AN INTEGRATED APPROACH TO BLOOD-BASED CANCER DIAGNOSIS AND BIOMARKER DISCOVERY

Martin Renqiang Min[ad*], Salim Chowdhury[bd], Yanjun Qi[a], Alex Stewart[c], and Rachel Ostroff[c]

[a] *NEC Labs America, Princeton, NJ 08540, USA,* [b] *Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA,* [c] *SomaLogic, Inc., Boulder, CO 80301, USA*

E-mail: renqiang@nec-labs.com, salim@cmu.edu, yanjun@nec-labs.com, {astewart, rostroff}@somalogic.com

[d] *These authors contributed equally to this work*

Disrupted or abnormal biological processes responsible for cancers often quantitatively manifest as disrupted additive and multiplicative interactions of gene/protein expressions correlating with cancer progression. However, the examination of all possible combinatorial interactions between gene features in most case-control studies with limited training data is computationally infeasible. In this paper, we propose a practically feasible data integration approach, QUIRE (QUadratic Interactions among infoRmative fEatures), to identify discriminative complex interactions among informative gene features for cancer diagnosis and biomarker discovery directly based on patient blood samples. QUIRE works in two stages, where it first identifies functionally relevant gene groups for the disease with the help of gene functional annotations and available physical protein interactions, then it explores the combinatorial relationships among the genes from the selected informative groups. Based on our private experimentally generated data from patient blood samples using a novel SOMAmer (Slow Off-rate Modified Aptamer) technology, we apply QUIRE to cancer diagnosis and biomarker discovery for Renal Cell Carcinoma (RCC) and Ovarian Cancer (OVC). To further demonstrate the general applicability of our approach, we also apply QUIRE to a publicly available Colorectal Cancer (CRC) dataset that can be used to prioritize our SOMAmer design. Our experimental results show that QUIRE identifies gene-gene interactions that can better identify the different cancer stages of samples, as compared to other state-of-the-art feature selection methods. A literature survey shows that many of the interactions identified by QUIRE play important roles in the development of cancer.

Keywords: Blood-based Cancer Diagnosis; Biomarker Discovery; Feature Interactions; Sparse Learning; Aptamer; SOMAmer Prioritization.

1. Introduction

In this paper, we focus on the task of biomarker discovery and cancer diagnosis directly based on patient blood samples in the setting of limited training data. Although cancer diagnosis based on microarray datasets has been extensively studied, blood-based cancer status prediction is still a challenging problem, because complex diseases like cancers are the results of multiple genetic and epigenetic factors and their manifestation in blood samples is even more complicated than in tumor samples. It is very difficult to identify these complicated factors solely based on limited information provided by training data. Previous studies on single gene markers can provide valuable information about disease status prediction, but they offer limited insight into the complex interplay among the molecular factors responsible for progression of complicated diseases such as cancers. So, recently, research in complex diseases shifts towards the identification of multiple genes that interact directly or indirectly in contributing their association to the target disease. Several complex interactive partners from previous

*To whom correspondence should be addressed.

studies have been shown to give important insight into the mechanism of breast cancer[1] and colorectal cancer,[2] but none of these approaches addressed the problem of disease diagnosis based on blood samples or considered the multiplicative effect of gene/protein expressions.

The identification of groups of genes that show differential behavior in the manifestation of complex diseases is computationally infeasible due to the combinatorial nature of the problem. Several recent methods propose to reduce the search space using orthogonal prior knowledge about connections amongst the genes, such as interactions collected from protein-protein interaction (PPI) network[3] or grouping information from functional annotations of proteins. One notable computational method named Group Lasso[4] incorporates such prior interaction or grouping among the genes to detect gene groups that contribute to human disease, by enforcing sparsity at the group level in a supervised regression framework. Group Lasso is extended by Jacob *et al.*[5] to a more general setting that incorporates groups whose overlaps are nonempty. Such overlaps in groups is biologically significant, because many genes participate in multiple pathways and manifest themselves in several biological processes. Although (Overlapping) Group Lasso is very useful in identifying biologically relevant groups of genes and proteins, it cannot capture complex combinatorial relationships among the features within and across the groups, and it often outputs too many features as biomarkers. Also, current PPI network data is inherently noisy due to experimental constraints.[6] Algorithmic approaches based solely on these noisy prior information can result in many false positive interactions which are absent in the real genome space.

Our goal in this paper is to identify the complex combinations of single genes and multiplicative pairwise interactions among genes that help us (1) better perform cancer diagnosis based on blood samples, and (2) gain novel insights into the mechanistic basis of the diseases. Since the total number of possible pairwise human gene interactions is huge, it is computationally infeasible to examine all possible combinatorial combinations of them when trying to understand their relevance to the diseases under consideration. We propose a two-stage approach in a sparse learning framework, named as QUIRE, i.e. to detect QUadratic Interactions among infoRmative fEatures which show differential behavior for diagnosing a target disease using protein or mRNA expressions. Based on our own experimentally generated data from patient blood samples using a novel SOMAmer technology,[7] we apply QUIRE to blood-based cancer diagnosis for RCC and OVC, and we also apply QUIRE to a publicly available CRC dataset that can be used to prioritize our SOMAmer design. QUIRE can discover complementary sets of markers and pairwise interactions that can better classify samples from different stages of cancer and predict post-cancerous events, like cancer recurrence and death from cancer with higher accuracy than other state-of-the-art feature selection methods. To the best of our knowledge, QUIRE is the first proposed method to identify combinatorial patterns among the pairs of informative genes for studying complex diseases like cancers. Subsequent functional analysis of the interactions identified by QUIRE reveals that it can indeed identify genes relevant to the progression of diseases under study.

2. Problem and Method

The identification of single gene markers in a genome-wide case-control study is an ill-posed problem, because the number of genes in human cells is much larger than that of available samples. For such problems, Lasso, proposed by Tibshirani *et al.*[8] is very popular for selecting a small number of features relevant to the problem under study. When a set of features are highly correlated to each other, Lasso selects one from that set, ignoring others. So there is a possibility that Lasso leaves out biologically relevant genes from its set of selected informative features.

Suppose we have a data set S containing n samples and p gene features $(\mathbf{x}^i, y^i)$ with response variable $y \in R$ and feature vector $\mathbf{x} \in R^p$, where $i \in \{1, \ldots, n\}$, and we assume that the feature values and the ys are centered in S. The Lasso approach optimizes the following objective function,

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} w_j x_j^i)^2,$$

$$\ell_{lasso}(\mathbf{w}) = \ell(\mathbf{w}) + \lambda \sum_{j=1}^{p} |w_j|, \tag{1}$$

where $\ell(\mathbf{w})$ is the loss function of linear regression, and $\mathbf{w}$ is the weight parameter. The ℓ_1 norm penalty in lasso induces sparsity in the weight space for selecting features. It is obvious that the sum of the least squared errors and the ℓ_1 norm are convex functions with respect to the weights $\mathbf{w}$. Lasso has a global optimum, which can be identified by any convex optimization technique.

In spite of the computational efficiency and the popularity of Lasso for feature selection, its formulation prevents it from capturing any prior information on possible group structures among the features. Group Lasso[4] proposed using $\ell_{2,1}$ penalty to select groups of input features which are partitioned into non-overlapping groups. The group penalty is the sum of the ℓ_2 norm on the features belonging to the same group. Overlapping Group Lasso[5] extends Group Lasso to handle groups of features with overlapping group members by duplicating input features belonging to multiple groups in the design matrix. Because a lot of real applications involve overlapping feature groupings, Overlapping Group Lasso is a more natural choice than Group Lasso for biomarker discovery. Suppose that we partition p features in data set S into q overlapping groups $G = \{g_1, g_2, \ldots, g_q\}$, the following objective function is minimized,[5]

$$\ell_{oglasso} = \ell(\mathbf{w}) + \lambda \sum_{g \in G} ||\mathbf{w}_g||, \tag{2}$$

where λ is the regularization parameter, $\mathbf{w}_g$ denotes the vector of weights associated with features in group g, and $||\cdot||$ is the Euclidean norm. The above optimization problem is separable, so we can use block coordinate descent to optimize the weights associated with each group g separately. Although considering grouping structure among input features is very important for feature selection, Overlapping Group Lasso only encourages sparsity at the feature group level and there is no sparsity penalty within feature groups. Therefore, Overlapping Group Lasso often outputs a much larger number of selected features than Lasso. Furthermore, Lasso

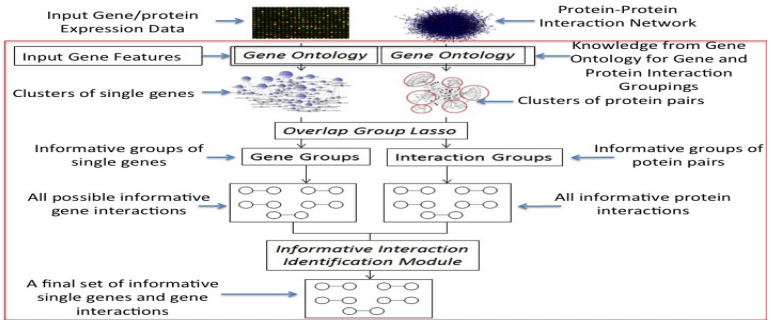

Fig. 1. Working model of QUIRE. QUIRE takes as input, gene or protein expression levels of a set of samples, disease status of those samples and physical interactions amongst the gene products. Then it uses gene ontology based functional annotation to group the genes and cluster the interaction network. Overlapping group lasso is run next on the expression and interaction space to identify informative set of genes and interactions. QUIRE then enumerates all pairwise binary interactions amongst the selected gene features. Finally the proposed novel objective function is applied on the selected single gene features, the informative protein protein interactions and the quadratic interactions amongst these genes to identify the final set of interactions and gene markers.

and Overlapping Group Lasso only consider single gene features for prediction, which is limited for disease status prediction and biomarker discovery as shown by our experimental results.

For cancer diagnosis and biomarker discovery from blood samples or tissue samples, we consider all possible combinations of single gene features and quadratic gene interaction features. Ideally, we want to optimize the following optimization problem to identify discriminative features given the dataset S,

$$\ell(\mathbf{w}, \mathbf{U}) = \sum_{i=1}^{n} (y^i - \sum_{j=1}^{p} w_j x_j^i - \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} U_{jk} x_j^i x_k^i)^2 + \lambda_1 \sum_{j=1}^{p} |w_j| + \lambda_2 \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} |U_{jk}|, \quad (3)$$

where $\mathbf{U}$ is the weights associated with pairwise feature interactions. However, the above model has $O(p^2)$ features and is not applicable to genome-wide biomarker discovery studies because p^2 is too large. Provided that the training data is often very limited, it is almost impossible to identify the discriminative single or quadratic interaction features by solving the above optimization problem. We propose QUIRE (QUadratic Interactions among infoRmative fEatures) to address these challenges, which is based on Overlapping Group Lasso and Lasso. And it takes advantage of both of these feature selection methods.

The underlying idea of QUIRE is to incorporate all possible complementary biological knowledge (see Figure 1) into the above intractable optimization problem to reduce search space. By restricting discriminative gene interactions to happen only between genes in some informative gene groups, we can use existing functional annotations of input genes to identify these groups thereby to throw away a lot of interaction terms during the optimization. In addition, available physical interactions between the protein products of input genes can also be used to cut the search space, although discriminative gene feature interactions for prediction do not always necessarily correspond to physical interactions. The general working model of QUIRE is shown in Figure 1. In details, QUIRE takes the expression profile of n samples over p genes (proteins), the physical interaction network among the genes products (i.e. protein-

protein interaction network) and the disease status of these samples as input, and it outputs a (small) set of discriminative genes and gene interactions with corresponding learned weights for predicting the disease status of any incoming test sample. The step-by-step working model of QUIRE is given below:

(1) *Functional group generation:*
   (a) QUIRE groups the p input gene features into q overlapping functional categories according to the existing Gene Ontology (GO) based functional annotations such as Cellular Colocalization (CC).
   (b) QUIRE clusters the given interaction network (i.e. PPI) into subsets of overlapping gene products based on CC.[b]

(2) *Informative genes and functional interactions selection:*
   (a) Given the GO functional grouping of input gene features, Overlapping Group Lasso is run to select m top discriminative genes for disease status prediction according to the absolute values of the learned weights of gene features.[c]
   (b) Overlapping group lasso is run on the clustered interaction network to select informative groups of protein-protein interactions. In this case, each cluster is considered as a group and the products of pairwise gene/protein feature values among the interacting proteins in a group are used as interaction feature values.

(3) *Selection of most informative interactions and genes:* QUIRE first enumerates all possible quadratic feature interactions among the informative genes selected at step 2(a). Then it takes these quadratic interactions, single informative gene features and the informative functional interactions identified at step 2(b) as input and it outputs the final selected gene interactions and single genes as biomarkers.

In order to identify the discriminative combinations of single gene features and quadratic interactions among pairwise informative genes, we define our proposed objective function for Lasso as follows,

$$\ell(\mathbf{w}, \mathbf{U}, \mathbf{R}) = \sum_{i=1}^{n}(y^i - \sum_{j=1}^{m} w_j x_j^i - \sum_{j=1}^{m-1}\sum_{k=j+1}^{m} U_{jk} x_j^i x_k^i - \sum_{l=1}^{r} R_l I_l)^2 + \lambda_1 \sum_{j=1}^{m}|w_j| + \lambda_2 \sum_{j=1}^{m-1}\sum_{k=j+1}^{m}|U_{jk}| + \lambda_3 \sum_{l=1}^{r}|R_l|, \quad (4)$$

where j and k index the m seed informative genes and l indexes the r informative protein-protein interactions selected by the Overlapping Group Lasso in the previous step, $\mathbf{U}$ and $\mathbf{R}$ are weights associated with feature interactions, and λ_1, λ_2, and λ_3 are regularization parameters[d]. The objective function contains ℓ_1 penalties at single gene level, pairwise gene interaction level, and protein interaction level. The intuition behind this formulation is that it captures

[b] We chose CC as final functional grouping of gene/protein features because it produces groups with reasonable size (see experiment section for details) and it is the most relevant annotation to blood-based diagnosis.
[c] m is selected by 5-fold cross validation.
[d] In our experiments, we make $\lambda_1 = \lambda_2 = \lambda_3$ and set it by 5-fold cross validation.

the interactions that are complementary to the individual informative genes. Because it is computationally infeasible to consider every pair of interaction in a genome-wide case-control study, QUIRE reduces the search space by using the features that are selected by Overlapping Group Lasso as the informative ones, and then it relies on Lasso with ℓ_1 penalties to identify the discriminative combination of informative individual gene features and gene interaction features, which provides an approximation to the problem of searching an exponential number ($O(2^{p+p^2})$) of all possible combinations of single features and pairwise interaction features. Our current implementation of QUIRE is based on the standard Lasso package from glmnet[9] and the Overlapping Group Lasso programs from Jacob *et al.*, 2009.[5]

3. Experimental Results and Discussion

In this section, we present experimental results of QUIRE on three different cancer datasets: blood-based cancer diagnosis and biomarker discovery for (1) Renal Cell Carcinoma (RCC) and (2) Ovarian Cancer (OVC) based on our private datasets, and cancer recurrence and death prediction for (3) Colorectal Cancer (CRC) based on a public microarray dataset, in which the identified markers can be used to prioritize our SOMAmer design. We compare the performance of QUIRE to the state-of-the-art feature selection techniques, Lasso, Overlapping Group Lasso and SVM. We then perform a literature survey and enrichment analysis of the informative interactions selected by QUIRE and show that they are relevant to the progression of the disease.

3.1. *Our Blood-based Datasets Generated by the SOMAmer Technology*

To predict cancer progression status directly from blood samples, we generated our own datasets[e]. All samples and clinical information were collected under Health Insurance Portability and Accountability Act compliance from study participants after obtaining written informed consent under clinical research protocols approved by the institutional review boards for each site. Demographic data was collected by self-report and clinical data by chart review. Blood was processed within 2 hours of collection according to established standard operating procedures. To predict RCC status, serum samples were collected at a single study site from patients diagnosed with RCC or benign renal mass prior to treatment. Definitive pathology diagnosis of RCC and cancer stage was made after resection. Outcome data was obtained through follow-up from 3 months to 5 years after initial treatment. To predict OVC status, plasma samples were collected from women with a suspicious pelvic mass prior to treatment. Definitive pathology diagnosis of ovarian cancer stage or benign mass was made after resection. CA-125 (Carbohydrate Antigen 125 also known as MUC16) was measured by a routine clinical laboratory assay. To generate RCC and OVC datasets, the SOMAmer based proteomic technology[7] is used to measure the concentration of a selected set of about 1000 proteins in Relative Fluorescence Unit. The CRC samples belong to a publicly available microarray dataset collected from gene expression omnibus (GEO), and referenced by accession number

[e]Due to conflict of interest, the datasets are not publicly available. Data requests should be sent to the last author of this paper.

$GSE17536$ (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17536).[10]

Our RCC dataset contains 212 RCC samples from benign and 4 different stages of tumor. Expression levels of 1092 proteins are collected. The number of Benign, Stage 1, Stage 2, Stage 3 and Stage 4 tumor samples are $40, 101, 17, 24$ and 31 respectively. Our OVC dataset contains 845 proteins' expressions for 248 samples across Benign and 3 different stages of ovarian cancer. The number of Benign, Stage 1, Stage 2 and Stage 3 tumor samples are $134, 45, 8$ and 61 respectively. The public CRC microarray dataset ($GSE17536$) contains 177 samples from 4 different stages (Stage 1 to Stage 4) of CRC. Expression levels of 20125 genes are collected. Besides stage information, this dataset also has records for each patient, the binary valued information of "Cancer Recurrence" and "Death from Cancer". Out of 177 patients, 55 had recurrence of cancer and 68 died from cancer.

In order to group the genes using gene ontology terms, we use the web based tool "Database for Annotation, Visualization, and Integrated Discovery" (DAVID, `http://david.abcc.ncifcrf.gov/`).[11] There are a set of parameters that can be adjusted in DAVID based on which the functional classification is done. This whole set of parameters is controlled by a higher level parameter "Classification Stringency", which determines how tight the resulting groups are in terms of association of the genes in each group. In general, a "High" stringency setting generates less number of functional groups with the member genes tightly associated and more genes will be treated as irrelevant ones into an unclustered group. We set the stringency level to "Medium" which results in balanced functional groups where the association of the genes are moderately tight. The total number of groups based on CC annotations for RCC and OVC datasets are 56 and 23 respectively, and the number of groups for the CRC dataset is 520.

Besides using it for selecting informative single gene features, we use Overlapping Group Lasso to select the informative protein protein interactions. We download the binary protein protein interactions (PPI) data from HPRD (`http://www.hprd.org/`). For each feature group G_i, we identify the pairs of member genes of G_i whose products interact directly with each other in the PPI network. The set of all such pairs where both interacting partners are members of G_i forms a group. For a pair of interacting genes x_j and x_k in a group, we use their quadratic interaction term $x_j x_k$ as their expression level. Usage of the quadratic interaction formulation in Overlapping Group Lasso helps us to integrate the resulting informative protein protein interactions into the formulation of QUIRE directly without any transformation. Thus the total number of groups are same in the case of interactions and single gene features. But the cardinality of each group and the expression levels of the members are different.

3.2. *Experimental Design*

We perform three stage-wise binary classification experiments using RCC samples: Classification of Benign samples from Stage $1-4$ samples, Classification of Benign and Stage 1 samples from Stage $2-4$ samples, and Classification of Benign, Stage $1, 2$ samples from Stage $3, 4$ samples. In the OVC dataset, *CA125* is a well-known marker in ovarian cancer.[12] Concentration of *CA125* is used to measure the progression of the disease. The suspicious cutoff level of *CA125* is 40 U/mL, meaning that concentration level above 40 of this marker might be

indicative of OVC. But *CA125* is not a good indicator of early detection of the disease onset, especially when the concentration of this biomarker is between 40 and 100.[13] So we use samples with *CA125* concentration level between 40 and 100 as our test set in this experiment. The remaining samples, with concentration of *CA125* below 40 and above 100 are used as training set. We perform the following experiments: Classification of Benign samples from Stage $1-3$ samples, Classification of Benign, Stage 1 samples from Stage $2,3$ samples, and Classification of Benign, Stage $1,2$ samples from Stage 3 samples. On the CRC dataset, we perform binary classifications to predict whether there is disease-free survival in the follow-up time or not for cancer recurrence prediction and whether there is death from CRC across all pathological stages of the disease for death from colorectal cancer prediction.

3.3. *Classification performance of QUIRE*

In this section, we report systematic experimental results on classifying samples from different stages of RCC and OVC and in predicting CRC recurrence and death from CRC. In the first stage of QUIRE, we use Overlapping Group Lasso to identify the biologically relevant groups of features and pairwise protein interactions, which in turn, is used in the subsequent stage to explore the set of informative markers and quadratic interactions. However, for the RCC and OVC datasets, we do not use protein protein interactions for prediction purpose. This is because, these datasets include only selected marker proteins distributed sparsely across the protein interaction network and thus most of them do not interact with each other directly.

After we run Overlapping Group Lasso on the gene groups, we sort the genes based on the weight value assigned to it by the algorithm. We used cross validation to select the optional parameter m in QUIRE from $\{100, 200, 300, 400, 500\}$, and $m = 200$ was selected for all our experiments. For classification of CRC samples, Overlapping Group Lasso on average selects 1000 PPIs as informative ones. We use this whole set of selected protein interactions as input to QUIRE to be considered besides the paired quadratic interactions.

The predictive performance of the markers and pairwise interactions selected by QUIRE is compared against the markers selected by Lasso, linear Support Vector Machine (SVM) and Overlapping Group Lasso. We use glmnet[9] and LiblineaR[14] packages for implementation of Lasso and SVM respectively. We use the Group Lasso implementation (with overlapping groups) from.[5] The overall performance of the algorithms are shown in Figure 2. In this figure, we report average AUC (Area Under the Curve) score for ten runs of five-fold cross validation experiments for cancer stage prediction in RCC (Figure 2(A)) and for predicting cancer recurrence and death from cancer in CRC(Figure 2(C)). In five fold cross validation experiments, we divide the samples equally into five disjoint sets or folds. We keep one fold for testing. On the remaining four folds, we use Overlapping Group Lasso to identify the informative set of markers and protein protein interactions (for CRC). We train QUIRE on these four folds using these markers to identify the pairwise interactions and markers and use the set-aside test set for prediction purpose. For each run, this procedure is repeated for each of the five folds and average AUC score is reported for ten such runs. For OVC, we report average AUC score (Figure 2(B)) for predicting the cancer stage of the samples with intermediate levels of *CA125* (concentration of *CA125* is between 40 and 100) using the

remaining samples for training and informative feature selection. In cancer stage prediction

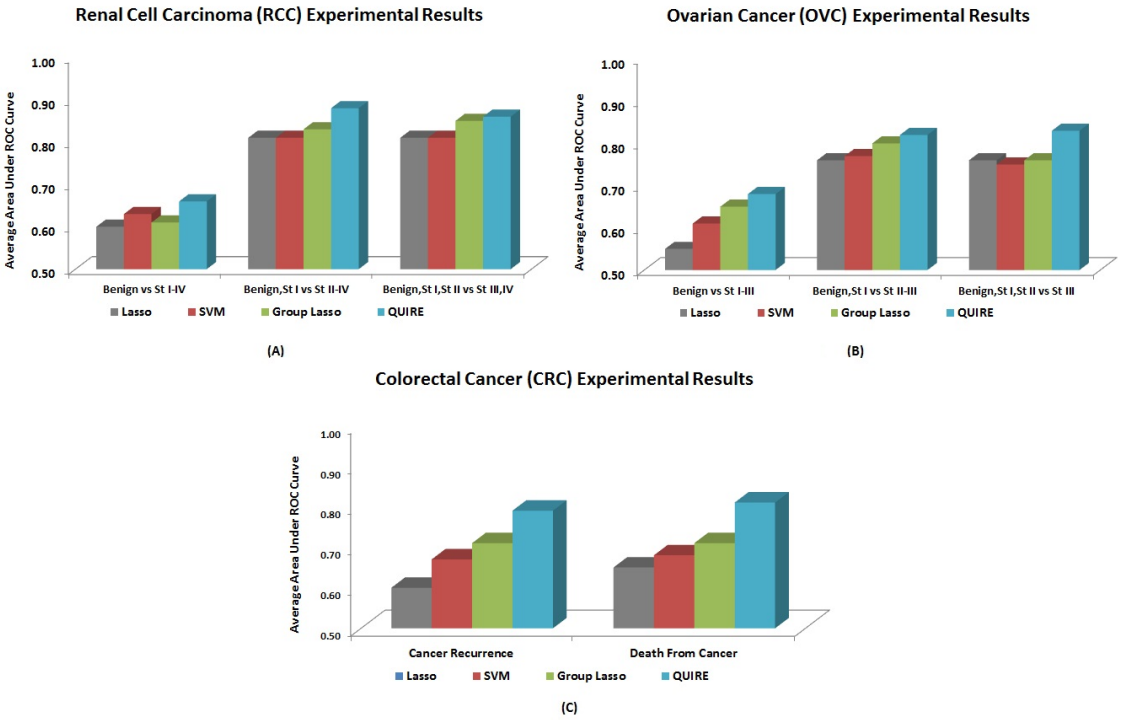

Fig. 2. Comparison of the classification performances of different feature selection approaches with QUIRE in identifying the different stages of (A)RCC , (B) OVC and (C) in predicting CRC recurrence and death from CRC. In (A) and (C), five fold cross validation is repeated ten times and average AUC score is reported. For (B), samples with *CA125* marker's expression level between 40 and 100 are used as test cases, while the remaining samples are used for training. This experiment is also repeated ten times and average AUC score is reported.

experiments for RCC and OVC, we see from Figure 2 that the combination of informative markers and pairwise interactions identified by QUIRE show better classification performance in every case, as compared to the markers selected by Lasso, SVM and Overlapping Group Lasso. For early detection of the diseases (classification of Benign, Stage 1 vs. rest of the samples), QUIRE achieves average AUC scores of 0.88 and 0.82 for RCC and OVC respectively. Overlapping group lasso shows next best performance with average AUC scores of 0.83 and 0.80 respectively. Lasso and SVM, which do not use any grouping or interaction information amongst the features, show the worst performance in all of the classification tasks apart from one. As QUIRE markers show consistently better performance across all the stages of RCC and OVC, they can be used for improved diagnosis and prognosis of these two different types of cancers. Also QUIRE helps better prediction of OVC progression for samples with intermediate levels of *CA125*; so it can be used by the physicians for early detection of this disease.

From Figure 2(C), we can see that gene-gene interactions help us better predict both CRC recurrence and death from CRC, as compared to the other feature selection mechanisms. In the events of cancer recurrence and death from cancer, the average AUC values achieved by features selected with QUIRE are 0.79 and 0.81 respectively, while markers identified by Overlapping Group Lasso show the next best performance with average AUC value of 0.71 in

both of these categories. Markers identified by Lasso show the worst performance in prediction of both of these events. The performance gap between QUIRE and the other three popular feature selection techniques hint to the fact that QUIRE can identify interactions that might help us better understand the mechanistic basis of CRC.

These experimental results show that QUIRE identifies markers and interactions that complement each other in such a way that they not only help better diagnosis and prognosis of cancer, but also can predict the advanced events of recurrence of cancer and survival after cancer with higher accuracy than other state-of-the-art algorithms. For each of these datasets, identification of informative pairwise interactions using brute-force enumerative technique is computationally impractical due to the huge dimensionality of the search space. QUIRE helps reducing this space by a large margin. The total running time of QUIRE is dominated by the Overlapping Group Lasso stage which takes around one hour to identify biologically relevant groups of genes and protein interactions in traditional desktop computers for the types of problems we study. After the dimensionality is reduced, QUIRE exhaustively enumerates all the pairwise interactions and use the protein interactions identified in the previous stage on this low dimensional space in a couple of minutes. QUIRE is computationally very fast considering that it identifies discriminative pairwise gene interactions at a genome-wide scale.

3.4. *Informative QUIRE markers and interactions associated with cancer*

Cancer is a genetic disease, which originates and develops through a process of mutations. Mutations in individual gene not only disrupts its own function, but also affects its interaction patterns with other genes. As complex diseases like cancer is a result of dysregulation in the interactions among the genes, researchers focus on identifying those relevant interactions to gain more insight into the molecular basis of the disease. On the CRC dataset, QUIRE selects about 120 quadratic interactions and single features in total on average as informative ones for both CRC recurrence and death from CRC. On the other hand, the average number of markers selected by Overlapping Group Lasso and Lasso on the same prediction tasks are about 1100 and 150 respectively. Therefore, Overlapping Group Lasso itself is unsuitable for biomarker discovery although it produced the second best performance.

An investigation of the pairwise interactions identified by QUIRE on CRC dataset reveals that many of these interactions are indeed relevant to the progression of cancer in general. Some of such interactions identified for prediction of CRC recurrence include *JAK2* - *LYN*,[15] Transforming growth factor beta ($TGF\beta$) - *SMAD*,[16] Epidermal growth factor receptor (*EGFR*) - Caveolin (*CAV*),[17] *TP53* - TATA binding protein (*TBP*),[18] Connective tissue growth factor (*CTGF*) - Vascular endothelial growth factor (*VEGF*),[19] Edoglin (*ENG*) - Transforming growth factor beta receptor ($TGF\beta R$).[20] Further investigations of the interactions identified by QUIRE might reveal novel gene partners associated with cancer and thus lead to testable hypothesis.

Disturbance in pairwise interactions among the genes affects the pathways in which they are located in. Cancer pathways are a set of pathways dysregulations in which have been shown to be associated with initiation and progression of the disease. A pictorial view of the well-known cancer pathways can be found in the KEGG database(`http://www.genome.jp/kegg/pathway/hsa/hsa05200.html`).[21] We per-

form a pathway enrichment analysis where we test if the set of the markers and interactions identified by QUIRE on the CRC dataset reside in the cancer pathways. As part of this experiment, we first use the partner genes identified by QUIRE as part of the informative interactions while predicting CRC recurrence. We use DAVID to identify the statistically significant pathways that are enriched in these genes. An investigation of the enriched pathways returned by DAVID indicates that many of them are indeed responsible for cancer or related to functions dysregulation in which results in cancer. Some of such KEGG pathways include Apoptosis (p-value $4.7\text{x}10^{-4}$), Focal adhesion (p-value $3\text{x}10^{-3}$), Cell adhesion molecules (p-value $9.2\text{x}10^{-4}$), p53 signaling pathway (p-value $1.3\text{x}10^{-2}$), Gap junction (p-value $1.3\text{x}10^{-2}$), MAPK signaling pathway (p-value $4.5\text{x}10^{-2}$), ErbB signaling pathway (p-value $5.8\text{x}10^{-2}$), Cell cycle (p-value $6.6\text{x}10^{-2}$), Pathways in Cancer (p-value $7.2\text{x}10^{-4}$), Colorectal cancer (p-value 10^{-3}). Repeating the same analysis on the interacting partners identified by QUIRE while predicting "Death from CRC" result in identification of similar pathways (data not shown).

3.5. *Significance of feature interactions in QUIRE*

We also perform classification experiments to observe the performance of PPIs and informative single features on predicting CRC recurrence and death from CRC without quadratic feature interactions. For this experiment, we use the single gene markers and the PPIs selected by Overlapping Group Lasso as input to QUIRE and enumeration of the pairwise interactions among the marker genes is avoided. For ten runs of five fold cross validation experiment on this modified feature set, we observe average AUC score of 0.71 for both classification tasks. If we only use informative single features with the same experimental setting, the average AUC score we got is 0.70. These results show that besides physical interactions and single features, indirect higher level interactions among the informative genes must be taken into account to understand the basic mechanism of complex diseases.

4. Conclusion

In this paper, we propose a computational approach, QUIRE, to identify combinatorial interactions among the informative genes in complex diseases, like cancer. Our algorithm uses Overlapping Group Lasso to identify functionally relevant gene markers and protein interactions associated with cancer. It then explores the pairwise interactions among these relevant genes within this reduced space exhaustively and the selected pairwise physical protein interactions to discover the combination of individual markers and gene-gene interactions that are informative for prediction of the disease status of interest. The application of QUIRE on three different types of cancer samples collected using two different techniques shows that our approach performs significantly better than the state-of-the-art feature selection methods such as Lasso and SVM for biomarker discovery while selecting a smaller number of features, and it also shows that our approach can capture discriminative interactions with high relevance to cancer progression. Further investigations show that QUIRE can identify markers and interactions that have been associated previously with pathways associated with cancer. Moreover, the good performance of QUIRE on the CRC dataset suggests that applications of QUIRE on genome-wide microarray experimental data can be used to help prioritize SOMAmer design

for blood-based cancer diagnosis. And QUIRE applied to blood-based experimental data has the great potential to impact the field of practical medical diagnosis.

Acknowledgement

We thank Hans Peter Graf for valuable comments and discussions.

References

1. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee and T. Ideker, *Mol. Syst. Biol.* **3**, p. 140 (2007).
2. S. A. Chowdhury, R. K. Nibbe, M. R. Chance and M. Koyuturk, *J. Comput. Biol.* **18**, 263 (Mar 2011).
3. S. Lee and E. P. Xing, *Bioinformatics* **28**, i137 (June 2012).
4. M. Yuan and Y. Lin, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49 (2006).
5. L. Jacob, G. Obozinski and J.-P. Vert, Group lasso with overlap and graph lasso, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (ACM, New York, NY, USA, 2009).
6. H. Yu, P. Braun, M. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis *et al.*, *Science* **322**, 104 (2008).
7. L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, J. Carter, A. B. Dalby, B. E. Eaton and T. Fitzwater et al., *PLoS ONE* **5**, p. e15004 (12 2010).
8. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, pp. 267 (1996).
9. J. H. Friedman, T. Hastie and R. Tibshirani, *Journal of Statistical Software* **33**, 1 (2 2010).
10. J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, S. Eschrich, C. Kis, S. Levy, M. K. Washington, M. J. Heslin, R. J. Coffey, T. J. Yeatman, Y. Shyr and R. D. Beauchamp, *Gastroenterology* **138**, 958 (Mar 2010).
11. G. Dennis Jr, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane and R. Lempicki, *Genome Biol* **4**, p. P3 (2003).
12. K. S. Suh, S. W. Park, A. Castro, H. Patel, P. Blake, M. Liang and A. Goy, *Expert Rev. Mol. Diagn.* **10**, 1069 (Nov 2010).
13. E. L. Moss, J. Hollingworth and T. M. Reynolds, *Journal of clinical pathology* **58**, 308 (March 2005).
14. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang and C. J. Lin, *Journal of Machine Learning Research* **9**, 1871 (2008).
15. A. Samanta, S. Chakraborty, Y. Wang, H. Kantarjian, X. Sun, J. Hood, D. Perrotti and R. Arlinghaus, *Oncogene* **28**, 1669 (2009).
16. W. Grady, *Clinical cancer research* **11**, 3151 (2005).
17. K. Dittmann, C. Mayer, R. Kehlbach, H. Rodemann *et al.*, *Mol Cancer* **7**, 17 (2008).
18. D. Crighton, A. Woiwode, C. Zhang, N. Mandavia, J. Morton, L. Warnock, J. Milner, R. White and D. Johnson, *The EMBO journal* **22**, 2810 (2003).
19. I. Inoki, T. Shiomi, G. Hashimoto, H. Enomoto, H. Nakamura, K. Makino, E. Ikeda, S. Takata, K. Kobayashi and Y. Okada, *The FASEB Journal* **16**, 219 (2002).
20. E. Fonsatti, M. Altomonte, P. Arslan and M. Maio, *Current drug targets* **4**, 291 (2003).
21. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi and M. Tanabe, *Nucleic acids research* **40**, D109 (2012).