2017 IEEE 33rd International Conference on Data Engineering

Integrated Theory- and Data-driven Feature Selection in Gene Expression Data Analysis

Vineet K. Raghu[1], Xiaoyu Ge[1], Panos K. Chrysanthis[1], Panayiotis V. Benos[1,2]
[1] Department of Computer Science, University of Pittsburgh
[2] Department of Computational and Systems Biology, University of Pittsburgh
[1]{vineet, xiaoyu, panos}@cs.pitt.edu, [2]{benos}@pitt.edu

***Abstract*—The exponential growth of high dimensional biological data has led to a rapid increase in demand for automated approaches for knowledge production. Existing methods rely on two general approaches to address this challenge: 1) the Theory-driven approach, which utilizes prior accumulated knowledge, and 2)the Data-driven approach, which solely utilizes the data to deduce scientific knowledge. Both of these approaches alone suffer from bias toward past/present knowledge, as they fail to incorporate all of the current knowledge that is available to make new discoveries. In this paper, we show how an integrated method can effectively address the high dimensionality of big biological data, which is a major problem for pure data-driven analysis approaches. We realize our approach in a novel two-step analytical workflow that incorporates a new feature selection paradigm as the first step to handling high-throughput gene expression data analysis and that utilizes graphical causal modeling as the second step to handle the automatic extraction of causal relationships. Our results, on real-world clinical datasets from The Cancer Genome Atlas (TCGA), demonstrate that our method is capable of intelligently selecting genes for learning effective causal networks.**

I. Introduction

Automated approaches for knowledge production from large-scale biological datasets are crucial to analyzing the exponentially increasing amount of publicly available data for analysis. Standard approaches can be categorized into two major classes: the Theory-driven approach, and the Data-driven approach [1]. The Theory-driven approach utilizes background knowledge accumulated through prior research to establish future knowledge, while the data-driven approach relies solely upon the data being analyzed to generate new scientific knowledge. Using either approach alone suffers from bias towards past/present knowledge and fails to incorporate all the information that is available to make future scientific claims.

One specific type of biological big data being readily analyzed is high-throughput gene expression data, which consists of measurements of the expression levels of the full human genome. To complement this expression data, clinical data (demographics, health status, etc.) provides useful information about patients. These heterogeneous sources together provide researchers with a way of uncovering meaningful relationships among genes and clinical outcomes to further the goal of individualized treatment plans. A major issue in analyzing this type of data is high dimensionality, which poses problems to pure data-driven analysis approaches. A way to mitigate this problem is feature selection; however, solely statistical techniques for feature selection do not incorporate background knowledge into this process.

For example, causal models were developed for automatic extraction of relationships between continuous and discrete variables; however, these algorithms are typically computationally expensive and cannot function well on very high dimensional datasets ($>$1000 variables) [2]. Feature selection for causal modeling approaches have recently started to emerge [3], [4] and using typical machine learning feature selection approaches for predictive models ignores theory-driven knowledge.

In this work, we present a novel paradigm for feature selection on high-throughput gene expression data, and we use this paradigm as the first step in a two-step analytical workflow. The second step of the workflow is causal modeling [5], which automatically extracts causal relationships from data under suitable assumptions [6].

Central to our paradigm for feature selection is a theory-driven analysis curated from a gene-disease relationship database used alongside data-driven analysis from the expression data itself to produce a single *importance* or *preference score* for each gene. We combine gene importance with differences between genes to select relevant and diverse features to be subsequently analyzed. This procedure avoids two major issues in standard modeling approaches: highly correlated variables and lack of prior knowledge. Highly correlated variables pose problems for graphical models that use independence relations between variables to extract causality. In addition, failing to utilize prior knowledge in feature selection limits the amount of useful information we can derive from our data.

We use the *Preferential Diversity* (PrefDiv) framework [7] to realize the first step of our analytical workflow. *PrefDiv* was proposed as an efficient solution to the Maximum Covering Diversified Top-K problem in traditional databases and used as a tool for big data exploration. Given a ranked set of objects, *PrefDiv* returns a set of K objects with maximized relevance and diversity. For the second step, we use the *MGM-PC Stable* causal discovery method that learns causal knowledge in high-dimensional datasets with discrete and continuous data [8].

Building upon the *PrefDiv* and Graphical Causal Modeling frameworks **our major contributions** are as follows:

- We define functions that serve as precise measures of importance and semantic distance for genes, and use these functions with *PrefDiv* to select relevant and diverse subsets of genes from expression data. (Sec. II)
- We combine this selected subset of genes with clinical data from The Cancer Genome Atlas (TCGA) as input to *MGM-PC Stable* to generate causal networks. (Sec. III)

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.223

- We provide an evaluation of the selected genes from our workflow as well as the learned causal networks on these genes and show the effectiveness of our proposed approach. In our evaluation, we use Top-K and Pref-Div gene sets produced by our approach and two baseline sets of randomly selected genes and genes produced based on maximum statistical variance. (Sec. IV)

II. Integrated Approach for Feature Selection

Our hypothesis is that each gene's role can be expressed by a single *importance score* and for different types of analyses, similar or dissimilar genes are the most preferred to be explored. There are several ways in which the importance score and gene distance can be used to select the most relevant genes. One example is the Top-K approach where the genes with the K highest importance scores are selected. Another is to select the K most diverse (dissimilar) genes. In this section, we describe our theory-driven and data-driven integrated approach for genetic feature selection. We first describe the data used to illustrate our approach, and then describe the computation of both the *importance* or *preference score* of genes as well as our *distance equation* between genes.

A. Datasets

In this work, we use a heterogeneous set of genetic information collected from a pathway information database, a gene expression dataset coupled with clinical information, a gene-disease similarity scoring dataset, and an informational dataset for each human gene. Each of these datasets is described below:

Pathway Information Data: Pathway Information data was obtained from the KEGG Pathway Database [9]. Three pathways (WNT-Signaling Pathway, TGF-β Signaling Pathway, and Breast Cancer Pathway) were used, as these are known to be relevant to breast cancer [10], [11]. The pathway information was trimmed to just the genes involved in each pathway, and these were used for downstream validation testing, which is further explained in Section IV-B.

Prediction Analysis of Microarray 50 (PAM50): PAM50 is a set of genes whose expression profile is used by clinicians to determine short-term treatment plans for breast cancer patients. These genes can also be used to inform breast cancer subtype classification. [12] In our work, the PAM50 gene list is used as further biological validation for our importance score.

Gene Expression Data: Gene expression data was provided by The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) project [13]. This data consists of RSEM normalized RNA-Seq gene expression data for 20,531 genes across 1,212 patients. In addition, clinical data (Age, Race, Gender, Vital Status, Tumor Stage, and Cancer Subtype) is available for each of these patients through TCGA-BRCA.

Gene-Disease Scoring Data: A gene-disease mapping dataset was provided by the DisGeNET database [14]. This data consists of 17,381 genes and 15,093 diseases composing 429,057 gene-disease associations with a score for each computed based upon the level of evidence present supporting the gene-disease relationship. In particular, the scoring function takes into account which organism the evidence was produced from (Human, Mouse Model, etc.), the level of curation of the data (Uniprot, GWAS, etc.), and the number of publications in which the gene-disease relationship appears.

Genetic Information Data: Genetic information data was obtained from the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) [15]. This database had data for 19,030 human genes including their full gene name, official gene symbol, chromosomal location, and gene family/functional grouping. This data was supplemented with specific chromosomal base pair location data for each gene from the PharmGKB database [16].

All of the information utilized by our approach for each gene is summarized in Table I.

B. Importance and Distance Equations

In this section, we define two central concepts for modeling the selection of genes, 1) the importance score of a gene G, denoted by $I(G)$, and 2) the distance between two genes G_1 and G_2, denoted by $dt(G_1, G_2)$.

Importance Score: The importance score for a particular gene G is given by the following equation:

$$I(G) = \alpha * \log_2(FC_G) + (1-\alpha) * \frac{\sum_{i=1}^{n}(SD_i)}{N} \quad (1)$$

Here, FC_G refers to the fold change of the gene expression values of G between control and experimental tissue samples. SD_i refers to the relevancy score between gene G and disease i. N refers to the number of diseases the user wishes to study (the sum ranges over all user-specified diseases of interest), and α is a positive user specified parameter. This parameter balances the trade-off between using expression data to determine gene preference versus using prior information about the gene-disease mapping to determine gene preference.

Gene Distance Metric: To measure the distance between two genes G_1 and G_2, we first measure the chromosomal distance between the two genes, which is defined by:

$$CD(G_1, G_2) = \frac{maxBase(G_1, G_2) - minBase(G_1, G_2)}{\text{\# of bases in Chromosome}} \quad (2)$$

Here, when G_1 and G_2 are located on the same chromosome. maxBase refers to the largest base position of either gene on the chromosome, while minBase refers to the smallest base position. If G_1 and G_2 are located on different chromosomes this function is equal to 1, denoting the maximum possible chromosomal distance.

In order to define the distance using the gene-disease mapping some additional notation is required. Let $GD_1(i)$ be the score of the relevancy of gene G_1 with disease i. Then the distance between two genes based on their disease relevancy scores is given by:

$$DD(G_1, G_2) = \sum_{i=1}^{n}(GD_1(i) - GD_2(i))^2 \quad (3)$$

Note that this is equivalent to the euclidean distance between the gene-disease relevancy vectors of the two argument genes.

Now, the overall distance equation between two genes G_1 and G_2 is given by the following equation:

TABLE I: Genetic Data Utilized

Attribute	Example	Description
Gene Symbol	CDH1	Official HUGO Gene Symbol
Chromosomal Location	chr16, 68761195 to 68872444	Chromosome number of the gene, along with specific base pair location
Gene Family	Type 1 Classical Cadherins	List of Gene families to which this gene belongs
Expression Values	$[x_1, ..., x_n]$	Vector of expression values across normal and tumor samples
Gene Disease Association	CDH1, Adenocarcinoma, 0.0165	Relevance Score for gene-disease pairs

$$\begin{aligned} dt(G_1, G_2) =&(a * CD(G_1, G_2)+ \\ & b * DD(G_1, G_2) + c * \rho_{exp})+ \\ & (1-(a+b+c)) * \delta(G_1, G_2) \end{aligned} \tag{4}$$

Here, a, b, and c are positive parameters where $(a+b+c = 1)$. $\delta(G_1, G_2)$ is an indicator function that is 1 when the two genes belong to different gene families and is 0 otherwise. ρ_{exp} refers to the Pearson product moment correlation between the full expression vectors of G_1 and G_2 across all samples.

III. Two-Step Workflow

As mentioned above, we use our integrated approach for the selection of genes as the first step in a two-step analytical workflow. We adopted *PrefDiv* [7] to realize this step because of its ability to select genes that are relevant yet diverse. The assumption is that often highly ranked genes are similar and do not contribute any additional information. In the second step, the selected genes become an input to the MGM PC-Stable [8] causal discovery method.

A. Relevance, Intensity, Diversity, and Similarity

We start by formally introducing the concepts of relevance and diversity, which are the two crucial elements for utilizing the *PrefDiv* algorithm.

Relevance: The Relevance of a gene essentially means the importance of this gene. Thus, the relevance of a gene G is represented by $I(G)$ (Equation 1).

Diversity: We measure the diversity of a set of genes S by measuring how dissimilar, i.e., the distance beyond a threshold, each gene in S is with respect to each other.

Definition 1: Dissimilarity Let S be the set of genes in the database. Two genes G_i and $G_j \in S$ are dissimilar to each other $dsm_\varrho(G_i, G_j)$, if $dt(G_i, G_j) > \varrho$, for a real number ϱ, where ϱ is a distance parameter, which we call radius.

Definition 2: Similarity Let S be the set of items. Two genes G_i and $G_j \in S$ are similar to each other, if $dt(G_i, G_j) \leq \varrho$ for a real number ϱ. We use $sim_\varrho(G_i, S)$ to denote a set of items in S that are similar to an gene G_i, such that $\forall G_j \in sim_\varrho(G_i, S), G_j \neq G_i$.

B. Preferential Diversity

In this section, we present the details of *Preferential Diversity* (*PrefDiv*) [7], which we have previously proposed as an efficient solution to the Diversified Top-k problem. *PrefDiv* is an iterative algorithm that utilizes a ranking model that produces an initial result set of genes for a given query and returns a set of k genes with maximized relevance and diversity. *PrefDiv* is shown in Algorithm 1 and its input parameters in Table II.

Parameter A is used to tune the balance between relevance and diversity in the returned result set. Specifically, A defines the distribution of the intensity values of genes in the final result set R. When $A = 1$, R would simply be the top k genes from the initial set, i.e., the genes with the k highest intensity values. When $A = 0$, R contains k dissimilar genes from the initial set. When A is between 0 and 1 and given that *PrefDiv* is an iterative algorithm, the final result will have at least $A * k$ genes from every iteration, and, in each iteration, A will be divided by half. For example, when $A = 0.5$ and $k = 20$, the first iteration will select at least $20 * 0.5$ genes for the final result set, the second iteration will select at least $20 * (0.5 * 0.5)$ genes, and so on.

The basic logic of *PrefDiv* is as follows: It first sorts the genes in the initial set $S = \{G_1, ..., G_n\}$ in descending order along with their intensity value and splits them into groups of k genes. In each iteration, it evaluates the genes in a group for diversity, starting with the first group with the highest intensity genes. The gene G_i with the highest $I(G_i)$ in the group T_S is moved into the final result set R, if there is no gene in R similar to G_i, i.e., $sim_\varrho(G_i, R)$ is empty; otherwise it is marked as "Eliminated". Also, all genes in $sim_\varrho(G_i, T_S)$ are marked as "Eliminated". While there are still genes left in T_O that are not marked as "Eliminated", it processes the next unmarked one G_j with the highest $I(G_j)$ in the same manner. It ends an iteration by finalizing the moved genes into R according to A, as mentioned above. If fewer than the required $A * k^{iteration}$ genes were moved in R, then the difference s is covered by moving the top-s genes with the highest intensity values that have been marked as "Eliminated" in T_S into R. The iterations continue until either k genes are selected ($|R| = k$), or if all genes in S are examined. If the size of R is still less than k, $k - |R|$ genes with the highest intensity values that have been marked as "Eliminated" will be selected and added into R.

PrefDiv is linear to the number of genes in the initial set. The initial candidate selection for the first iteration takes $O(k^2)$ and each subsequent iteration costs $O(k^2)$ as well. As there are at most $\frac{N}{k}$ iterations, Algorithm 1 has an overall worst case complexity of $O(kN)$.

C. Graphical Causal Models

Graphical causal modeling is a data analysis methodology to infer causal relationships from a dataset of observations of random variables. These algorithms typically fall into two

Algorithm 1 *PrefDiv*

```
Require:
    One set of genes $S$, a size $k$, a relevancy parameter $A$, and a radius $\varrho$
Ensure:
    One subset $R$ of $S$
$T \leftarrow \emptyset$
$turnCounter = 0$
while there exists unmarked items in $S$ and $|R| < k$ do
    Increase $turnCounter$ by $1$
    $T \leftarrow$ Pick $k$ items with highest intensity from $S$
    for all genes $G_i \in R$ do
        for all genes $G_j \in T$, s.t. $G_j \in sim_\varrho(G_i, T)$ do
            Mark $G_j$ as ”Eliminated”
        end for
    end for
    while there exists unmarked items in $T$ do
        $R = R \cup G_i$, s.t. $G_i \in T$ is unmarked and $I(G_i) \geq I(G_j) : \forall G_j \in T$
        for all unmarked $G_u \in T$ do
            if $G_u \in sim_\varrho(G_i, T)$ then
                mark $G_u$ as “Eliminated”
            end if
        end for
    end while
    while number of unmarked items in $T < A \cdot k$ do
        $R = R \cup G_i$, s.t. $G_i \in S$ is unmarked and $I(G_i) \geq I(G_j) : \forall G_j \in T$
    end while
    $A = A \cdot 0.5$
    if $turnCounter == 1$ then
        create new set $N \leftarrow \forall G_j \in T$, s.t. $G_j$ is marked
    end if
    $S = S - (S \cap T)$
end while
if $|R| < k$ and $\forall G_j \in S$, s.t. $G_j$ are marked then
    while $|R| < k$ do
        $R = R \cup G_j$, s.t. $G_j \in N$ and $I(G_j) \geq I(G_i) : \forall G_i \in N$
    end while
end if
Return $R$
```

TABLE II: Parameters of *PrefDiv*

Par.	Range	Usage
S	$1 \leq \|S\|$	Set of genes with intensity values
k	$1 \leq k$	Size of the result set
ϱ	$0 \leq \varrho \leq M^1$	Determines whether a pair of genes are similar.
A	0 ≤A≤ 1	Determines the number of genes to be promoted to the result set at each iteration.

[1] M = Max distance of dataset

major categories: constraint-based algorithms, and score-based approaches [17]. In this work, we focus on the constraint-based approach as these algorithms have recently been extended to handle both continuous and categorical variables that are present in our heterogeneous data source [18].

Constraint-based algorithms utilize conditional independence tests to determine direct causal influences between observed variables. The de facto standard for constraint-based causal inference has been the PC Algorithm, and more recently an order independent version called PC-Stable [19]. PC-Stable begins its search with a fully connected graphical structure and performs conditional independence tests in order to delete adjacencies from the graph. The space of all causal graphs is super-exponential in the number of variables, so to avoid this search PC-Stable searches by performing conditional independence tests in increasing size of the conditioning set (starting with unconditional independence tests). The algorithm finds the provably correct equivalence class of causal graphs given that the conditional independence test always outputs the true conditional independencies in the underlying causal graph. For full details of this method we refer the reader to [19] and [6].

In order to improve both the speed and accuracy of the PC-Stable algorithm on mixed data, the Mixed Graphical Models (MGM) [20] preprocessor was used. MGM is a characterization of a joint distribution over discrete and continuous variables given in Equation 5.

$$p(x, y; \theta) \propto exp\Bigg(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\Bigg) \quad (5)$$

Here, x_s represents the s^{th} of p continuous variables and y_j represents the j^{th} of q discrete variables. β_{st} represents the potential for an edge between continuous variables s and t, α_s represents the potential for a node of a continuous variable, ρ_{sj} represents the potential for an edge between continuous variable s and discrete variable j, and finally Φ_{rj} represents the potential for an edge between discrete variables r and j. This model has the favorable property that its conditional distributions are given by Gaussian linear regression and Multiclass Logistic Regression for continuous and discrete variables respectively. However, as many next generation genome sequencing datasets consist of continuous variables that do not satisfy a normality assumption, we employ a non-paranormal transformation of the data to render our method suitable for any mRNA sequencing method.

Optimizing this distribution exactly is computationally intractable, and thus a pseudolikelihood approach is used. To avoid overfitting and to ensure a sparse causal structure a penalized form of the pseudolikelihood is used, and this penalty is given by a separate regularization parameter for each type of edge (λ_{CC} for Continuous-Continuous edges, λ_{CD} for Continuous-Discrete edges, and λ_{DD} for Discrete-Discrete edges). This penalized pseudolikelihood is given by Equation 6.

The nonzero parameters from the MGM model serve as indicators of the existence of an edge between two variables, and so MGM learns an undirected graph representing the adjacencies of the variables in the undirected causal graph. The graph produced by MGM then serves as a starting point for the PC-Stable algorithm (instead of the fully connected graph), thus resulting in runtime savings and improved precision.

$$\begin{aligned}\underset{\Theta}{\text{minimize}}\, l_\lambda(\Theta) = \widetilde{l}(\Theta) &+ \lambda_{CC}\sum_{s=1}^{p}\sum_{t=1}^{s-1}|\beta_{st}| \\ &+ \lambda_{CD}\sum_{s=1}^{p}\sum_{j=1}^{q}||\rho_{sj}||_2 \qquad (6) \\ &+ \lambda_{DD}\sum_{j=1}^{q}\sum_{r=1}^{j-1}||\phi_{rj}||_F\end{aligned}$$

IV. Experimental Evaluation

Here we demonstrate our workflow on real expression and clinical data from the TCGA-BRCA project [13]. We validate our gene selection results and the predictive modeling that results from using the selected genes. We test four different gene selection methods: Random, Highest variance, Top-K, and Pref-Div. Highest variance refers to choosing the K genes with the largest statistical variance, Top-K refers to choosing the K genes with the largest intensity score, and Pref-Div uses the *PrefDiv* algorithm to choose a diverse set of K genes with high intensity. We perform three evaluations of our approach: a sensitivity analysis of the parameters of our importance score and *PrefDiv*, an evaluation of the relevance of the selected genes to breast cancer, and an examination of the predictive accuracy of the selected genes to important clinical variables.

A. Parameter Selection

The initial segment of our experimental workflow is to determine the parameter α for our gene importance score (Equation 1). A reasonable choice for α is one that balances the theory-driven approach through the gene-disease relationships with the data-driven approach through fold changes from the expression data. Our exploration of the Top 50 genes selected by intensity score over varying α values is given by Figure 1. From this figure, it is clear that the pure data-driven and theory-driven approaches select entirely different sets of genes. Thus, we chose a balanced approach between the two extremes and set the α parameter to be 0.25.

Due to a large number of parameters for the distance between two genes (Equation 4), we chose to leave these parameters equal. We leave an examination of the distance parameters for future work.

The aforementioned α parameter is the only necessary parameter to set for the Top-K approach. The next portion of the workflow involves finding the appropriate tradeoff between intensity and diversity for the *PrefDiv* approach through the Accuracy parameter (A). We choose to use a similar visual procedure to find a reasonable value. Figure 2 displays a visualization of the selected genes across the various values of the A parameter. Here we choose the parameter value to be 0.5 to balance high intensity scores (A = 1) with fully diverse sets (A = 0).

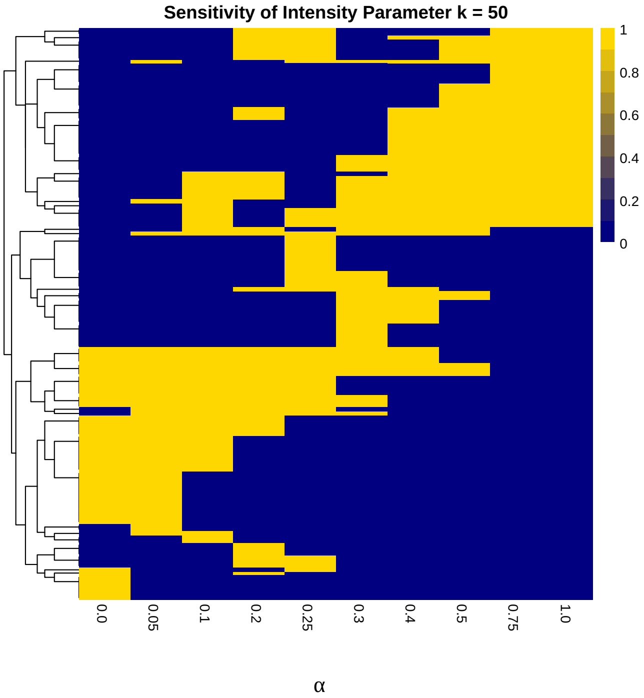

Fig. 1: Heatmap of Selected Genes, varying Intensity Parameter α, using the Top-K Approach

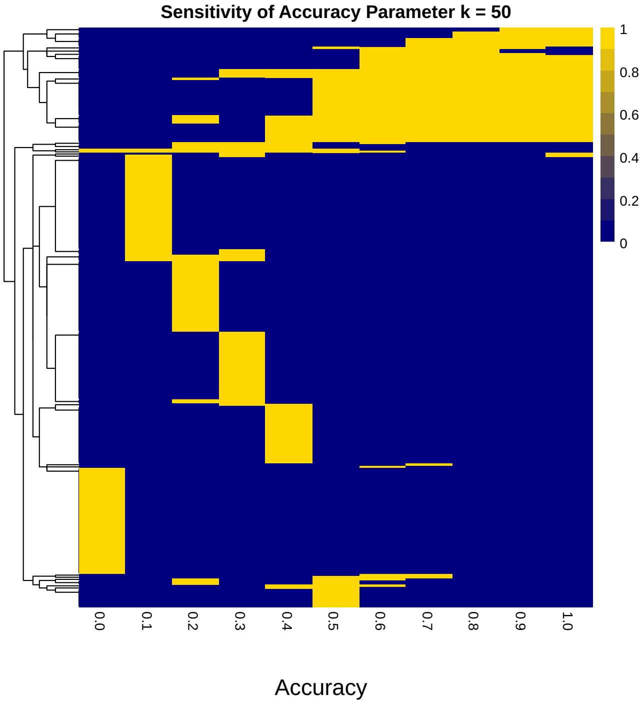

Fig. 2: Heatmap of Selected Genes, varying Accuracy Parameter A for PrefDiv

B. Biological Relevance

Next, we analyze the selected genes for their biological relevance. We find that 6 of our top 50 genes with the highest intensity score are found in the PAM50 (p = 1.55×10^{-7}). Upon examination of the top 50 highest scoring genes, we find that the top five genes with high intensity score (BRCA1, BRCA2, TP53, FGFR2, TOX3) have all been found to be relevant to breast cancer in recent publications. BRCA1 and BRCA2 genes encode tumor suppressing proteins, and it has been shown that a mutation in these genes increases susceptibility to breast cancer [21], [22]. FGFR2 and TOX3 mutations have both been linked to an increase in breast cancer risk though their molecular mechanisms are not well understood [23], [24]. Finally, TP53 is an important gene in many cancers, and it has been shown that a change in the p53 pathway leads to more aggressive breast cancer. [25]

In addition, we determine the biological relevance of the selected genes of all three approaches using pathway data. Our analysis focuses upon three KEGG Pathways relevant to Breast Cancer: WNT-Signaling, TGF-β Signaling, and Breast Cancer. The goal of this study is to determine whether underlying theory supports the relevance of the selected genes to breast invasive carcinoma. The results are given by Figure 3, which shows the percentage of selected genes belonging to at least one of the three relevant pathways for each method. The figure clearly demonstrates that using our intensity score provides more theory-driven gene selections than the data-driven highest variance baseline or randomly selecting genes; however, this effect does tend to dissipate when selecting a very large subset of genes (K = 500). This is due to the fact that the total number of unique genes in the three pathways is 203, thus the maximum possible percentage decreases when selecting 500 genes. It is also interesting to note that both the Pref-Div and Top-K genes have similar relevance scores when computed in this manner, implying that the relevant genes selected by Top-K do not tend to be replaced with irrelevant genes when diversity is also desired. Overall, this study displays theory-driven support for our approach, especially when selecting smaller subsets of genes.

C. Predictive Modeling of Outcomes

1) Approach: Our final experiment attempted to quantify the predictive value of the full workflow. We chose to study the predictive value of the workflow because evaluating causal associations is a difficult problem on real biological data due to a lack of ground truth knowledge. Thus, our experimental design is as follows (outlined in Figure 4). First, we use each of our feature selection approaches to choose a subset of genes as in the previous experiment. Then we use the expression data from this subset merged with clinical variables from the TCGA to produce a full dataset. The causal modeling algorithm (MGM PC-Stable) is then used to find the causal associations between the genetic and clinical variables. For a particular variable of interest, T, we use the Markov Blanket of T in the causal graph as the relevant predictors. The Markov Blanket of a target variable T refers to the set containing the parents of T, the children of T, and the parents of the children of T. This set of variables contains all of the causal

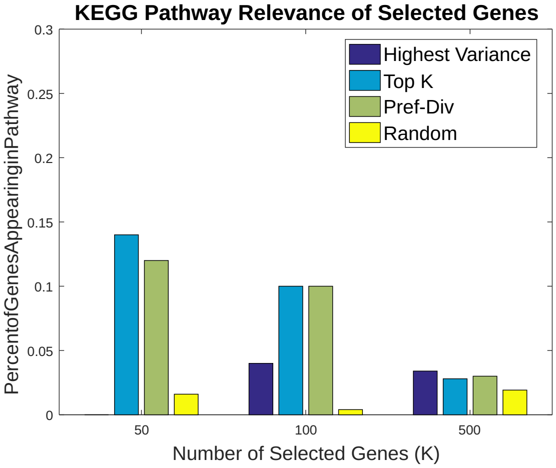

Fig. 3: Percentage of Genes selected by each method appearing in a relevant KEGG Pathway. (Note: Highest Variance had 0 correct genes for K = 50)

information about T that the data can provide. The data from these relevant features are fed to a Support Vector Machine classification algorithm, and 5-fold cross validation is done to find the predictive accuracy of the model. In this manner, we can evaluate the selected features for their predictive relevance to important target variables in Breast Cancer.

2) Results: In Figure 5 the results of predicting whether a given sample is from tumor or normal tissue is displayed by F_1 Score. Clearly, for the smaller sets of variables choosing the Top-K variables with the highest intensity score provides the best predictive accuracy. However, as shown in Figure 6, by selecting only the most relevant variables the method produces relatively small Markov Blankets and thus prevents biologists from potentially finding all relevant genes for the disease of interest. Adding diversity to the Top-K using the *PrefDiv* framework tends to have similar predictive accuracy as the Highest Variance method, but this approach is able to find significantly larger Markov Blankets. This could be due to the fact that having a diverse set of variables prevents one variable from shielding away other genes from causal relevance to the target variable. Thus, we conclude the Top-K and Highest Variance approaches both perform well when predictive accuracy is the goal, while the *PrefDiv* approach finds a large number of relevant features.

Figures 7 and 8 demonstrate results for a similar workflow for predicting the Breast Cancer subtype of each sample. Four breast cancer subtypes were present in this dataset (Luminal-A, Luminal-B, Basal, and HER-2). Since this involved a multi-class classification problem, these results are given as misclassification rate instead of F_1 Score. Here we clearly see similar predictive performance between Top-K, Highest Variance, and PrefDiv approaches. However, we see a significant change in performance when using PrefDiv with a large number of genes selected (e.g., 500) for Causal Modeling. Although, PrefDiv still finds the most variables relevant to the target, these lead to overfitting as nearly 150 predictors were found in the Markov

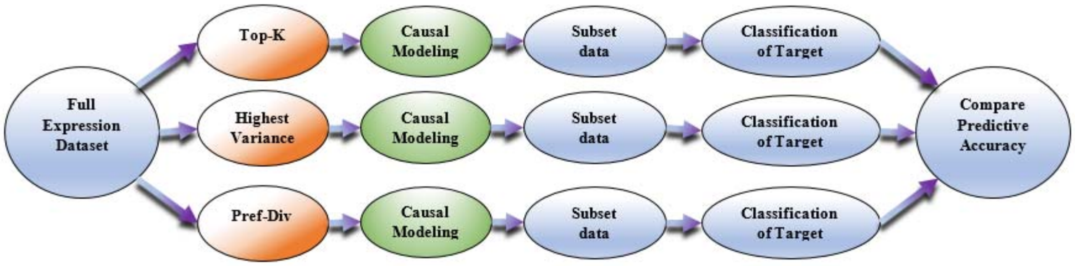

Fig. 4: Illustration of experimental workflow for Predictive Modeling

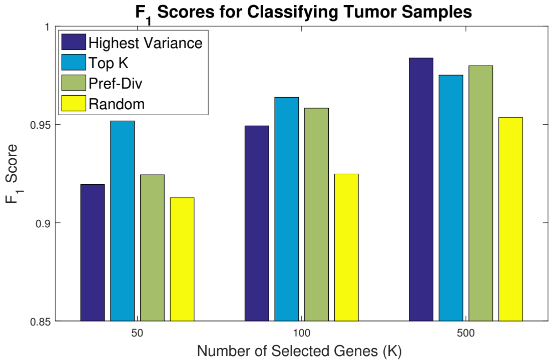

Fig. 5: F_1 Score for Classification of Tumor vs. Normal Samples. Genes selected using Markov Blanket of causal graph learned on variables selected by each method.

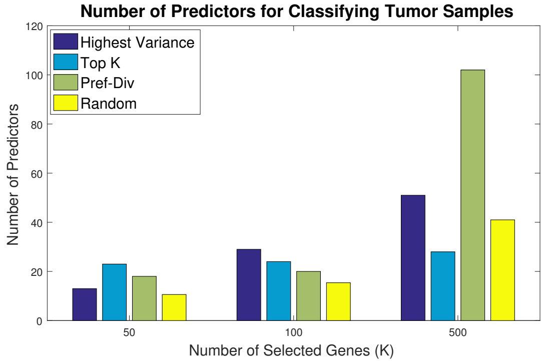

Fig. 6: Number of Predictors found in the Markov Blanket of the Tumor variable for each dimensionality reduction method.

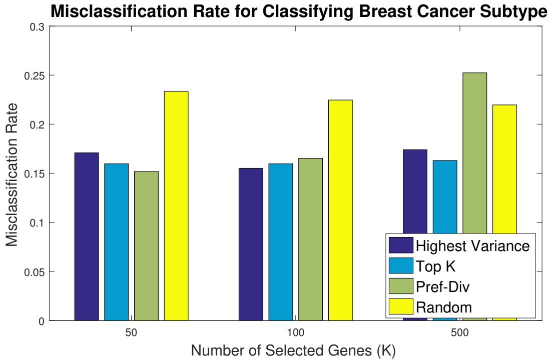

Fig. 7: Misclassification Rate for Classification of Different Cancer Subtypes. Genes selected using Markov Blanket of causal graph learned on variables selected by each method.

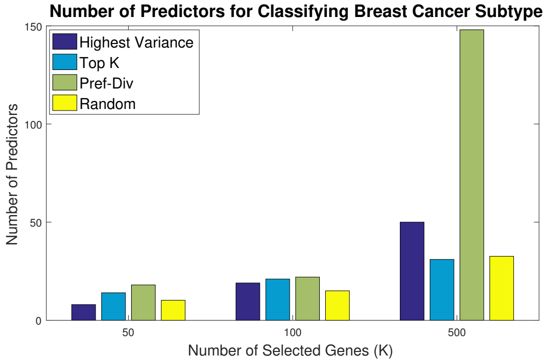

Fig. 8: Number of Predictors found in the Markov Blanket of the Cancer Subtype variable for each dimensionality reduction method.

Blanket of the Subtype variable when using PrefDiv. With these detailed analyses, we established a distinction between predictive and causal models, which can be used to determine the best method based on different use cases.

V. Conclusion

In this work, we have presented a two-step analytical workflow for theory and data-driven feature selection in high-throughput gene expression data. We formulated an importance score for genes as the first step, and we then used graphical causal models to extract variable relationships. Our results show that our approach balances selecting predictive features with finding causally relevant features for knowledge generation compared to the baseline approaches. The Top-K approach appears to be better suited to predictive modeling applications whereas diversity allows for improved causal knowledge generation.

Our work differs from the current state of the art in feature selection, as most approaches have focused on data-driven quantitative methods [26]. However, some methods have incorporated prior knowledge into their dimensionality reduction techniques. In [27] and [28], the authors use a constraint-based feature hierarchy given by a domain expert to inform feature selection techniques. This is difficult to apply to our case as it is unclear how to partition the human genome into layers of relevance to the target variable. In [29], the author uses a domain expert provided pairwise dissimilarity score, similar in nature to ours to project the data into smaller dimensions for clustering. In our work, we do not aim to find clusters of the original data though this could be an added method of evaluation for our dissimilarity score for future work.

Furthermore, our method is the first to study using feature selection methods prior to the causal modeling in an analytical workflow. The second part of our workflow (using causal models for further feature selection) has been lightly examined. Sun et al. use Granger causality for feature selection in time series data [30]. In [31], the authors provide an extensive evaluation of the effectiveness of causal discovery methods to inform predictive variables. And, in [32], the authors examine causal feature selection in instances where the relationships between input and response variables can change between training and testing data. Despite these related works, ours is the first method to utilize prior domain knowledge in a simple manner to inform causal discovery approaches.

In the future, we will work to integrate more types of data to better inform our intensity score. In addition, we intend to examine more closely the biological relevance of portions of the discovered causal networks. Finally, we are also interested to see a more thorough evaluation of the predictive accuracy and causal relevance of our method as compared with more traditional machine learning feature selection approaches. The difficulty in evaluating causal discovery methods is the lack of ground truth causal knowledge, so finding alternative evaluation methods would be a meaningful contribution.

Acknowledgment

Research reported in this publication was supported by the National Institutes of Health under Awards Number U54HG008540 , R01LM012087, T32CA082084 and a gift from EMC. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or EMC. Causal Modeling was performed using the Tetrad V software.

References

[1] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery," *arXiv preprint arXiv:1612.08544*, 2016.

[2] E. Sokolova, P. Groot, T. Claassen, and T. Heskes, "LNAI 8754 - Causal Discovery from Databases with Discrete and Continuous Variables," pp. 442–457, 2014.

[3] G. T. Huang, I. Tsamardinos, V. Raghu, N. Kaminski, and P. V. Benos, "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. NIH Public Access, 2015, p. 431.

[4] L. C. e. a. Villaruz, "Microrna expression profiling predicts clinical outcome of carboplatin/paclitaxel-based therapy in metastatic melanoma treated on the ecog-acrin trial e2603," *Clinical epigenetics*, vol. 7, no. 1, p. 58, 2015.

[5] P. Spirtes, "Introduction to Causal Inference," *Journal of Machine Learning Research*, pp. 1–3, 2011.

[6] P. Spirtes, C. Glymour, and R. Scheines, "Causation, Prediction, and Search," *Technometrics*, vol. 45, no. 3, pp. 272–273, 2003.

[7] X. Ge, P. K. Chrysanthis, and A. Labrinidis, "Preferential diversity," in *ExploreDB*, 2015.

[8] A. Sedgewick, "Graphical models for de novo and pathway-based network prediction over multi-modal high throughput biological data," Ph.D. dissertation, University of Pittsburgh, 2016.

[9] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[10] G.-B. Jang, J.-Y. Kim, S.-D. Cho, K.-S. Park, J.-Y. Jung, H.-Y. Lee, I.-S. Hong, and J.-S. Nam, "Blockade of wnt/β-catenin signaling suppresses breast cancer metastasis by inhibiting csc-like phenotype," *Scientific reports*, vol. 5, p. 12465, 2015.

[11] M. B. Buck and C. Knabbe, "Tgf-beta signaling in breast cancer," *Annals of the New York Academy of Sciences*, vol. 1089, no. 1, pp. 119–126, 2006.

[12] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, 2009, pMID: 19204204.

[13] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumors," *Nature*, vol. 490, no. 7418, p. 61, 2012.

[14] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, p. bav028, 2015.

[15] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, "Genenames. org: the hgnc resources in 2015," *Nucleic acids research*, p. gku1071, 2014.

[16] M. Whirl-Carrillo, E. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. Altman, and T. E. Klein, "Pharmacogenomics knowledge for personalized medicine," *Clinical pharmacology and therapeutics*, vol. 92, no. 4, p. 414, 2012.

[17] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3, no. 1, 2016, p. 3.

[18] A. J. Sedgewick, I. Shi, R. M. Donovan, and P. V. Benos, "Learning mixed graphical models with separate sparsity parameters and stability-based model selection," *BMC Bioinformatics*, vol. 17, no. S5, p. 175, 2016.

[19] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *Annals of Statistics*, vol. 40, no. 1, pp. 294–321, 2012.

[20] J. Lee and T. Hastie, "Learning the Structure of Mixed Graphical Models," *Stanford.Edu*, pp. 1–32, 2013. [Online]. Available: http://www.stanford.edu/˜hastie/Papers/structmgm.pdf

[21] P. Futreal, Q. Liu, D. Shattuck-Eidens, C. Cochran, K. Harshman, S. Tavtigian, L. Bennett, A. Haugen-Strano, J. Swensen, Y. Miki *et al.*, "Brca1 mutations in primary breast and ovarian carcinomas," *Science*, vol. 266, no. 5182, pp. 120–122, 1994.

[22] J. M. Lancaster, R. Wooster, J. Mangion, C. M. Phelan, C. Cochran, C. Gumbs, S. Seal, R. Barfoot, N. Collins, G. Bignell *et al.*, "Brca2 mutations in primary breast and ovarian cancers," *Nature genetics*, vol. 13, no. 2, pp. 238–240, 1996.

[23] M. N. Fletcher, M. A. Castro, X. Wang, I. De Santiago, M. OReilly, S.-F. Chin, O. M. Rueda, C. Caldas, B. A. Ponder, F. Markowetz *et al.*, "Master regulators of fgfr2 signalling and breast cancer risk," *Nature communications*, vol. 4, 2013.

[24] J. O. Jones, S.-F. Chin, L.-A. Wong-Taylor, D. Leaford, B. A. Ponder, C. Caldas, and A.-T. Maia, "Tox3 mutations in breast cancer," *PloS one*, vol. 8, no. 9, p. e74102, 2013.

[25] M. Gasco, S. Shami, and T. Crook, "The p53 pathway in breast cancer," *Breast Cancer Research*, vol. 4, no. 2, p. 70, 2002.

[26] A. M. C.O.S.Sorzano, J.Vargas, "A survey of dimensionality reduction techniques," *ArXiv*, vol. abs/1403.2877.

[27] W. C. Groves, "Toward automating and systematizing the use of domain knowledge in feature selection," Ph.D. dissertation, University of Minnesota, 2015.

[28] S. Radovanovic, M. Vukicevic, A. Kovacevic, G. Stiglic, and Z. Obradovic, "Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction," in *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 2015, pp. 96–100.

[29] I. Davidson, "Knowledge driven dimension reduction for clustering," *IJCAI*, 2009.

[30] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, "Using causal discovery for feature selection in multivariate numerical time series," *Machine Learning*, vol. 101, no. 1-3, pp. 377–395, 2015.

[31] I. Guyon, C. Aliferis, and A. Elisseeff, "Causal feature selection," *Computational methods of feature selection*, pp. 63–82, 2007.

[32] G. C. Cawley, "Causal and non-causal feature selection for ridge regression." *JMLR*, vol. 3, 2008.