garcia-Casado Z, Hansson J et al. 2014. Inherited variation in the PARP1 gene and survival from melanoma. *Int J Cancer*.

de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M. 2012. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The Lancet Oncology* **13**(6): 607-615.

Delaney C, Wang L, Kyle S. 2000a. Potentiation of Temozolomide and Topotecan Growth Inhibition and Cytotoxicity by Novel Poly(adenosine Diphosphoribose) Polymerase Inhibitors in a Panel of Human Tumor Cell Lines. *Clinical cancer …*.

Delaney CA, Wang LZ, Kyle S, White AW, Calvert AH, Curtin NJ, Durkacz BW, Hostomsky Z, Newell DR. 2000b. Potentiation of temozolomide and topotecan growth inhibition and cytotoxicity by novel poly(adenosine diphosphoribose) polymerase inhibitors in a panel of human tumor cell lines. *Clinical cancer research : an official journal of the American Association for Cancer Research* **6**(7): 2860-2867.

Efron B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM.

El-Khamisy SF, Masutani M, Suzuki H, Caldecott KW. 2003. A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. *Nucleic Acids Res* **31**(19): 5526-5533.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2004. MicroRNA targets in Drosophila. *Genome biology* **5**(1): R1-R1.

Eustermann S, Videler H, Yang JC, Cole PT, Gruszka D, Veprintsev D, Neuhaus D. 2011. The DNA-binding domain of human PARP-1 interacts with DNA single-strand breaks as a monomer through its second zinc finger. *Journal of molecular biology* **407**(1): 149-170.

Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C et al. 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**(7035): 917-921.

Fellinghauer B, Buehlmann P, Ryffel M, von Rhein M, Reinhardt JD. 2013. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. *Computational Statistics and Data Analysis* **64**: 132-152.

Fellinghauer B, Bühlmann P. 2011. Stable Graphical Model Estimation with Random Forests for Discrete, Continuous, and Mixed Variables. *arXiv preprint arXiv: …*.

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*: 432-441.

Friedman RC, Farh KK-H, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19**(1): 92-105.

Geiger D, Verma T, Pearl J. 1990. Identifying independence in Bayesian networks. *Networks* **20**(5): 507-534.

Goncalves A, Finetti P, Sabatier R, Gilabert M, Adelaide J, Borg JP, Chaffanet M, Viens P, Birnbaum D, Bertucci F. 2011. Poly(ADP-ribose) polymerase-1 mRNA expression in human breast cancer: a meta-analysis. *Breast Cancer Res Treat* **127**(1): 273-281.

Guzman ML, Rossi RM, Karnischky L, Li X, Peterson DR, Howard DS, Jordan CT. 2005. The sesquiterpene lactone parthenolide induces apoptosis of human acute myelogenous leukemia stem and progenitor cells. *Blood* **105**(11): 4163-4169.

Heacock ML, Stefanick DF, Horton JK, Wilson SH. 2010. Alkylation DNA damage in combination with PARP inhibition results in formation of S-phase-dependent double-strand breaks. *DNA repair* **9**(8): 929-936.

Helleday T. 2011. DNA repair as treatment target. *European journal of cancer* **47 Suppl 3**: S333-335.

Heneghan HM, Miller N, Lowery AJ, Sweeney KJ, Newell J, Kerin MJ. 2010. Circulating microRNAs as novel minimally invasive biomarkers for breast cancer. *Annals of surgery* **251**: 499-505.

Hopkins TA, Shi Y, Rodriguez LE, Solomon LR, Donawho CK, DiGiammarino EL, Panchal SC, Wilsbacher JL, Gao W, Olson AM et al. 2015. Mechanistic Dissection of PARP1 Trapping and the Impact on In Vivo Tolerability and Efficacy of PARP Inhibitors. *Molecular cancer research : MCR* **13**(11): 1465-1477.

Horton JK, Stefanick DF, Naron JM, Kedar PS, Wilson SH. 2005. Poly(ADP-ribose) polymerase activity prevents signaling pathways for cell cycle arrest after DNA methylating agent exposure. *The Journal of biological chemistry* **280**(16): 15773-15785.

Huang GT, Athanassiou C, Benos PV. 2011. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res* **39**(Web Server issue): W416-423.

Huang GT, Tsamardinos I, Raghu V, Kaminski N, Benos PV. 2014. T-ReCS: Stable selection of dynamically formed groups of features with application to prediction of clinical outcomes. In *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, Vol 20, pp. 431-442. World Scientific.

Ikejima M, Noguchi S, Yamashita R, Ogura T, Sugimura T, Gill DM, Miwa M. 1990. The zinc fingers of human poly(ADP-ribose) polymerase are differentially required for the recognition of DNA breaks and nicks and the consequent enzyme activation. Other structures recognize intact DNA. *The Journal of biological chemistry* **265**(35): 21907-21913.

Jang SY, Jang S-W, Ko J. 2012. Regulation of ADP-ribosylation factor 4 expression by small leucine zipper protein and involvement in breast cancer cell migration. *Cancer letters* **314**(2): 185-197.

Jansson MD, Lund AH. 2012. MicroRNA and cancer. *Molecular Oncology* **6**: 590-610.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**(24): 2938-2939.

Kedar PS, Stefanick DF, Horton JK, Wilson SH. 2012. Increased PARP-1 association with DNA in alkylation damaged, PARP-inhibited mouse fibroblasts. *Molecular cancer research : MCR* **10**(3): 360-368.

Kobayashi H, Tomari Y. 2016. RISC assembly: Coordination between small RNAs and Argonaute proteins. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1859**(1): 71-81.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* **40**(10): 1253-1260.

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M. 2005. Combinatorial microRNA target predictions. *Nature genetics* **37**(5): 495-500.

Lauritzen SL, Wermuth N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*: 31-57.

Lee J, Hastie T. 2013. Structure Learning of Mixed Graphical Models. *Journal of Machine Learning Research* **31**: 388-396.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell* **120**(1): 15-20.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**(1): 1.

Lin WC, Lin FT, Nevins JR. 2001. Selective induction of E2F1 in response to DNA damage, mediated by ATM-dependent phosphorylation. *Genes & development* **15**(14): 1833-1844.

Liu H, Lafferty J, Wasserman L. 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research* **10**: 2295-2328.

Liu H, Roeder K, Wasserman L. 2010. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. In *Advances in Neural Information Processing Systems*, pp. 1432-1440.

Liu XS, Haines JE, Mehanna EK, Genet MD, Ben-Sahra I, Asara JM, Manning BD, Yuan ZM. 2014. ZBTB7A acts as a tumor suppressor through the transcriptional repression of glycolysis. *Genes & development* **28**(17): 1917-1928.

Loh P-L, Bühlmann P. 2014. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research* **15**(1): 3065-3105.

Lokk K, Vooder T, Kolde R, Valk K, Vosa U, Roosipuu R, Milani L, Fischer K, Koltsina M, Urgard E et al. 2012. Methylation markers of early-stage non-small cell lung cancer. *PloS one* **7**(6): e39813.

Lu T-P, Lee C-Y, Tsai M-H, Chiu Y-C, Hsiao CK, Lai L-C, Chuang EY. 2012. miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PloS one* **7**: e42390.

Luo X, Kraus WL. 2012. On PAR with PARP: cellular stress signaling through poly(ADP-ribose) and PARP-1. *Genes & development* **26**(5): 417-432.

Macgregor S, Montgomery GW, Liu JZ, Zhao ZZ, Henders AK, Stark M, Schmid H, Holland EA, Duffy DL, Zhang M et al. 2011. Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. *Nat Genet* **43**(11): 1114-1118.

Malumbres M. 2013. MiRNAs and cancer: An epigenetics view. *Molecular Aspects of Medicine* **34**: 863-874.

Mannino DM, Valvi D, Mullerova H, Tal-Singer R. 2012. Fibrinogen, COPD and mortality in a nationally representative US cohort. *COPD: Journal of Chronic Obstructive Pulmonary Disease* **9**(4): 359-366.

Masson M, Niedergang C, Schreiber V, Muller S, Menissier-de Murcia J, de Murcia G. 1998. XRCC1 is specifically associated with poly(ADP-ribose) polymerase and negatively regulates its activity following DNA damage. *Molecular and cellular biology* **18**(6): 3563-3571.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et biophysica acta* **405**(2): 442-451.

Mazur W, Linja-Aho A, Ronty M, Toljamo T, Bergmann U, Kinnula V, Ohlmeier S. 2012. Sputum Proteomics Identifies New Potential Markers For Chronic Obstructive Pulmonary Disease (COPD). *Am J Respir Crit Care Med* **185**: A3748.

Meinshausen N, Buehlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. *Ann Statist* **34**(3): 1049-1579.

Meinshausen N, Bühlmann P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**: 417-473.

Melo SA, Esteller M. 2011. Dysregulation of microRNAs in cancer: Playing with fire. *FEBS Letters* **585**: 2087-2099.

Miknyoczki S, Jones-Bolin S. 2003. Chemopotentiation of Temozolomide, Irinotecan, and Cisplatin Activity by CEP-6800, a Poly(ADP-Ribose) Polymerase Inhibitor. *Molecular cancer …*: 371-382.

Miknyoczki SJ, Jones-Bolin S, Pritchard S, Hunter K, Zhao H, Wan W, Ator M, Bihovsky R, Hudkins R, Chatterjee S et al. 2003. Chemopotentiation of temozolomide, irinotecan, and cisplatin activity by CEP-6800, a poly(ADP-ribose) polymerase inhibitor. *Molecular cancer therapeutics* **2**(4): 371-382.

Molinete M, Vermeulen W, Burkle A, Menissier-de Murcia J, Kupper JH, Hoeijmakers JH, de Murcia G. 1993. Overproduction of the poly(ADP-ribose) polymerase DNA-binding domain blocks alkylation-induced DNA repair synthesis in mammalian cells. *EMBO J* **12**(5): 2109-2117.

Mortusewicz O, Ame JC, Schreiber V, Leonhardt H. 2007. Feedback-regulated poly(ADP-ribosyl)ation by PARP-1 is required for rapid response to DNA damage in living cells. *Nucleic Acids Res* **35**(22): 7665-7675.

Murai J, Huang SY, Das BB, Renaud A, Zhang Y, Doroshow JH, Ji J, Takeda S, Pommier Y. 2012. Trapping of PARP1 and PARP2 by Clinical PARP Inhibitors. *Cancer research* **72**(21): 5588-5599.

Murai J, Huang SY, Renaud A, Zhang Y, Ji J, Takeda S, Morris J, Teicher B, Doroshow JH, Pommier Y. 2014a. Stereospecific PARP trapping by BMN 673 and comparison with olaparib and rucaparib. *Molecular cancer therapeutics* **13**(2): 433-443.

Murai J, Zhang Y, Morris J, Ji J, Takeda S, Doroshow JH, Pommier Y. 2014b. Rationale for poly(ADP-ribose) polymerase (PARP) inhibitors in combination therapy with camptothecins or temozolomide based on PARP trapping versus catalytic inhibition. *J Pharmacol Exp Ther* **349**(3): 408-416.

Ohtsuka M, Ling H, Doki Y, Mori M, Calin G. 2015. MicroRNA Processing and Human Cancer. *Journal of Clinical Medicine* **4**: 1651-1667.

Pandit KV, Corcoran D, Yousef H, Yarlagadda M, Tzouvelekis A, Gibson KF, Konishi K, Yousem SA, Singh M, Handley D. 2010. Inhibition and role of let-7d in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* **182**(2): 220-229.

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z et al. 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**(8): 1160-1167.

Ramsey J, Zhang J, Spirtes PL. 2006. Adjacency-Faithfulness and Conservative Causal Inference. In *Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006)*, MIT, Cambridge, Massachussetts, USA.

Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y. 2012. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research* **72**(14): 3499-3511.

Riganti C, Kopecka J, Panada E, Barak S, Rubinstein M. 2015. The role of C/EBP-beta LIP in multidrug resistance. *Journal of the National Cancer Institute* **107**(5).

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*: gkv007.

Rivas MA, Venturutti L, Huang Y-W, Schillaci R, Huang TH-M, Elizalde PV. 2012. Downregulation of the tumor-suppressor miR-16 via progestin-mediated oncogenic signaling contributes to breast cancer development. *Breast cancer research : BCR* **14**: R77.

Romero V, Rumi R, Salmeron A. 2006. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning* **42**(1-2): 54-68.

Rosas IO, Richards TJ, Konishi K, Zhang Y, Gibson K, Lokshin AE, Lindell KO, Cisneros J, MacDonald SD, Pardo A. 2008. MMP1 and MMP7 as potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. *PLoS medicine* **5**(4): e93.

Satoh MS, Lindahl T. 1992. Role of poly(ADP-ribose) formation in DNA repair. *Nature* **356**(6367): 356-358.

Schreiber V, Dantzer F, Ame JC, de Murcia G. 2006. Poly(ADP-ribose): novel functions for an old molecule. *Nature reviews Molecular cell biology* **7**(7): 517-528.

Schwarz G. 1978. Estimating the dimension of a model. *The annals of statistics* **6**(2): 461-464.

Sedgewick AJ, Benz SC, Rabizadeh S, Soon-Shiong P, Vaske CJ. 2013. Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. *Bioinformatics* **29**(13): i62-i70.

Sedgewick AJ, Shi I, Donovan R, Benos PV. 2016. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*: accepted.

Shah MY, Calin GA. 2011. MicroRNAs miR-221 and miR-222: a new level of regulation in aggressive breast cancer. *Genome Med* **3**: 56.

Shah RD, Samworth RJ. 2013. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1): 55-80.

Spirtes P, Glymour C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review* **9**(1): 62-72.

Spirtes P, Glymour CN, Scheines R. 2000. *Causation, Prediction, and Search*. MIT Press.

Strom CE, Johansson F, Uhlen M, Szigyarto CA, Erixon K, Helleday T. 2011. Poly (ADP-ribose) polymerase (PARP) is not involved in base excision repair but PARP inhibition traps a single-strand intermediate. *Nucleic Acids Res* **39**(8): 3166-3175.

Su W, Xia J, Chen X, Xu M, Nie L, Chen N, Gong J, Li X, Zhou Q. 2014. Ectopic expression of AP-2alpha transcription factor suppresses glioma progression. *Int J Clin Exp Pathol* **7**(12): 8666-8674.

Tentori L, Portarena I, Barbarino M, Balduzzi A, Levati L, Vergati M, Biroccio A, Gold B, Lombardi ML, Graziani G. 2003. Inhibition of telomerase increases resistance of melanoma cells to temozolomide, but not to temozolomide combined with poly (adp-ribose) polymerase inhibitor. *Molecular pharmacology* **63**: 192-202.

Toba H, Cortez D, Lindsey ML, Chilton RJ. 2014. Applications of miRNA Technology for Atherosclerosis. *Current atherosclerosis reports* **16**: 386.

Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E. 2003. Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS Conference*, Vol 2.

Tsamardinos I, Brown LE, Aliferis CF. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn* **65**(1): 31-78.

Tur I, Castelo R. 2011. Learning mixed graphical models from data with p larger than n. In *UAI*, pp. 689-697.

-. 2012. Learning high-dimensional mixed graphical models with missing values. In *Probabilistic Graphical Models (PGM) 2012*, Granada, Spain.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**(12): i237-245.

Walsh AA, Szklarz GD, Scott EE. 2013. Human cytochrome P450 1A1 structure and utility in understanding drug and xenobiotic metabolism. *Journal of Biological Chemistry* **288**(18): 12932-12943.

Wang C, Chen Q, Hamajima Y, Sun W, Zheng YQ, Hu XH, Ondrey FG, Lin JZ. 2012a. Id2 regulates the proliferation of squamous cell carcinoma in vitro via the NF-kappaB/Cyclin D1 pathway. *Chinese journal of cancer* **31**(9): 430-439.

Wang L, Mason Ka, Ang KK, Buchholz T, Valdecanas D, Mathur A, Buser-Doepner C, Toniatti C, Milas L. 2012b. MK-4827, a PARP-1/-2 inhibitor, strongly enhances response of human lung and breast cancer xenografts to radiation. *Investigational new drugs* **30**: 2113-2120.

Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. 2013. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**(2): 149-159.

Wang Y, Wiltshire T, Senft J, Reed E, Wang W. 2007. Irofulven induces replication-dependent CHK2 activation related to p53 status. *Biochemical pharmacology* **73**(4): 469-480.

Yang E, Baker Y, Ravikumar P, Allen G, Liu Z. 2014. Mixed graphical models via exponential families. *J Mach Learn Res* **33**: 1042-1050.

Yang X, Martin TA, Jiang WG. 2013. Biological influence of brain−derived neurotrophic factor (BDNF) on colon cancer cells. *Experimental and therapeutic medicine* **6**(6): 1475-1481.

Zhang J, Le TD, Liu L, Liu B, He J, Goodall GJ, Li J. 2014. Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data. *Bioinformatics*: btu489.

Zhang L, Kim S. 2014. Learning Gene Networks under SNP Perturbations Using eQTL Datasets. *PLoS computational biology* **10**(2): e1003420.

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. 2012. The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research* **13**(1): 1059-1062.

Zou H, Hastie T, Tibshirani R. 2007. On the "degrees of freedom" of the lasso. *The Annals of Statistics* **35**(5): 2173-2192.