Identifying Gene Subnetworks Associated with Clinical Outcome in Ovarian Cancer Using Network Based Coalition Game

Abolfazl Razi[1], Fatemeh Afghah[2], Vinay Varadan[1]
[1]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH;
[2]ECE Department, North Carolina A&T State University, Greensboro, NC

***Abstract*— The problem of identifying interacting genes that jointly are associated with a phenotype is considered. When the number of features are extremely large compared to the number of samples, there may be several subsets of features that provide acceptable levels of predictability. This is particularly true in cancer genomics, where we are interested in finding functionally related gene sets likely to jointly drive cancer phenotypes.**

In this paper, a novel game theoretic solution is proposed by modeling genes as players of a Coalition Game. This method discovers and develops informative gene subnetworks by integrating gene expression profiling of cancer tissues with protein-protein interaction (PPI) networks. These subnetworks are gradually developed by selective addition of candidate genes that present maximal Shapely values in coalition with subnetworks of genes. We applied the proposed algorithm to an ovarian cancer dataset (N = 201), in order to identify optimal subnetworks that can predict cancer progression risk in response to platinum-based therapy. We show improved predictive power of the proposed method when compared to state-of-the-art feature selection methods, with the added advantage of identifying potentially functional gene subnetworks that may provide insights into the mechanisms underlying cancer progression.

I. INTRODUCTION

A critical problem in cancer research involves the identification of functionally connected sets of genes that jointly drive cancer development, progression and therapy response. Large-scale cancer profiling studies have enabled the measurement of thousands of genes across samples, but only a few tens of genes at most are likely to be relevant to the underlying biological process.

Dimensionality reduction using discriminative component analysis methods such as CCA, LDA and IDA are developed to project data into new subspaces, where a few components bear the most discriminative information about data, hence simplifying data storage, prediction and interpretation [1]. Although very efficient in dimensionality reduction, these methods are not ideally suited for the identification of genes driving cancer progression, since the predictors are provided in the transformed subspace.

Explicit feature selection methods are divided to wrapper methods and filtering methods. In filtering methods, the predictors are chosen based on their strong connection to the labels with less connection among the features using various geometric or information theoretic measures, whereas in wrapper methods the features are chosen based on their impact to the classifier. Wrapper methods require exhaustive search, thereby are computationally expensive. The filter methods with geometric distance measures are very fast but not capable of capturing non-linear relations. On the other hand, information theoretic filtering methods are very powerful but become computationally expensive. Further, they require large number of samples in order to obtain reliable empirical information-theoretic measures [2]. In cancer genomics, we are interested in methods that incorporate gene-interaction information such as protein-protein interaction (PPI) or biological pathway network databases in the feature-selection process, in order to identify functionally related sets of genes that jointly discriminate between phenotypes [3], [4].

In this work, we develop a Game theoretic solution that develops pathways emerging from a seed gene set in PPI network by traversing the network to discover the most informative pathways associated with a desired outcome. This algorithm reports a set of compact subnetworks that collectively modulate specific biological processes associated with the outcome, thus facilitating the development of biomarkers using core representative nodes within the identified subnetworks, as opposed to measuring all the genes individually.

II. COALITION GAMES REVIEW

In this section, we review the coalition games and their application in choosing subset of features considering their synergic predictive power. Coalition game is a class of games, where the players cooperate with one another by forming coalitions [5]. Coalition games have been recently utilized in feature selection problems to account for the relevance among potentially effective combinations of the features as well as providing a quantitative measure of the impact of each feature on the overall prediction [6] [7]. In this work, we propose a novel Network Based Coalition Game (NBCG) algorithm, where the game players are gene subnetworks extracted from the networks. In this algorithm, the game players are subnetworks which are not fixed, but rather developing identities over the game iterations by picking up new genes from the PPI network.

Let N_L be the number of players, $P = \{P_1, \dots, P_N\}$ be the set of players and S denote a coalition set, $S \subseteq P$. The total payoff that can be gained by the members of coalition S is defined by characteristic function $v(S)$. The game solutions are determined with possible scenarios that the players can form coalitions and how the total payoff of a coalition is divided amongst the coalition members.

Different possible coalitions of genes and pathways are examined to recognize the optimal classification features. Payoff of each coalition *S*, *v(S)*, measures the contribution for a coalition to the performance of the classifier (e.g. success rate in supervised learning). If feature i joins a coalition S, it

may improve the classification capability of this coalition. This is called marginal importance and is defined as;

$$\Delta_i(S) = v(S \cup \{i\}) - v(S). \qquad (1)$$

This marginal importance of a player does not reflect a fair share of the player from the characteristic function, since it depends on the order of the players in forming the coalition. Shapely value assigns a fair quantity for each player based on the average contribution of the player among all possible coalitions with all possible permutations [8]. Formally, the Shapley value of player $i \in P$ denoted by $\gamma_i(v)$ is defined as the expected marginal importance of player i to the set of players who precede this player.

$$\gamma_i(v) = \frac{1}{N_L!} \sum_{\pi \in \Pi} \Delta_i(S_i(\pi)), \qquad (2)$$

where Π is the set of all $N_L!$ permutations of P and $S_i(\pi)$ is the set of players (features) preceding player i in subset S with permutation π. Since the order of features inside a coalition does not change the coalition power, the calculations in (2), can be further simplified by excluding the permutation inside coalitions in the average, resulting in the following equation:

$$\gamma_i(P, v) = \frac{1}{N_L!} \sum_{S \subseteq N_L \backslash i} \Delta_i(S) |S|_i (N_L - |S| - 1)!, \quad (3)$$

where $S \subseteq N_L \backslash i$ presents the coalitions to which player i does not belong. $|S|_i$ and $|S|_i(N_L - |S| - 1)$ correspond to permutations of the preceding players and the subsequent players, respectively.

In applications with a large number of players, computation of Shapley value for all possible feature coalitions may be computationally intensive. Therefore, we utilize the multi-perturbation Shapley value (MSA) measurement, which is determined using an unbiased estimator based on Shapley value by using sampled permutations of players that form coalitions up to size N_L'. The idea behind this method is that coalitions of size $N_L' < N_L$ are capable of capturing the synergic power of players and larger coalitions only present additive power [8]. Therefore, we use

$$\gamma_i'(v) = \frac{1}{\left|\Pi_{N_L'}\right|} \sum_{\pi \in \Pi_{N_L'}} \Delta_i(S_i(\pi)), \qquad (4)$$

where $\Pi_{N_L'}$ denotes the sampled permutation on sub-groups of players of size N_L'. In this work, we use the approximate method of MSA with $N_L' = 4$.

III. Proposed Algorithm

In order to develop our algorithm, we note that the genes and their corresponding protein products directly or indirectly interact with one another as part of underlying biological processes. Protein-Protein Interaction (PPI) networks are developed to translate these complex biological processes into an undirected Boolean graph, where the nodes are genes (or their corresponding product proteins) and the edges represent biological interactions between the connected nodes. We use the human PPI network [9] and represent it as a $G \times G$ binary matrix A, *where G =12126* is the number of genes.

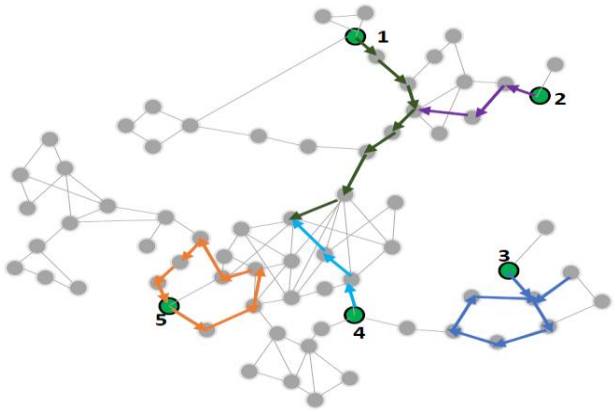


Fig. 1: An example of developing 5 modulated subnetworks over a PPI network. The initialization seed genes are marked with green color. The paths may develop chain, star and loop configurations.

A. *Initializations*

To select the subset of enriched genes with a phenotype, we start with initial set of $N_L - 1$ genes, from which the subnetworks emerge. The seed genes may be chosen randomly or using prior biological knowledge. For instance, one may choose the genes that are most frequently mutated in the cancer being studied, or genes whose expression levels are highly correlated with the phenotype. While these initializations may provide good initial gene sets, they perhaps prevent the discovery of subnetworks that are formed by genes that collectively but not individually impact the desired phenotype. The other option is to be agnostic of the phenotype but rather choose the genes from hot-spot points in the PPI network using degree or in betweenness measures. This initialization has the advantage of more flexibility and shorter access paths from seed genes to the informative subnetworks. Therefore, we use this method and choose the seed genes randomly among top 100 hot-spot genes in the PPI network.

B. *Coalition Evaluation*

The proposed algorithm uses classification accuracy as characteristic function to evaluate the impact of a coalition of genes on predicting therapy response, as detailed in Section IV. To evaluate a subset of features, we train a classifier using a training dataset and then evaluate classifier success rate using the unseen test dataset. The prediction success rate defines the payoff of the test gene set, from which the marginal importance and Shapley value for each player is obtained.

C. *Subnetwork Development Based on Shapley Value*

The game starts with N_L seed genes as players, as depicted in Fig. 1. Then, a path is emerged from each gene sequentially. For each seed gene $i \in \{1,2, \dots, N_L\}$, we form a set of directly connected neighbors in the PPI network $\Omega_i = \{k | A(i,k) = 1\}$. Each node in Ω_i is a candidate to evolve a path from that corresponding seed gene. Then, we evaluate the Shapley value for each candidate gene $j \in \Omega_i$, as the expected marginal importance of the gene j when forming coalitions with all of the other players. The gene k with the maximum Shapley value $\underset{k \in \Omega_i}{\operatorname{argmax}} \gamma_k(v)$ is chosen to join the subnetwork i. Once

a subnetwork is evolved from a seed gene, it takes over the role of player *i* in the game replacing the corresponding seed gene. Therefore, the number of players does not increase with game evolution. We note that the computational complexity of the algorithm is exponentially proportional to the number of players, hence the proposed method of representing players with N_L subnetworks rather than G individual genes significantly decrease the computational cost, since we have $(2^{N_L} \ll 2^G)$. In fact, the coalition-based game solution using genes as players is computationally infeasible. The idea behind this method is that we do not exclude genes from an already formed subnetwork, since it breaks down the subnetwork which is contradictory to the goal of finding informative subnetworks.

We repeat this procedure for the rest of the seed genes until they are replaced by two-node subnetworks. Then, this process is continued using the subnetworks as new game players. At each iteration, each subnetwork collects a single gene from its neighborhood set that presents a maximum Shapley value. The subnetworks are allowed to join one another to develop star configurations. We repeated the procedure until the stop criteria are met for all developed subnetworks. If the Shapley value for all neighbor genes are negative for a path, we stop extending the subnetwork and set link stop flag *LS(i)=1* to avoid collecting uninformative genes. The subnetwork evolution is also stopped if it crosses itself by choosing a gene which is already part of the subnetwork. We also terminate each subnetwork if the number of its nodes L_i reaches a predefined parameter L_{max}. The algorithm stops, when all subnetworks are frozen or the collective prediction power of all subnetwork gene members denoted by *Acc* reaches a predefined accuracy level, $Acc_{\max}$. This algorithm is summarized below.

Algorithms NBCG: Reports Informative Subnetworks

Inputs: N_L, L_{max}, Acc_{max}, A

Output: N_L Subnetworks P_i , Acc

---

1) Randomly choose N_L seed genes g_i from top 100 hot-spot nodes in A

2) For i=1 to N_L

  a) *Set* $P_i = \{g_i\}$, *set LS(i)=0;* $L_i = 1$*; end*

3) End

4) Loop

  a) For i=1 to N_L

    i) Generate Left and Right Neighbor List Ω_i^L , Ω_i^R

    ii) Incerement Number if iteration: *L_i =L_i+1*

    iii) Find best left neighbor : $l_{opt} = \underset{l \in \Omega_i}{argmax}\, \gamma_l(v)$

    iv) If $\gamma_{l_{opt}}(v) \geq 0$

      $P_i \leftarrow l_{opt} \cup P_i$

    v) If $\gamma_{l_{opt}}(v) < 0$ or $L_i > L_{max}$

      Set $LS(i)_i \leftarrow 1$

    vi) Calculate Acc=Acc($P_1 \cup \ldots \cup P_{N_L}$)

  b) End For

  c) If $\prod_{i=1}^{N_L} LS(i) > 0$ *or* $Acc \geq Acc_{max}$

    i) Report P_i and Acc

    ii) Exit Loop

5) End Loop

IV. RESULTS

In this section, the proposed algorithm is utilized to find the gene subnetworks that significantly impact therapy response in ovarian cancer. The data is obtained from The Cancer Genome Atlas (TCGA) dataset [10] and includes 201 cancer samples with their gene expression levels and clinical response. The clinical response data includes survival information (death or cancer progression) after platinum-based chemotherapy. We first divide the samples into two cohorts of poor and good survival rates. The poor survival cohort includes samples with events during the first 6 months of receiving platinum therapy, excluding patients who left the study (censored samples). Patients who survive at least 6 months without cancer progression are included in the good survival cohort.

We run the proposed algorithm using the following parameters of the game set as: number of subnetworks $N_L = 5$, maximum group size $N_L^{'} = 3$, and maximum subnetwork length $L_{max} = 20$. The algorithm reports a collection of subnetworks that are highly associated with the survival outcomes. The proposed solution can be integrated with any classification method, where the prediction power of the classifier is used as the characteristic function of a coalition under test. In this work, we arbitrarily use the SVM classification with RBF kernel. However, the obtained results is not sensitive to the choice of classifier and the numerical results show negligible change when using other classifiers (such as Random Forest, Naive Bayes, Bayes Net, and KNN).

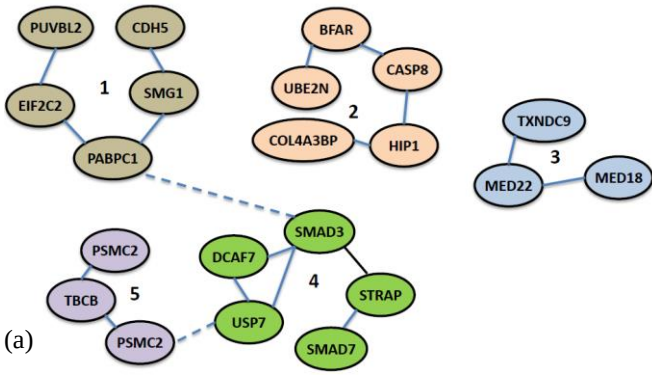
We compare the results with the same number of genes obtained using two benchmark solution categories. We apply state-of-the-art feature selection methods including Correlation based subset evaluation (CFS), Chi-square test based subset evaluation (Chi-Square) and mutual-information based subset evaluation method (Gain-Ratio). Additionally, two representative wrapper methods including Best First Search (BFS) method with Naive-Bayes and ranker method with SVM classifier were also applied. These methods report the most informative genes that may or may not belong to connected subnetworks. We also compare the proposed method with a network-based traversal method, where the subnetworks are initiated from the same initial gene seeds as in our proposed method. Instead of using the Shapely value, genes from the connected subnetwork in proximity of the seed genes are selected using a random walk until it collects the same number of genes as the proposed method. This whole procedure is repeated 100 times and the set of gene-subnetworks which provides the best prediction accuracy is selected for comparison with the proposed methodology.

In order to compare the relevance of the obtained gene sets across methods, in addition to the classification accuracy based on phenotype survival rates, we also compare the discriminative power of the gene sets in terms of survival probability. Therefore, patients are clustered using K-means clustering based on the gene expression data for the selected genes by the different methods. Then, we estimate the survival probability for each cluster using the standard method of Kaplan-Meier estimation followed by survival difference estimation using the log-rank test method that provides the probability of obtaining such a difference purely by chance (p-value).

Table I. Comparison of genes selected using the proposed method and other state-of-the-art feature selection methods based on its prediction accuracy and survival outcome separation. The results are corresponding to the first 18 genes reported by each method.

Method	Log-rank test p-value	Prediction Success Rate
CFS	0.01814	0.6488
Chi-Square	0.25505	0.6667
Gain-Ratio	0.47773	0.5179
Best First Search	0.07646	0.5714
SVM: Ranker	0.09190	0.5714
Optimal Random-Walk	0.08060	0.6190
Proposed NBCG	**0.00004**	**0.7262**

The results of these comparisons are provided in Table. 1. For competitor methods, where the sorted list of genes are provided based on various geometric distance or information based measures, we choose the same number of the top-genes that are reported by our method, which is 18. It is seen that the proposed method outperforms the other methods. The main cause is that the proposed coalition-based solution considers the collective power of gene sets based on Shapley value concept. This is in particular interesting, since the competitor methods are not restricted to choose the genes from connected subnetworks. The proposed solution provides more insightful and clinically relevant gene subnetworks.



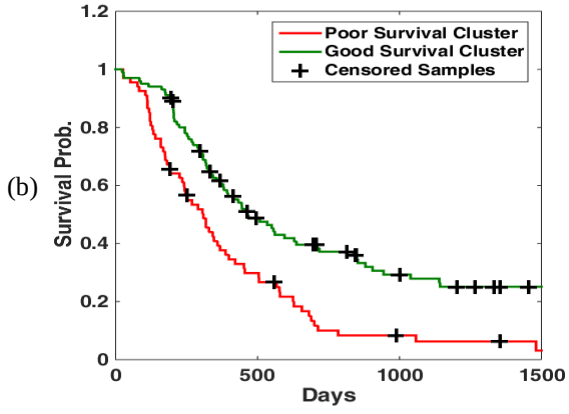
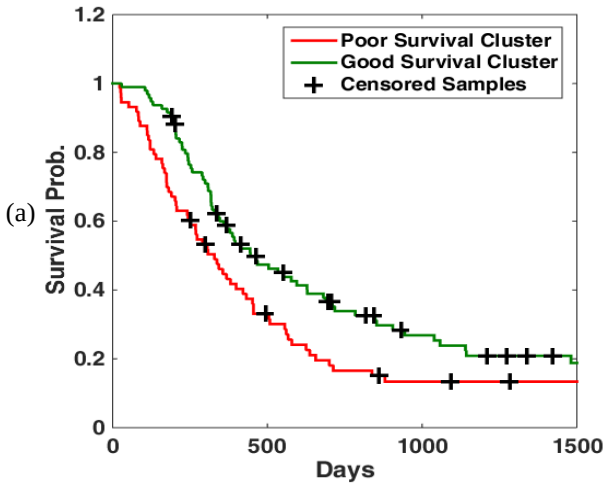


Fig. 2. (a) Sample subnetworks reported by the proposed algorithm. (b) Platinum-free survival, p-value= 4×10^{-5}.



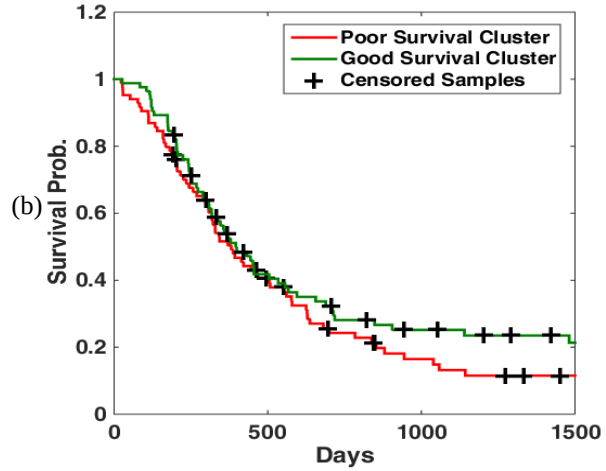


Fig. 3: Survival probability obtained by Kaplan-Meier Estimate for the cancer samples clustered using the genes that are selected by (a) best classification method (CFS), p-value=0.018 and (b) optimal network based random walk method, p-value = 0.08.

The subnetworks identified by our proposed algorithm for $N_L = 5$ are depicted in Fig. 2a. Subnetworks 1,2, and 4 correspond to i) vascular endothelial regulation, ii) cell cycle progression and apoptosis and iii) TGFb signaling pathways, respectively. These pathways belong to well-known hallmarks of cancer, thus suggesting that our proposed methodology is able to identify potentially functional pathways mediating therapy resistance. The subnetwork 5 joins subnetwork 4 at iteration 3 (dashed line) and subnetwork 3 stops extending at iteration 3, since no new informative neighbor genes were available.

Fig. 2b presents the survival curves for patients clustered into two groups using K-means clustering based on the expression level of the genes obtained from the proposed game-theoretic method (Fig 2a). The result demonstrates that the proposed solution can identify gene subnetworks with higher survival discriminatory power as compared to the estimates from the best feature selection method, which is CFS method for this case (Fig. 3a), and the Optimal Network-based Random-walk solution (Fig. 3b).

V. CONCLUSIONS

A novel Coalition based algorithm is developed using PPI networks to identify gene subnetworks that predict therapy-response in ovarian cancer. This method overcomes the benchmark feature selection methods and provides a collection of subnetworks that predict patient survival after platinum-based chemotherapy. The proposed method has the advantage of using PPI networks to identify functionally related gene sets that jointly discriminate patient outcomes. Additionally, this approach takes into account the collective power of subnetworks using the concept of Shapley value, as opposed to techniques that grow each subnetwork individually. The resulting subnetworks identified by the method could allow for the identification of functionally related subnetworks that are associated with cancer phenotypes, thus enabling the discovery of novel biomarkers and therapeutic targets in cancer.

REFERENCES

[1] A. Sarveniazi, "An Actual Survey of Dimensionality Reduction," American J. Comput. Mathematics, vol. 4, no. 2, pp. 55-72, 2014.

[2] Y. Saeys, I. Inza and P. Larranage, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[3] J. Soul, T. Hardingham, R. Boot-Handford and J. Schwartz, "PhenomeExpress: A refined network analysis of expression datasets by inclusion of known disease phenotypes," Scientific Reports, 2015.

[4] H. Chuang, E. Lee, Y. Liu, D. Lee and I. T, "Network-based classification of breast cancer metastasis," Molecular Systems Biology, vol. 3, no. 140, 2007.

[5] D. Ray, A Game-Theoretic Perspective on Coalition Formation, New York: Oxford University Press, 2007.

[6] A. Razi, F. Afghah, A. Belle and K. Ward, K. Najarian, "Blood Loss Severity Prediction using Game Theoretic Based Feature Selection," in IEEE-EMBS Int. Conf. on Biomedical and Health Informatics (BHI), pp. 776-780, 2014.

[7] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen and X. and Liu, "Feature Evaluation and Selection with Cooperative Game Theory," Pattern Recognition, vol. 45, no. 8, p. 2992–3002, 2012.

[8] L. S. Shapley, "A value for n-person games," in Contributions to the Theory of Games, volume II, H.W. Kuhn and A.W. Tucker (eds.), Princeton University Press, 1953, pp. 307-317.

[9] S. Razick, G. Magklaras, I. Donaldson, "iRefIndex: a consolidated protein interaction database with provenance," BMC Bioinformatics, vol. 9, no. 405, , 2008.

[10] "The Cancer Genome Atlas," [Online]. Available: http://cancergenome.nih.gov/.