PRER: A Patient Representation with Pairwise Relative Expression of Proteins on Biological Networks

Halil İbrahim Kuru[1], Mustafa Buyukozkan[1], and Oznur Tastan[2], *

Department of Computer Engineering, Bilkent University, 06800 Ankara
Faculty of Natural Sciences and Engineering, Sabanci University, 34956 Istanbul
⋆ Corresponding author, otastan@sabanciuniv.edu

Abstract. Alterations in protein and gene expression levels are often used as features to predictive models such as clinical outcome prediction. A common strategy to combine signals on individual proteins is to integrate alterations with biological knowledge. In this work, we propose a novel patient representation where we integrate the expression levels of proteins with the biological networks. Patient representation with PRER (Pairwise Relative Expressions with Random walks) operates in the neighborhood of a protein and aims to capture the dysregulation patterns in protein abundance for proteins that are known to interact. This neighborhood of the source protein is derived using a biased random-walk strategy on the network. Specifically, PRER computes a feature vector for a patient by comparing the protein expression level of the source protein with other proteins' levels in its neighborhood. We test PRER's performance through a survival prediction task in 10 different cancers using random forest survival models. PRER representation yields a statistically significant predictive performance in 8 out of 10 cancer types when compared to a representation based on individual protein expression. We also identify the set of proteins that are important not because of alteration of its expression values but due to the alteration in their pairwise relative expression values. The set of identified relations provides a valuable collection of biomarkers with high prognostic value. PRER representation can be used for other complex diseases and prediction tasks that use molecular expression profiles as input. PRER is freely available at: https://github.com/hikuru/PRER

Keywords: Cancer, Patient representation, Expression, Protein-protein interaction network, Survival prediction

1 Introduction

With the advances in the sequencing technologies, large scale molecular profiling of patients has become possible. The comprehensive profiling of cancer patients, along with the available patient clinical data, presents an opportunity to gain deeper insights into cancer and develop prediction tools for diagnostic, prognostic, and therapeutic purposes. Machine learning has been an instrumental tool for realizing this aim. In these studies, patients are often represented with their molecular data, such as gene expression profiles encoded as numerical feature vectors. For example, Yuan et al. [1] assess the utility of different types of molecular alterations for survival prediction. While using miRNA, protein, or mRNA expression, they use the expression values of these entities as input. Others follow a similar approach for different clinical outcome prediction tasks [2, 3, 4].

Molecular entities such as genes and proteins interact to carry out their functional roles in the cell, and phenotypes arise from these functional interactions. Based on this basic principle, alternative approaches, where the patient molecular profiles are integrated with the prior knowledge of molecular interactions, have been proposed (reviewed in [5] and [6]). A network of interactions helps to aggregate the signals attached to a single gene or a protein in a biologically principled way. Integration of the expression profiles of genes and their interactions are used in multiple studies [7, 8, 9, 10]. Chuang et al. [7] first identify discriminant and highly altered subnetworks of interactions using gene expression data and use the activity summaries of genes on these subnetworks as features for metastasis prediction. By assessing the association of pathways and transcription factors with overall survival as opposed to individual genes, Crijns et al. [9] identify signaling pathways and transcription factors that contribute to the clinical outcome of ovarian cancer. Taylor et al. [8] integrate a PPI network with a co-expression profile and report that the genes with dysregulated neighbors in that PPI network are potential prognostic markers. NetBank [11] uses gene expressions and the prior knowledge network to rank the genes in their relevance to the outcome of pancreatic cancer. All of these studies aggregate the alterations in a subnetwork, pathway, or network by summing or diffusing them in the network without the relative expression changes.

There is a limited number of studies that use the pairwise comparisons of molecular alterations instead of aggregation. Geman et al. report a method that uses the pairwise ranks of mRNA expression levels for classifying gene expression profiles in tumor identification, disease detection, and treatment response. Magen et al. use pairwise combinations of expression dysregulations to predict survival-related gene pairs. These methods, however, do not make use of the prior knowledge available in biological networks.

In this work, we combine the pairwise rank idea with the idea of integrating with biological networks. Pairwise Rank Expressions with Random walks (PRER) is a novel molecular representation method where only pairwise ranks in the known neighborhood of the proteins in the PPI are considered. The proposed model considers the relative expression of a protein within its neighborhood on the PPI network. For a given protein, its neighborhood is defined based on a biased random walk search on the PPI network. PRER also allows interpretability. The pairwise relationships of interacting neighborhood molecules offer a direct interpretation of molecular dysregulation patterns in the context of a known biological network. We also present methods to analyze which pairs become predictive due to their relations instead of their expression levels.

We use PRER for survival prediction in different cancer types. Survival prediction is conducted with PRER features calculated on protein expressions and input a random forest survival model. PRER yields a significant improvement in 8 of the 10 cancer types when compared to the representation of patients with their protein expression features. Additionally, PRER unveils predictive features concerning the known PPIs. In this regard, proteins that are deemed important solely with respect to their interactions are further investigated considering their higher prognostic potential within the known biological interactions.

2 Methods

PRER constructs a vector-based patient representation to be used in subsequent prediction tasks by integrating the patients' molecular expression profiles and the PPI network. The molecular expressions can be the mRNA expressions or protein expressions. Since not all protein expressions are reflected as changes at the protein expression level, in this work, we choose to use protein expression data as input to PRER.

Let $G = (V, E)$ be the given PPI network, where V is the set of vertices representing the proteins, and E is the set of edges that exist between proteins if known to interact. Let $U \subset V$ be the proteins for which protein expression values are available for all patients in the data set. The nodes with the protein expression data, U, constitute the source proteins, and we will denote the number of such proteins with m. Given G, U, and patient expression data over U, the output of PRER for a patient k is a feature vector, $x^{(k)} \in R^s$, that contains the pairwise comparisons encoded with 1 and -1's. Here, s denotes the size of the pairwise comparisons, which will be clarified in the following sections. Below we detail the steps of PRER.

Step 1. Obtaining a Protein's Neighborhood on the Protein Interaction Network: For each source protein in U, we first define a neighborhood, N_u, which is the set of proteins that are proximal to the source protein u on G. To obtain the neighborhood of a node in the graph, a set of random walks is generated. For every source node $u \in U$, we sample neighbors of the source node with a strategy similar to the one in the node2vec [14] algorithm. A random walk with a fixed length of l starting at source node u is generated based on the following distribution:

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & \texttt{if } if(v,x) \quad \epsilon \quad E \\ 0 & \texttt{otherwise} \end{cases} \tag{1}$$

Here, c_i denotes i_{th} node in the walk and $c_0 = u$. $\mathbf{Z}$ is the normalization constant. $P(c_i = x \mid c_{i-1} = v)$ is the transition probability on edge (v, x), where the current node is v, the next node to visit is x, and

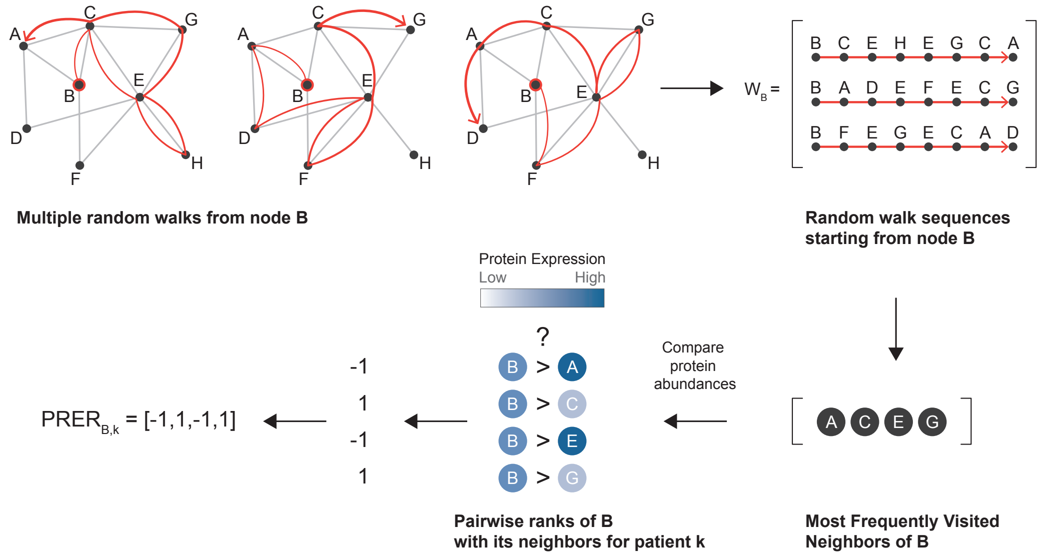

A
B
C
D
E
F
G
H
Multiple random walks from node B
W_B =
B C E H E G C A
B A D E F E C G
B F E G E C A D
Random walk sequences starting from node B
Protein Expression
Low
High
?
B > A
B > C
B > E
B > G
-1
1
-1
1
Compare protein abundances
PRER_B,k = [-1,1,-1,1]
A C E G
Pairwise ranks of B with its neighbors for patient k
Most Frequently Visited Neighbors of B

Fig. 1: Illustration to show how the PRER representation is obtained for a single source node, node B. The nodes in the graph are proteins, edges exist if they interact in the PPI network. First, several random walks are generated that starts at node B as in [14]. These random walks are stored in W_B and used to define the neighborhood of B, N_B. Only the most frequently visited nodes are included in the set of neighbors of B. Then, the pairwise comparison of the neighborhood proteins in terms of their protein expression quantities is used to form a representation of the patient for node B and its neighborhood. The figure shows the features generated for a single protein. This procedure is repeated for all source proteins, and the resulting vectors are concatenated.

the previous node is t. The transition probability depends on the function π, and it is defined as:

$$\pi_{vx} = \alpha_{pq}(t, x) * w_{vx} \tag{2}$$

, where w_{vx} is the edge weight between nodes v and x. However, in this work, we use an unweighted PPI network and, thus, we set $w_{vx} = 1$. $\alpha_{pq}(t, x)$ is the random walk bias which is defined by equation 3 based on the parameters p and q and the shortest path distance between nodes t and x, d_{tx}

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} \text{ if } d_{tx} = 0 \\ 1 \text{ if } d_{tx} = 1 \\ \frac{1}{q} \text{ if } d_{tx} = 2 \end{cases} \tag{3}$$

This bias controls the different search strategies to sample the next visited nodes. We use two different search methods: depth-first sampling (DFS) and breadth-first sampling (BFS), as in [14]. BFS samples the nodes from the nearby nodes, whereas DFS samples the nodes sequentially by gradually increasing the distance from a source node. p and q parameters control the connection between BFS and DFS approaches. With a high q value, sampled nodes in the random walk are aligned to BFS and get a local view over the source node. Small q value aligns random walk to DFS so that a global view of the network is explored. p controls the chance of revisiting the nodes. A high value of p decreases the probability of sampling of the already visited nodes while a small value of p aligns random walk to return the source node.

This biased random walk strategy has two further parameters: (i) walk length l and (ii) the number of random walks r. We select these parameters based on the parameter sensitivity analysis at node2vec [14]. The parameters p and q are used as $p = 0.25, q = 0.25$ in our random walk generation. When $p = 1, q = 1$ uniform random walks are generated without any bias as stated in Grover and Leskovec. A small q value is used to bias the random walks to capture the global view of the network, while a small p value is used to capture the community around the source node u. With the given values, random walks are inclined to see the communities inside the network. By using fixed-length ($l = 100$) random walks, we sample a neighborhood for a given source node, u. To be consistent and to decrease the variance, multiple random walks per source node are applied so that different neighborhoods are sampled for each node. We sampled random walks 18 times and these are stored in W_B (see Figure 1). The frequency of nodes in the multiple neighborhoods are calculated, and the nodes that are involved in more than one random walk are selected as the neighborhod genes.

Step 2. Feature Representation based on Pairwise Rank of Neighborhood Genes: At the end of step one, we arrive at the neighborhood of the protein i, which we denote as N_i. Some neighbors lack measurements, and we define the subset of neighbor proteins with accompanying measurements as $M_i =\in N_i \cap U$. Next, for a protein i, we generate pairwise rank features with every protein $i \in M_i$ as follows.

Let $X_i^{(k)}$ and $X_j^{(k)}$ denote the expression quantities for protein i and j for patient k. Protein i is the source protein, and protein j is a protein in the neighborhood of i. The pairwise rank expression representations (PRER) for this patient is defined as:

$$X_{i,j}^{(k)} = \begin{cases} 1 \texttt{ if } X_i^{(k)} > X_j^{(k)} \\ -1 \texttt{ otherwise} \end{cases} \tag{4}$$

$X_{i,j}^{(k)} = 1$ indicates that the molecule i is more upregulated with respect to molecule j for this patient, whereas $X_{i,j}^{(k)} = -1$ indicates otherwise. For every i in U and for every j in M_i, we define a pairwise rank order for the protein pair. If the protein i's phosphorylated state or states are measured, their comparison with i is also included. This representation constitutes a nonlinear interaction feature mapping among original features that aims to capture expression dysregulations among proteins that are interacting.

2.1 Survival Prediction

Problem Description and the Survival Model: We apply the PRER representation for the survival prediction problem. For each cancer type, the data is of the form, $D = \{\mathbf{x}^{(i)}, \mathbf{S}^{(i)}, \delta^{(i)}\}_{i=1}^{n}$; n is the number of patients. For each patient, $\mathbf{x}$ is the derived features from protein expression data, $\mathbf{S}$ is the overall survival time, and δ denotes censoring. We use random survival forests for the problem. Random Survival Forest(RSF)[15] is a non-parametric method and has been shown to perform well in survival prediction. It is an ensemble method wherein the base learner is a tree, and each tree is grown on a randomly drawn bootstrap sample. Furthermore, in growing a tree, at each node of the tree, a randomly selected subset of features is chosen as the candidate features for splitting. The node is split with the feature among the candidate features that maximizes survival difference between child nodes. We used the default values for the `rfsrc` package [15], where the number of trees is 1000, the number of random splits to consider for each candidate splitting variable is set to 10, and the default splitting rule for a node implements log-rank splitting [16, 17].

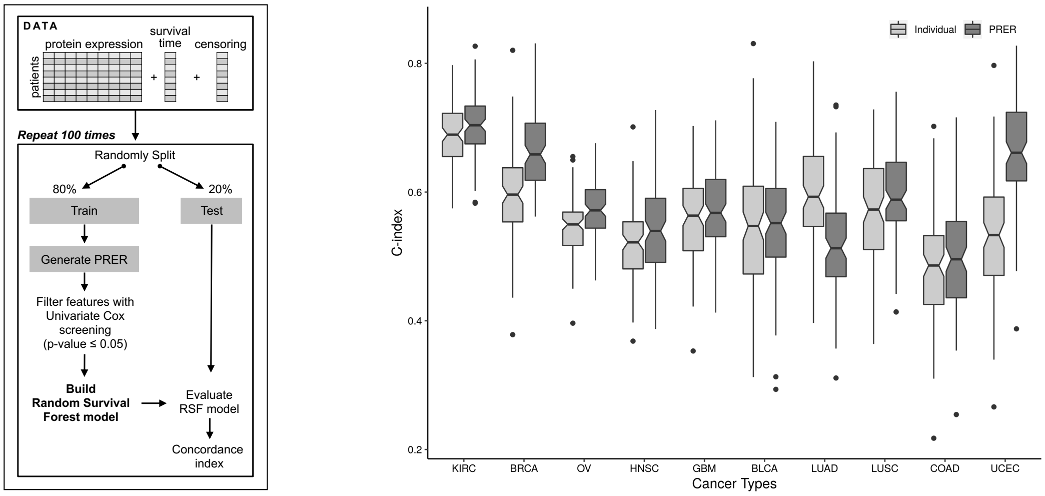
Molecular and Clinical Data: We test the method on ten different cancer types: ovarian adenocarcinoma (OV), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), uterine corpus endometrial carcinoma (UCEC). For each cancer type, the number of patients is given at Supplementary Table 1. We obtained The Cancer Genome Atlas protein expression data and patient survival data from USCS Cancer Browser (`https://genome-cancer.ucsc.edu`) (April 11, 2017). The protein expression is quantified by reverse-phase protein array (RPPA). The features in RPPA data are the expression values of multiple proteins and some phosphorylated versions of proteins. For example, RPPA data include STAT3 and STAT3PY705, where STAT3 is Signal Transducer And Activator Of Transcription 3 protein, and STAT3PY705 is the phosphorylation of STAT3 at tyrosine 705 residue.

Protein-Protein Interaction Network: We obtained the protein-protein interaction (PPI) network from the InBio Map platform (April 11, 2017). InBio Map specifies a confidence score for each edge, representing the support of the interaction in the literature. The interactions that have lower than 0.1 confidence cut-off are eliminated from the network. The final network used in this study includes 17.653 proteins and 625.641 interactions between those proteins.

3 Results and Discussion

To understand if PRER representation captures the molecular expression profiles better than the individual protein expression values, we use these representations for survival prediction task and build two survival prediction models for the 10 cancer types. In these two models, only the feature representations differ. In the first one, we use the protein expression values as input, which is the typical approach taken in survival prediction. In contrast, in the second one, we use the proposed PRER representation.

In all the models trained, we randomly split the samples into train and test groups: 80% as the training set and 20% as the test set. We train 100 such models in 100 test runs. In each of these models, we perform a univariate feature selection based on the hazard ratio of the Cox model [18]. We use the p-values of the likelihood ratio test to quantify the significance of hazard ratio, and features with p-value ≤ 0.05 are retained for model training. Finally, the models are evaluated by the Concordance-Index (C-index) [19] on the test data. The pipeline of the model training and evaluation is summarized in Figure 2a.



(a)

(b)

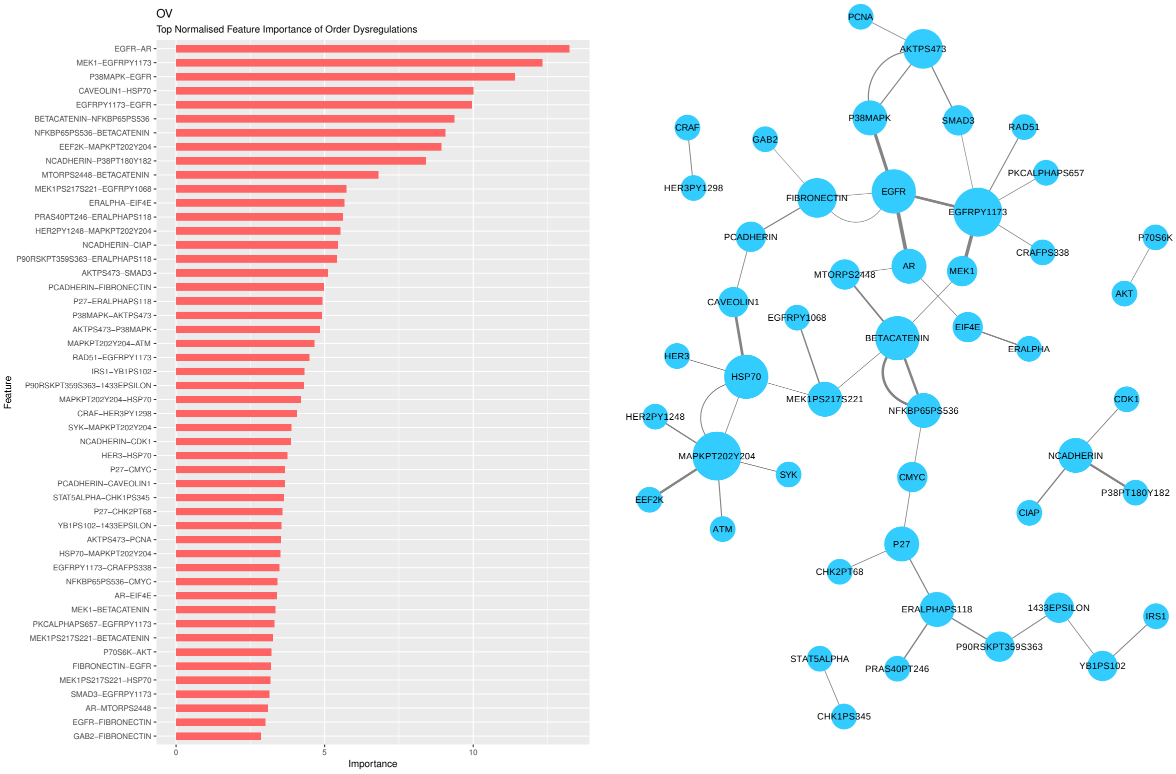
Fig. 2: (a) The pipeline for survival prediction. The step that involves generating PRER is skipped when the experiment is run with the alternative method of individual expression values. **(b)** Comparison of RSF model performances that are trained with individual proteins and pairwise ranking representations for different cancer types. The distribution is over 100 models trained that have different random train and test splits. The performances of the models that use the individual expression values as features (Individual) and PRER representation as features (PRER) are compared in each case.

3.1 Survival Prediction Performance of PRER

Figure 2b compares the distribution of C-indices for 100 models trained with two feature representations for the 10 different cancer types. In 8 of 10 cancer types, PRER representation yields significant improvements (Wilcoxon signed-rank test, (p-value < 0.05)), in one case, the results are promising (BLCA, p-value$= 0.09$) . The C-index quantiles of 100 bootstrap results and corresponding *p-values* are listed in Supplementary Table 2. The best improvements are found in *UCEC*, *BRCA*, *KIRC* and *OV*.

3.2 Predictive PRER Features

We seek to determine the features which are ranked as significant in the RSF models trained with PRER features. Note that in these models, pairs of proteins constitute the features. The importance of a particular feature is quantified by the performance difference between the models trained with the original feature vector and the case where the feature vector values are permuted [20]. A significant difference indicates a feature whose absence degrades the model performance. As there are 100 models trained on the repeatedly split data, we calculate the overall feature importance scores over these models as the sum of the scores. We show the normalized feature importance scores for ovarian cancer (OV) in Figure 3a. The feature importance scores for other cancer types are available in Supplementary Figures 1-9).



(a)

(b)

Fig. 3: (a) The variable importance of significant pairwise ranking representations for ovarian cancer. **(b)** Nodes represent proteins that appear in the top 50 pairwise ranking representations for ovarian cancer; each edges indicate that two proteins participate in a pairwise rank order feature together. For cases where the expression value pertains to the phosphorylated state of the protein, the ids include the phosphosite's residue position and the amino acid type of the phosphosite.

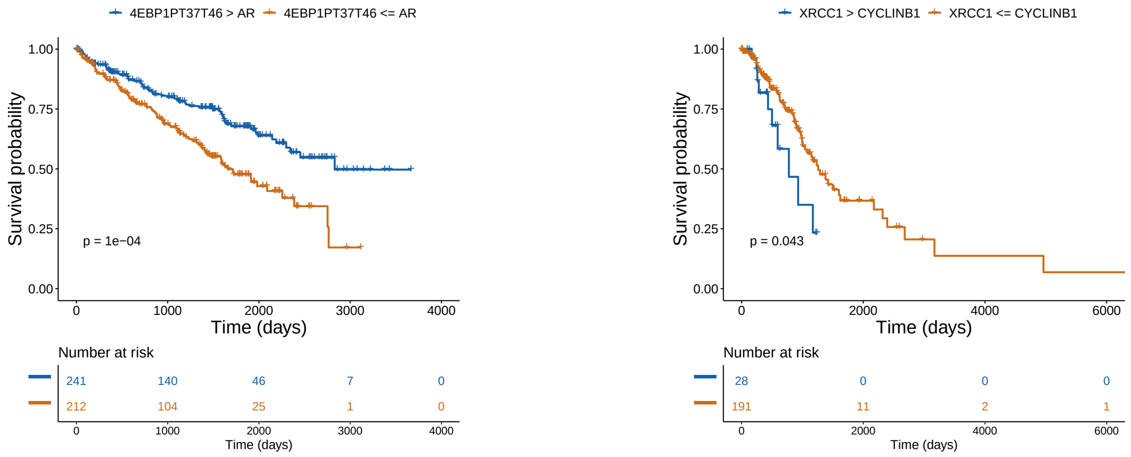
As shown in Figure 3a, some proteins repeatedly show up as partners in the list of important genes. To analyze these relationships, we form a network where the nodes represent proteins that participate in the top 50 PRER features. Edges are formed when a given protein pair is found to be partners in a PRER feature. Figure 3b demonstrates that some proteins emerge as important in many pairs. Several studies support these genes to ovarian cancer. Epidermal growth factor receptor protein (EGFR) and its phosphorylated state EGFRPY1173 are among the top features in PRER representation. EGFR is a receptor protein that receives and transmits signals from the environment to the cell and is the target of drugs in therapies for many cancer types, including ovarian cancer [21, 22]. Marozkina et al. provide results that changes in expression of EGFR may lead to ovarian carcinoma. Others [24, 25, 26] also claim that up-regulation of EGFR expression promotes ovarian cancer. Interestingly, Li et al. and Ilekis et al.

demonstrate that the levels of EGFR and androgen receptor (AR), which constitute the top feature of PRER in Figure 3a, are interacted in ovarian cancer.

Cancer	**Top Rank PRER Protein Pair**
BLCA	NCADHERIN-SRCPY416
BRCA	DVL3-P38MAPK
COAD	MRE11-HER3PY1298
GBM	NF2-EGFR
HNSC	ECADHERIN-PAXILLIN
KIRC	4EBP1T37T46-AR
LUAD	XRCC1-CYCLINB1
LUSC	PAXILLIN-YAP
OV	EGFR-AR
UCEC	EIF4E-AKT

Table 1: The top PRER feature in each cancer type. The relative expression level of this feature is found to be important in the RSF model. The gene symbols of the corresponding gene are listed. The letter P after the gene symbol indicates that this is the phosphorylated version of the protein. The type of the phosphosite and its residue number are provided.

Another important protein that participates in important features is Caveolin-1 (CAV1). CAV1 takes on critical roles in cell survival, cell proliferation, cell migration and programmed cell death [28]. An earlier study by Wiechen et al. report that CAV1 is dysregulated among ovarian cancer patients based on microarray expression data. Others also report that CAV1 is reported to be dysregulated in different cancer types and its role in chemotherapy resistance [30, 31].



(a) KIRC

(b) LUAD

Fig. 4: Kaplan-Meier plots for a) KIRC and b) LUAD based on overall survival. Number at risk denotes the number of patients at risk at a given time, and p-value is calculated with the log-rank test.

We list the top-ranked PRER pairs for each cancer in Table 1. We provide the Kaplan-Meier (KM) plots of the top feature for KIRC and UCEC based on overall survival in Figure 4. Based on only one feature, the patients can be grouped into groups that differ significantly in their survival distributions. We provide the KM plots of top-ranked features for the other cancers in Supplementary Figure 19.

We should note that many of the proteins that are reported in the RPPA assay in the TCGA study are selected due to their relevance to cancer. Thus, these important genes are likely to exhibit the individual importance of PRER partners. Therefore, we suggest an alternative way to exclusively analyze those features which emerge as important in the next section.

3.3 Proteins that Emerge as Important only in the PRER Representation

Since many of the proteins that are in the protein expression data are cancer-related, it is not surprising that they are found to be relevant to cancer. However, proteins that emerge as important in the PRER representation but are not highly ranked in the models trained with individual protein expression values would be interesting. These sets of proteins will reveal proteins whose relative expression states to their neighbors are important as opposed to the expression level being up or down-regulated. To identify these proteins, we first assign a feature importance score to each protein in the PRER representation. As the features are pairs of proteins in the PRER models, we calculate the feature importance of a protein by averaging the importance of the PRER feature importance in which this protein contributes. Let $f_{i,j}$ denotes the feature importance score of the protein pair i and j. We calculate the individual feature importance score for molecule i as follows:

$$s_i = \frac{1}{\|N_i\|} \sum_{j \epsilon N_i} f_{i,j} \tag{5}$$

where N_i is the set of all pairwise ranking representations that include molecule i. s_i represents the average importance of molecule i concerning the expression levels of other proteins in its neighborhood. We get the rank order of each protein based on s_i, and a lower rank indicates that the protein is important. Let r_p be the protein's rank in the models with PRER representation and let r_q be the rank order in the models trained with individual protein expressions. To find the proteins whose ranks are low in the models trained with protein expression but are highly ranked in the PRER models, we measure the differences of feature ranks, $r_q - r_p$. Table 2 lists the top 10 proteins in each cancer based on this $r_q - r_p$ difference. We provide the full list of the ranks and differences in Supplementary Table 3. A large positive difference points to those proteins for which the relative expression relations of this protein to other proteins in its neighborhood carry prognostic value as opposed to its expression value.

BLCA	BRCA	COAD	GBM	HNSC
SRCPY416	YB1PS102	RAD50	EGFR	YAP
YB1	STAT5ALPHA	MRE11	PI3KP110ALPHA	STATHMIN
JNKPT183Y185	CKIT	NF2	P38PT180Y182	SMAD4
YB1PS102	CYCLINB1	MTORPS2448	PDK1PS241	LKB1
RAD51	CHK2PT68	TUBERIN	NFKBP65PS536	NCADHERIN
NCADHERIN	PTEN	NCADHERIN	PRAS40PT246	PDK1PS241
STATHMIN	YAPPS127	MIG6	PTEN	P38MAPK
XRCC1	YB1	JNKPT183Y185	MRE11	P27
NF2	EEF2	PI3KP110ALPHA	ERALPHAPS118	PKCDELTAPS664
TUBERIN	P53	HER3PY1298	NOTCH1	PKCALPHAPS657

KIRC	LUAD	LUSC	OV	UCEC
SMAD1	XRCC1	YAP	EGFR	ASNS
DJ1	YB1	P38PT180Y182	PRAS40PT246	PRAS40PT246
NF2	ASNS	P70S6K	YB1	STATHMIN
KU80	STAT3PY705	LKB1	RAD51	P27PT157
STAT3PY705	YAPPS127	RAD50	PCADHERIN	RAD51
4EBP1PS65	PTEN	XRCC1	HER3	MIG6
GSK3ALPHABETA	YAP	MTOR	PKCALPHAPS657	PCADHERIN
EEF2K	EGFR	SMAD4	SMAD3	P90RSKPT359S363
PR	PEA15	ERALPHAPS118	CIAP	SMAD4
STATHMIN	STATHMIN	BIM	EIF4E	YB1PS102

Table 2: Top-10 rank differentiated features in each cancer with PRER.

We analyze a subset of the proteins in Table 2. The relevance of the relative expressions of proteins for survival is not reported. Some proteins that are known to be cancer drivers and perturbed in cancers such as PTEN or EGFR do not rank high in the model wherein the protein expression data is used as input, but in PRER models, they emerge as important. For example, EGFR is ranked as the 16^{th} most important feature for ovarian cancer in the models trained with PRER, while it is ranked as the least significant one in the models trained with individual expressions only. Similarly, for GBM, EGFR is ranked as the least significant protein in individual expression models, while it is ranked as the 5^{th} most significant feature in PRER. Thus, the PRER models actually highlight that the dysregulation of EGFR expression with respect to its neighbors is an important feature. Below we mention other interesting observations in Table 2.

STAT3PY705 (STAT3 phosphorylation at tyrosine 705), phosphorylated state of STAT3 (Signal Transducer and Activator of Transcription 3) protein, and STAT5ALPHA (Signal Transducer And Activator Of Transcription 5A) also appear in multiple cancer types. While we observe STAT3PY705 as significant in KIRC and LUAD, STAT5ALPHA appears in BRCA in Table 2. Activation in the STAT family is reported, especially for STAT3 and STAT5, in several cancer cell lines including head and neck, breast, kidney, ovarian and colorectal[32, 33, 34, 35].

YAPPS127 and YAP proteins, which are encoded with the YAP1 (Yes-associated protein 1) gene, found important in BRCA, HNSC, LUAD, and LUSC cancer types in Table 2. YAP1 is involved in the Hippo signaling pathway that is associated with the growth, development and repair of the cells, and influences the survival of multiple cancers [36]. Poma et al. reports that 17 genes (out of 32) in the Hippo pathway have effects on survival in more than 20 different cancer types and conclude that YAP1 is relevant to the survival of head and neck carcinoma, hepatocellular, lung adenocarcinoma, gastric, pancreatic and colorectal cancers. Further, other studies also suggest that survival for different cancer types is associated with the expression level of YAP1 and its differential expression is considered as a biomarker for bladder urothelial carcinoma (BLCA) [38], breast invasive carcinoma (BRCA) [39, 40, 41, 42], ovarian serous cystadenocarcinoma (OV) [43, 44].

The upregulation of STATHMIN is linked with poor survival for primary HNSC [45], and Kouzu et al. suggest that it may be used for the prognosis and a therapeutic target for oral squamous-cell carcinoma, which is the most common type of HNSC. Likewise, the upregulation of STATHMIN is significantly correlated with several cancer types such as LUAD [47], gastric cancer [48, 49], UCEC [50], OV [51] and BRCA [52, 53, 54].

YB1 and its phosphorylated state YB1PS102 show correlation with many genes that have functions such as resistance to drugs, transcription and translation of cancerous cells [55]. Although the downregulation of YB1 is found to be correlated with the reduction in progression, development of cell and programmed cell death at various cancer cells such as breast, colon, lung, prostate and pediatric glioblastoma by some studies [56, 57], there are studies [58, 59, 60, 61, 62] showing the association between overexpression of YB1 and different cancer types such as breast, colorectal, glioblastoma, lung, liver, ovarian cancers.

4 Conclusion and Future Work

Accurate prediction of clinical outcomes such as survival success remains to be a challenge for cancer patients. If achieved, it can guide the decision-making process for choosing optimal treatment and surveillance strategies among alternative options. Typically, clinical or pathological features such as the age of the patient, tumor stage, or grade are employed to predict the clinical outcomes. With the advent of high-throughput technologies, molecular descriptions of the tumors for a large number of patients across many cancer types have become available. However, it remains a significant challenge to use this data due to the high level of genomic heterogeneity among patients. In this study, we propose a novel patient representation method, PRER, based on molecular expression patterns on PPI. PRER is based on a pairwise comparison of the expression values of a protein with the other proteins in its neighborhood. In this way, the relative expression level patterns with respect to the proteins in their neighborhood can be captured.

We showcase PRER in the task of survival prediction for ten different cancer types. PRER with Random Survival Forest (RSF) model achieves significant improvements compared to the models with individual expression values in 8 of the 10 cancer types. We also suggest ways to delineate the importance of proteins not through their individual up or down-regulation patterns, but their relative expressions compared to their neighbors. Such an analysis can provide fundamental mechanistic insights into the studied diseases.

One limitation of the current study is that we use a generic protein expression network, disregarding whether the protein is expressed in the given cancer type tissue. We can improve the survival models with tissue-specific PPI networks. Additionally, since we aim to assess the PRER representation power, we only use features related to expression. The survival model can be further improved with clinical features such as age, duration of the follow-up, and cancer stage. PRER representation can be used with other data types, such as mRNA expression. However, we should note that the number of features increases quadratically with the size of the original features. In this case, a more stringent feature filtering step or a regularized prediction model will be helpful.

5 Acknowledgements

Authors thank Bilkent and Sabanci Universities internal funds. OT acknowledges support from Science Academy of Turkey BAGEP program. The results here are in part based upon data generated by the TCGA Research Network.

Bibliography

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

[2] Z. Jagga and D. Gupta, "Classification models for clear cell renal carcinoma stage progression, based on tumor rnaseq expression trained supervised machine learning algorithms," in *BMC proceedings*, vol. 8, no. S6. Springer, 2014, p. S2.

[3] Z. Ding, S. Zu, and J. Gu, "Evaluating the molecule-based prediction of clinical drug responses in cancer," *Bioinformatics*, vol. 32, no. 19, pp. 2891–2895, 2016.

[4] C. Suphavilai, D. Bertrand, and N. Nagarajan, "Predicting cancer drug response using a recommender system," *Bioinformatics*, vol. 34, no. 22, pp. 3907–3914, 2018.

[5] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, p. 551, 2017.

[6] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[7] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, 2007.

[8] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, p. 199, 2009.

[9] A. P. Crijns, R. S. Fehrmann, S. de Jong, F. Gerbens, G. J. Meersma, H. G. Klip, H. Hollema, R. M. Hofstra, G. J. te Meerman, E. G. de Vries *et al.*, "Survival-related profile, pathways, and transcription factors in ovarian cancer," *PLoS medicine*, vol. 6, no. 2, p. e1000024, 2009.

[10] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," *PLoS computational biology*, vol. 9, no. 3, p. e1002975, 2013.

[11] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert *et al.*, "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes," *PLoS computational biology*, vol. 8, no. 5, p. e1002511, 2012.

[12] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mrna comparisons," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–19, 2004.

[13] A. Magen, A. D. Sahu, J. S. Lee, M. Sharmin, A. Lugo, J. S. Gutkind, A. A. Schäffer, E. Ruppin, and S. Hannenhalli, "Beyond synthetic lethality: Charting the landscape of pairwise gene expression states associated with survival in cancer," *Cell reports*, vol. 28, no. 4, pp. 938–948, 2019.

[14] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, vol. 2016, pp. 855–864, 2016.

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

[16] M. R. Segal, "Regression trees for censored data," *Biometrics*, pp. 35–47, 1988.

[17] M. LeBlanc and J. Crowley, "Survival trees by goodness of split," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 457–467, 1993.

[18] T. Therneau, "A package for survival analysis in s. r package version 2.37-4. 2013," 2013.

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

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

[21] L. G. Hudson, R. Zeineldin, M. Silberberg, and M. S. Stack, "Activated epidermal growth factor receptor in ovarian cancer," in *Ovarian Cancer*. Springer, 2009, pp. 203–226.

[22] J. A. Wilken, T. Badri, S. Cross, R. Raji, A. D. Santin, P. Schwartz, A. J. Branscum, A. T. Baron, A. I. Sakhitab, and N. J. Maihle, "Egfr/her-targeted therapeutics in ovarian cancer," *Future medicinal chemistry*, vol. 4, no. 4, pp. 447–469, 2012.

[23] N. V. Marozkina, S. M. Stiefel, H. F. Frierson Jr, and S. J. Parsons, "Mmtv-egf receptor transgene promotes preneoplastic conversion of multiple steroid hormone-responsive tissues," *Journal of cellular biochemistry*, vol. 103, no. 6, pp. 2010–2018, 2008.

[24] I. Dimova, B. Zaharieva, S. Raitcheva, R. Dimitrov, N. Doganov, and D. Toncheva, "Tissue microarray analysis of egfr and erbb2 copy number changes in ovarian tumors," *International Journal of Gynecological Cancer*, vol. 16, no. 1, pp. 145–151, 2006.

[25] J. V. Ilekis, J. P. Connor, G. S. Prins, K. Ferrer, C. Niederberger, and B. Scoccia, "Expression of epidermal growth factor and androgen receptors in ovarian cancer," *Gynecologic oncology*, vol. 66, no. 2, pp. 250–254, 1997.

[26] I. Skirnisdóttir, B. Sorbe, and T. Seidal, "The growth factor receptors her-2/neu and egfr, their relationship, and their effects on the prognosis in early stage (figo i-ii) epithelial ovarian carcinoma," *International Journal of Gynecological Cancer*, vol. 11, no. 2, pp. 119–129, 2001.

[27] A. J. Li, D. R. Scoles, K. U. Armstrong, and B. Y. Karlan, "Androgen receptor cytosine-adenine-guanine repeat polymorphisms modulate egfr signaling in epithelial ovarian carcinomas," *Gynecologic oncology*, vol. 109, no. 2, pp. 220–225, 2008.

[28] C. Boscher and I. R. Nabi, "Caveolin-1: role in cell signaling," in *Caveolins and Caveolae.* Springer, 2012, pp. 29–50.

[29] K. Wiechen, L. Diatchenko, A. Agoulnik, K. M. Scharff, H. Schober, K. Arlt, B. Zhumabayeva, P. D. Siebert, M. Dietel, R. Schäfer *et al.*, "Caveolin-1 is down-regulated in human ovarian carcinoma and acts as a candidate tumor suppressor gene," *The American journal of pathology*, vol. 159, no. 5, pp. 1635–1643, 2001.

[30] L. A. Carver and J. E. Schnitzer, "Caveolae: mining little caves for new cancer targets," *Nature Reviews Cancer*, vol. 3, no. 8, pp. 571–581, 2003.

[31] M. Zhang and S. Luo, "Gene expression profiling of epithelial ovarian cancer reveals key genes and pathways associated with chemotherapy resistance," *Genet Mol Res*, vol. 15, no. 1, p. 11, 2016.

[32] R. Buettner, L. B. Mora, and R. Jove, "Activated stat signaling in human tumors provides novel molecular targets for therapeutic intervention," *Clinical cancer research*, vol. 8, no. 4, pp. 945–954, 2002.

[33] H. Yu and R. Jove, "The stats of cancernew molecular targets come of age," *Nature Reviews Cancer*, vol. 4, no. 2, p. 97, 2004.

[34] A. Lavecchia, C. Di Giovanni, and E. Novellino, "Stat-3 inhibitors: state of the art and new horizons for cancer treatment," *Current medicinal chemistry*, vol. 18, no. 16, pp. 2359–2375, 2011.

[35] I. Souissi, I. Najjar, L. Ah-Koon, P. O. Schischmanoff, D. Lesage, S. Le Coquil, C. Roger, I. Dusanter-Fourt, N. Varin-Blank, A. Cao *et al.*, "A stat3-decoy oligonucleotide induces cell death in a human colorectal carcinoma cell line by blocking nuclear transfer of stat3 and stat3-bound nf-κb," *BMC cell biology*, vol. 12, no. 1, p. 14, 2011.

[36] E. Lorenzetto, M. Brenca, M. Boeri, C. Verri, E. Piccinin, P. Gasparini, F. Facchinetti, S. Rossi, G. Salvatore, M. Massimino *et al.*, "Yap1 acts as oncogenic target of 11q22 amplification in multiple cancer subtypes," *Oncotarget*, vol. 5, no. 9, p. 2608, 2014.

[37] A. M. Poma, L. Torregrossa, R. Bruno, F. Basolo, and G. Fontanini, "Hippo pathway affects survival of cancer patients: extensive analysis of tcga data and review of literature," *Scientific reports*, vol. 8, no. 1, p. 10623, 2018.

[38] J.-Y. Liu, Y.-H. Li, H.-X. Lin, Y.-J. Liao, S.-J. Mai, Z.-W. Liu, Z.-L. Zhang, L.-J. Jiang, J.-X. Zhang, H.-F. Kung *et al.*, "Overexpression of yap 1 contributes to progressive features and poor prognosis of human urothelial carcinoma of the bladder," *BMC cancer*, vol. 13, no. 1, p. 349, 2013.

[39] F. Cheng, J. Zhao, A. B. Hanker, M. R. Brewer, C. L. Arteaga, and Z. Zhao, "Transcriptome-and proteome-oriented identification of dysregulated eif4g, stat3, and hippo pathways altered by pik3ca h1047r in her2/er-positive breast cancer," *Breast cancer research and treatment*, vol. 160, no. 3, pp. 457–474, 2016.

[40] L. Cao, P.-L. Sun, M. Yao, M. Jia, and H. Gao, "Expression of yes-associated protein (yap) and its clinical significance in breast cancer tissues," *Human pathology*, vol. 68, pp. 166–174, 2017.

[41] S. K. Kim, W. H. Jung, and J. S. Koo, "Yes-associated protein (yap) is differentially expressed in tumor and stroma according to the molecular subtype of breast cancer," *International journal of clinical and experimental pathology*, vol. 7, no. 6, p. 3224, 2014.

[42] H. M. Kim, W. H. Jung, and J. S. Koo, "Expression of yes-associated protein (yap) in metastatic breast cancer," *International journal of clinical and experimental pathology*, vol. 8, no. 9, p. 11248, 2015.

[43] C. He, X. Lv, G. Hua, S. M. Lele, S. Remmenga, J. Dong, J. S. Davis, and C. Wang, "Yap forms autocrine loops with the erbb pathway to regulate ovarian cancer initiation and progression," *Oncogene*, vol. 34, no. 50, p. 6040, 2015.

[44] Y. Xia, T. Chang, Y. Wang, Y. Liu, W. Li, M. Li, and H.-Y. Fan, "Yap promotes ovarian cancer cell tumorigenesis and is indicative of a poor prognosis for ovarian cancer patients," *PloS one*, vol. 9, no. 3, p. e91770, 2014.

[45] H. Wu, W.-W. Deng, L.-L. Yang, W.-F. Zhang, and Z.-J. Sun, "Expression and phosphorylation of stathmin 1 indicate poor survival in head and neck squamous cell carcinoma and associate with immune suppression," *Biomarkers in medicine*, vol. 12, no. 7, pp. 759–769, 2018.

[46] Y. Kouzu, K. Uzawa, H. Koike, K. Saito, D. Nakashima, M. Higo, Y. Endo, A. Kasamatsu, M. Shiiba, H. Bukawa *et al.*, "Overexpression of stathmin in oral squamous-cell carcinoma: correlation with tumour progression and poor prognosis," *British journal of cancer*, vol. 94, no. 5, pp. 717–723, 2006.

[47] L. Yurong, R. Biaoxue, L. Wei, M. Zongjuan, S. Hongyang, F. Ping, G. Wenlong, Y. Shuanying, and L. Zongfang, "Stathmin overexpression is associated with growth, invasion and metastasis of lung adenocarcinoma," *Oncotarget*, vol. 8, no. 16, p. 26000, 2017.

[48] T. Jeon, M. Han, Y. Lee, Y. Lee, G. Kim, G. Song, G. Hur, J. Kim, H. Kim, S. Yoon *et al.*, "Overexpression of stathmin1 in the diffuse type of gastric cancer and its roles in proliferation and migration of gastric cancer cells," *British journal of cancer*, vol. 102, no. 4, pp. 710–718, 2010.

[49] X. Liu, H. Liu, J. Liang, B. Yin, J. Xiao, J. Li, D. Feng, and Y. Li, "Stathmin is a potential molecular marker and target for the treatment of gastric cancer," *International journal of clinical and experimental medicine*, vol. 8, no. 4, p. 6502, 2015.

[50] W. Xi, W. Rui, L. Fang, D. Ke, G. Ping, and Z. Hui-Zhong, "Expression of stathmin/op18 as a significant prognostic factor for cervical carcinoma patients," *Journal of cancer research and clinical oncology*, vol. 135, no. 6, pp. 837–846, 2009.

[51] D. Su, S. M. Smith, M. Preti, P. Schwartz, T. J. Rutherford, G. Menato, S. Danese, S. Ma, H. Yu, and D. Katsaros, "Stathmin and tubulin expression and survival of ovarian cancer patients receiving platinum treatment with and without paclitaxel," *Cancer*, vol. 115, no. 11, pp. 2453–2463, 2009.

[52] L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmström, L. Memeo *et al.*, "Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7564–7569, 2007.

[53] G. Brattsand, "Correlation of oncoprotein 18/stathmin expression in human breast cancer with established prognostic factors," *British journal of cancer*, vol. 83, no. 3, pp. 311–318, 2000.

[54] R. Golouh, T. Cufer, A. Sadikov, P. Nussdorfer, P. A. Usher, N. Brünner, M. Schmitt, R. Lesche, S. Maier, M. Timmermans *et al.*, "The prognostic value of stathmin-1, s100a2, and syk proteins in er-positive primary breast cancer patients treated with adjuvant tamoxifen monotherapy: an immunohistochemical study," *Breast cancer research and treatment*, vol. 110, no. 2, pp. 317–326, 2008.

[55] Y. Basaki, K.-i. Taguchi, H. Izumi, Y. Murakami, T. Kubo, F. Hosoi, K. Watari, K. Nakano, H. Kawaguchi, S. Ohno *et al.*, "Y-box binding protein-1 (yb-1) promotes cell cycle progression through cdc6-dependent pathway in human cancer cells," *European journal of cancer*, vol. 46, no. 5, pp. 954–965, 2010.

[56] Y. Basaki, F. Hosoi, Y. Oda, A. Fotovati, Y. Maruyama, S. Oie, M. Ono, H. Izumi, K. Kohno, K. Sakai *et al.*, "Akt-dependent nuclear localization of y-box-binding protein 1 in acquisition of malignant characteristics by human ovarian cancer cells," *Oncogene*, vol. 26, no. 19, p. 2736, 2007.

[57] A. Lasham, W. Samuel, H. Cao, R. Patel, R. Mehta, J. L. Stern, G. Reid, A. G. Woolley, L. D. Miller, M. A. Black *et al.*, "Yb-1, the e2f pathway, and regulation of tumor cell growth," *Journal of the National Cancer Institute*, vol. 104, no. 2, pp. 133–146, 2011.

[58] R. C. Bargou, K. Jürchott, C. Wagener, S. Bergmann, S. Metzner, K. Bommert, M. Y. Mapara, K.-J. Winzer, M. Dietel, B. Dörken *et al.*, "Nuclear localization and increased levels of transcription factor yb-1 in primary human breast cancers are associated with intrinsic mdr1 gene expression," *Nature medicine*, vol. 3, no. 4, p. 447, 1997.

[59] T. Kamura, H. Yahata, S. Amada, S. Ogawa, T. Sonoda, H. Kobayashi, M. Mitsumoto, K. Kohno, M. Kuwano, and H. Nakano, "Is nuclear expression of y box-binding protein-1 a new prognostic factor in ovarian serous adenocarcinoma?" *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 85, no. 11, pp. 2450–2454, 1999.

[60] K. Shibao, H. Takano, Y. Nakayama, K. Okazaki, N. Nagata, H. Izumi, T. Uchiumi, M. Kuwano, K. Kohno, and H. Itoh, "Enhanced coexpression of yb-1 and dna topoisomerase ii α genes in human colorectal carcinomas," *International journal of cancer*, vol. 83, no. 6, pp. 732–737, 1999.

[61] K. Shibahara, K. Sugio, T. Osaki, T. Uchiumi, Y. Maehara, K. Kohno, K. Yasumoto, K. Sugimachi, and M. Kuwano, "Nuclear expression of the y-box binding protein, yb-1, as a novel marker of disease progression in non-small cell lung cancer," *Clinical cancer research*, vol. 7, no. 10, pp. 3151–3155, 2001.

[62] M. Yasen, K. Kajino, S. Kano, H. Tobita, J. Yamamoto, T. Uchiumi, S. Kon, M. Maeda, G. Obulhasim, S. Arii *et al.*, "The up-regulation of y-box binding proteins (dna binding protein a and y-box binding protein-1) as prognostic markers of hepatocellular carcinoma," *Clinical cancer research*, vol. 11, no. 20, pp. 7354–7361, 2005.