IET Systems Biology

Research Article

IET Journals
The Institution of Engineering and Technology

Biclustering-based association rule mining approach for predicting cancer-associated protein interactions

ISSN 1751-8849
Received on 19th March 2019
Revised 3rd June 2019
Accepted on 18th June 2019
E-First on 2nd August 2019
doi: 10.1049/iet-syb.2019.0045
www.ietdl.org

Lopamudra Dey[1] ✉, *Anirban Mukhopadhyay*[2]

[1]*Department of Computer Science and Engineering, Heritage Institute of Technology, 994 Madurdaha, Kolkata 700 107, West Bengal, India*
[2]*Department of Computer Science and Engineering, University of Kalyani, Nadia, Kalyani 741235, West Bengal, India*
✉ *E-mail: lopamudra.dey1@gmail.com*

Abstract: Protein–protein interactions (PPIs) have been widely used to understand different biological processes and cellular functions associated with several diseases like cancer. Although some cancer-related protein interaction databases are available, lack of experimental data and conflicting PPI data among different available databases have slowed down the cancer research. Therefore, in this study, the authors have focused on various proteins that are directly related to different types of cancer disease. They have prepared a PPI database between cancer-associated proteins with the rest of the human proteins. They have also incorporated the annotation type and direction of each interaction. Subsequently, a biclustering-based association rule mining algorithm is applied to predict new interactions with type and direction. This study shows the prediction power of association rule mining algorithm over the traditional classifier model without choosing a negative data set. The time complexity of the biclustering-based association rule mining is also analysed and compared to traditional association rule mining. The authors are able to discover 38 new PPIs which are not present in the cancer database. The biological relevance of these newly predicted interactions is analysed by published literature. Recognition of such interactions may accelerate a way of developing new drugs to prevent different cancer-related diseases.

1 Introduction

Protein–protein interactions (PPIs) play a key role in different physiological processes and signal pathways in the human body. Most of the biological activities in a cell are controlled by proteins. The proteins are also important for the functional and structural development of different organs of the body [1, 2]. It is observed that some of the proteins perform their operations independently but a vast number of proteins interact with each other to complete a proper biological function.

Cancer, one of the deadliest diseases of mankind, is nothing but the abnormal growth of the cells. Genes inside a cell, raise the instructions to protein so that the protein can execute the appropriate function for the cell [3]. The Cancer Genome Atlas (http://cancergenome.nih.gov) [4] and The International Cancer Genome Consortium (http://icgc.org/ icgc/cgp) [5] have discovered that cancer disease alters the genome characteristics. This alteration changes the proteins' features that regulate cell growth. As a result, cells increase maniacally and turn into cancerous. So, identification of these proteins and its interaction with other proteins can give insight knowledge of cancer disease and its pathway. In [6], the authors have shown that alternation of protein interaction network and abnormal cell growth led to tumorigenesis and some other disease progression.

Recently, cancer-related research work has gained a lot of interest. As a result, target PPI for anticancer strategy has also increased. However, those investigations reveal only a small part of the total possible PPIs. The total number of potential PPI within one cell is large. So, verification of each and every interaction experimentally is not feasible. This leads to developing computational methods to predict large PPIs and validate them by further biological experiments [7]. A structure-based approach is applied in the paper [8] for acquiring cancer-related hub proteins. Here the genes are classified according to their phenotypes. The paper proves that the interfaces of the cancer-related proteins are different from non-cancer interfaces within a phenotype. In [9], the authors have prepared a cancer PPI network, termed as OncoPPi network using time-resolved Förster resonance energy transfer technology. They have discovered around 260 new high confidence PPIs, mostly for lung cancer, which is not identified previously. A system biology-based method is applied in [10] to predict cancer PPI, especially lung, breast, and ovarian cancer. In [11], the authors have done a global comparative analysis on different cancer data sets of bladder, colon, kidney, and thyroid cancers. They have identified that a set of molecular functions (GO-MF) and biological processes (GO-BP) are similar in all these cancers with the help of the GEO database (Gene Expression Omnibus). A cancer PPI database based on the pathway was constructed in [12]. The authors investigated that the cancer-related proteins' alterations tend to cluster in modules and these modules are closely linked to particular biological pathways. In [13], the authors examined the dynamic structure of the human protein interaction network and analysed the intermolecular and intramolecular hub proteins. In intermolecular hub proteins, the interacting proteins are co-expressed in a tissue-restricted manner. On the other hand, the proteins in the intramolecular set are co-expressed with their interacting proteins in all or most tissues. They have applied this method for breast cancer diagnosis in [13]. Note that utilisation of these data sets is limited due to lack of experimental data and conflicting PPI data among different available databases. In addition, none of these cancer-focused literature studies have considered the annotation type and direction of interactions.

Motivated by this, a database consists of PPI between different cancer disease-related proteins and other human proteins with annotation type and direction have been prepared. All the interactions are collected from STRING www.string-db.com [14, 15], and cross-validated by BioGRID version 3 https://thebiogrid.org [16] and DIP https://dip.doe-mbi.ucla.edu database [17]. Then biclustering-based association rule mining (ARM) algorithm is applied to this data set to predict new interactions. ARM has been applied to many studies for prediction including different PPI databases [18–20]. However, in this article, we are using this algorithm to predict new interactions of cancer-associated proteins which is not done previously as per our knowledge.

The traditional classifiers require both positive and negative data set to design the training model. However, for any PPI database, experimentally validated, i.e. positive data set is

available. However, there is no 'gold standard' of preparing negative data set is addressable. Therefore, in most of the studies, negative data sets are generated by choosing random samples based on the hypothesis that random pairs are less expected to interact physically [21, 22]. As a consequence, the performance of the classifier extremely depends on the selection of the negative samples. The ARM algorithm alleviates the need for choosing negative samples (non-interacting pairs of proteins) which is needed for building a classifier and can predict PPI from the information of experimentally validated PPIs only. In addition, traditional classification-based approaches cannot predict the annotation type and direction of the interactions. In the proposed approach, we are able to predict the type and direction of the interactions, and thus it is easier to validate them through experiments. Classical Apriori algorithm for ARM algorithm generates frequent itemsets first and then extracts rules from these frequent itemsets. As a result, the time complexity of the algorithm is usually high for large data sets. Moreover, as any subset of a frequent itemset is also frequent, the algorithm requires huge memory to store redundant frequent itemsets. Therefore, the biclustering algorithm is used to improve the space and time complexity of ARM. We have applied the biclustering algorithm to generate the frequent closed itemsets (FCIs) from the PPI matrix and then the rules are extracted from the FCIs. Note that the number of FCIs is much less than the number of frequent itemsets, and they contain non-redundant information. The time complexity comparison of these algorithms is also included in a separate section.

The proposed biclustering-based ARM algorithm generates a set of new rules and using these rules, we are not only predicting new PPI interactions, but also the type and direction of the interactions. As the number of proteins having common interaction type and direction is very less, we have discovered only 38 new interactions which are not present in published databases. The biological relevance of these interactions is studied in various works of literature. The study will give helpful data to establish new relationships among cancer-related proteins and accelerate the discovery of new medications.

Table 1 Different cancer diseases and the proteins responsible for the particular types of cancer

Various types of cancer	Responsible proteins
acute myeloid leukemia	RUNX1, HIST3H3, KRAS, NPM1, CBFA2T2
adenocarcinoma	S100A4
B-cell lymphoma	BCL6
bladder cancer	AMFR
breast cancer	BRCA1, CXCR1, DMTF1, DEFB106, LMO4
colorectal cancer	ANKS1A, S100A4, KRAS
esophageal cancer	SPINK7,PICK1
fibrosarcoma	ETV6
insulinoma	INSM1
laryngeal cancer	BAG1
leukemia	MLLT3,CRKL,KAT6A
ewing sarcoma	EWSR1
liver cancer	DLC2
lung cancer	KEAP1
lymphoma	BCL2L1,CSF3
melanoma	MIA,S100B
nephroblastoma/Wilms' tumor	BASP1
ovarian cancer	BRCA1,DAB2
pancreatic cancer	AGR2
prostate cancer	MMP23B
renal cancer	STIP1
rhabdomyosarcoma	TTN
epithethial cancer	MUC1

Table 2 Five fields of protein interaction database associated with cancer disease

Protein 1	Protein 2	Annotation type	Weight	Direction
CSF3	SELL	activation	0.8	1
CSF3	SELL	expression	0.8	1
CSF3	CXCR4	activation	0.8	1
ERBB2	CXCR4	activation	0.8	1
CXCL12	CXCR4	expression	0.8	1
BAG1	HSPA8	binding	0.999	0
KEAP1	CUL3	binding	0.937	0

The first two columns represent the names of two interacting proteins, the third column represents the type of interaction, the fourth column represents the weight or confidence of the interaction (between 0 and 1), and the last column represents the direction of interaction (1 for first protein to second protein, −1 for second protein to first protein, and 0 for both direction)

2 Methods

2.1 Disease-Protein mapping database

Researchers have officially found proteins that are responsible for various cancer diseases. The information is gathered from http://www.bmrb.wisc.edu/data_library/Diseases/. We have placed our findings in Table 1 (File D1). Protein plays key roles in different cellular processes like to regulate all processes, cascading signals, accelerate some chemical reactions etc. [23]. So, identifying those proteins as well as the interaction between these proteins with the rest of the proteins of the human body is very important for the comprehension of the functionality of a living cell. This will help the researchers to find the specific pathways of other diseases that may result due to PPI.

2.2 PPI database with annotation type and direction

All the human proteins that interact with these cancer-related proteins, mentioned in Table 1, are extracted from STRING [14, 15] database. The STRING database contains more complete annotation type and direction information compared to other databases like Biogrid and DIP. As our main focus of the paper is finding annotation types and directions of predicted interactions, we have prepared the cancer PPI data from the STRING database. To construct a high throughput cancer PPI database, a high threshold value (0.7) is considered for extracting PPI from STRING (File D2). All these interactions are cross-validated by BioGRID version 3 [16] and DIP database [17] by eliminating the interactions which are not present in BioGRID or DIP.

We retrieved a total of 680 cancer-focused PPI between cancer-related human proteins and other human proteins. A snapshot of the database is shown in Table 2. The direction field plays a very important role in prediction. Basically, a cell responds to stimuli through different signaling pathways. Then the proteins in a cell interact with each other to transmit those signals. Now, the direction of interaction represents the direction of signal flow. So, it gives the information which protein activates/reacts/expresses/catalyses or post-translates which protein. Three types of direction (0, 1, and − 1) of the interactions based on a directed graph are considered here. These three values (0, 1, and − 1) indicate that the relation between the two proteins is either uni-directional or bi-directional [24]. If a signal flows in only one direction, for example, from protein 1 to protein 2, then the direction will be uni-directional, i.e. 1 or −1. However, if the signal can flow in both directions, then the direction will be bi-directional, i.e. 0. From Table 2, it can be seen that CSF3 interacts with SELL. The interaction type is activation and direction is 1. So, from this, we can conclude that CSF3 activates SELL. In the same way, BAG1 and HSPA8 bind with each other as the direction of interaction is 0.

There are six unique interaction types found in the cancer data set. They are binding, expression, catalysis, reaction, activation, and post-translation. The frequency of each interaction is shown in Fig. 1 in descending order. It can be noted from the figure that most of the interactions have annotation type binding (304), whereas, only 25 PPIs have annotation type activation.

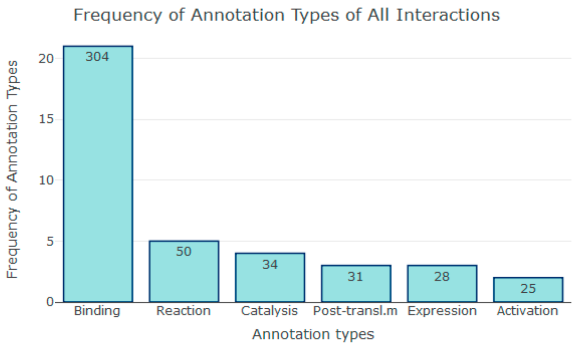

Fig. 1 *Distribution of the frequency of interaction types in the whole Cancer Data set. It can be seen from the figure that most of the interactions have annotation type binding (304), whereas, only 25 PPIs have annotation type activation*

2.3 Association rule mining and bimax biclustering algorithm

In data science, ARM is utilised to discover interesting relevant relationships covered up in extensive information sets. We can represent all unexposed interactions in the form of association rules [25]. A mathematical model can be used to address the problem of mining association rules. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literal, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Let X be a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. An association rule can be expressed by the form $X => Y$, $X \subset T$, $Y \subset I$, and $X \cap Y = \emptyset$. There are two parameters are associated with ARM-confidence, and support. Confidence denotes the strength of implication and support indicates the frequencies of the occurring patterns in the rule. Rules that satisfy both user-defined minimum support (*minsup*) and minimum confidence (*minconf*) are referred to as strong rules. The main task of the algorithm is to identify strong association rules in large databases. This can be done by producing FCIs. If support of an itemset is greater than or equal to some threshold (*minsup*), then the itemset is called frequent itemset. An itemset A is called closed if there exists no proper super-itemset B such that B has the same support value as A. Now an itemset is called closed frequent itemset if it satisfies the criteria of both frequent and closed itemset.

However, the generation of FCI from ARM is a two-step process. First, from all defined itemsets, common sets of items are gathered that have at least a minimum support(*minsup*). Then, high confidence rules are generated from each frequent itemset. As any subset of a frequent itemset is also frequent, frequent itemset mining using ARM is computationally expensive. It incurs a huge memory overhead [26]. To minimise the complexity of ARM, biclustering technique is used to directly mine the closed frequent itemsets reducing the time and space required for mining all frequent itemsets.

Biclustering is a special clustering technique which is very useful for synchronous grouping of the rows and columns of a matrix [20]. This is different from the normal clustering approach. Clustering technique can be applied to either rows or columns of matrix separately but biclustering performs clustering in both dimensions simultaneously. This will generate sub-matrices having a unique pattern as clusters. Each of these submatrices is called a bicluster. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of rows and $Q = \{q_1, q_2, \ldots, q_m\}$ be a set of columns of a data matrix $B = [b_{ij}]$, where $1 \leq i \leq n$ and $1 \leq j \leq m$. A submatrix with a subset I of rows and subset J of columns is called a bicluster if all the values of rows and columns follow consistent patterns where $I \subseteq P$ and $J \subseteq Q$. There are many biclustering algorithms in the literature. Here we have used Bimax (Binary inclusion-Maximal) algorithm to perform biclustering in our data set. From a binary matrix, the Bimax algorithm generates all biclusters that contain 1's only. This is achieved by applying the divide-and-conquer strategy recursively on binary data. All these generated biclusters are maximal biclusters in nature [27]. A maximal bicluster is one which is not included within another bicluster. Thus the set of columns of a maximal bicluster corresponds to a closed itemset because the addition of any column with this set will only reduce the support of the column set. Hence a maximal bicluster actually represents a closed frequent itemset.

2.4 Proposed methodology

The proposed methodology in the paper is comprised of two phases: (i) finding all maximal biclusters and maximal FCIs using Bimax algorithm, (ii) extract the important rules using Apriori ARM approach.

The sets of data items which appear together in most of the occurrences are called frequent itemset. Various study show that only a few proteins act alone. Most of the proteins interact with each other to execute a proper biological function. So, PPI can be viewed as a frequent itemset problem. Frequently closed itemset concept comes from economic market basket analysis. It tracks the frequent combinations and associations of items purchased together with the objective of understanding the activity of retail customers [28]. In this study, the Bimax algorithm is used to extract the maximal frequent itemsets from the cancer PPI database. Note that the Bimax algorithm operates only on binary data [29]. Therefore, we have first converted the cancer PPI database into an adjacency matrix (Table 3), where the rows specify the unique proteins and the columns correspond to the interacting proteins with annotation types and directions of the interactions written in Interacting ProteinName_AnnotationType_Direction format. Therefore, the values of the table become binary. An entry '1' in the matrix refers to the presence of the connection between the corresponding protein pairs with that particular annotation and direction, and an entry of '0' denotes the absence of any information between the association of the comparing proteins. This adjacency matrix helps to find out the FCIs so that the new interactions can be predicted. As there are a very small number of proteins which share common interacting protein with the same annotation type and direction, the adjacency matrix is sparse in nature. As an illustration, a sample of the adjacency matrix corresponding to Table 2 data has been shown in Table 3.

Bimax algorithm is applied to this preprocessed data using Biclustering Analysis Toolbox (BicAT), [30], openly accessible from http://www.tik.ethz.ch/sop/bicat/ to identify all maximal biclusters. The columns for each maximal bicluster correspond to a closed frequent itemset. The whole procedure can be described by following algorithm 1 (Fig. 2). The codes and flowchart are given in Appendix (File D8).

2.5 Predictive performance comparison

The adjacency matrix prepared in the previous step is sparse in nature. However, PPI prediction for the sparse network is a challenging task. In many studies, adjacency matrix factorisation (AMF) has been used for link prediction for sparse data set [31]. In [32], Yokoi *et al.* implemented the link prediction method using incidence matrix factorisation (IMF) and showed that this algorithm performed better compared to adjacency matrix factorisation when the network became sparser. However, the functioning of IMF deteriorates for complex real-world networks having high clustering coefficient (>0.01). In a graph, the clustering coefficient is defined as the ratio of the number of edges

Table 3 Adjacency matrix corresponding to Table 2 data, where rows are cancer proteins and columns are interacting proteins with annotation type and direction of interaction written in ProteinName_AnnotationType_Direction format

Protein_protein	SELL_Activation_1	SELL_Expression_1	CXCR4_Activation_1
CSF3	1	1	1
ERBB2	0	0	1
CXCL12	0	0	0
BAG1	0	0	0
KEAP1	0	0	0

It can be seen that the matrix discussed here is sparse matrix. Most of the values of the matrix are zero. Only a few proteins interact with each other with common annotation type and direction

Protein_protein	HSPA8_Binding_0	CUL3_Binding_0	CXCR4_Expression_1
CSF3	0	0	0
ERBB2	0	0	0
CXCL12	0	0	1
BAG1	1	0	0
KEAP1	0	1	0

Require: A set of new rules based on association rule mining with $minsup$ and $minconf$

Ensure: Protein - protein interaction database with annotation and direction

Step 1: Prepare the Adjacency matrix (A_{ij}) from the PPI dataset, where i= unique proteins and j= interacting proteins with annotation and direction.

Step 2: Apply Bimax to A_{ij} and store all the frequent closed itemsets.

Step 3: Apply Apriori algorithm with $minsup$ and $minconf$ to generate all possible rules from frequent closed itemsets.

Step 4: Remove the redundant and confidence-1 rules and assemble a set of useful rules required for predicting new interactions.

Fig. 2 *Algorithm 1: algorithm for extracting association rules*

Table 4 Performance of the IMF, AMF, and the proposed algorithm to predict artificially removed links

% of Removed links	Performance metric of IMF in %	Performance metric of AMF in %	Performance metric of Bimax based ARM in %
10%	72	**74**	72
20%	62	64	**82**
30%	**78**	74	70
40%	70	**82**	78
50%	66	78	**84**
average	69.6	74.4	**77.2**

For each row, the best result is shown in bold

among the neighbours of nodes to the maximum number of edges that could potentially exist between the nodes. We have noted that the clustering coefficient (as defined in Section 4) of our cancer PPI data set is 0.291, which signifies the scale-free property of the network. IMF is not expected to give good results for networks having such high clustering coefficient value. Before applying the proposed method to predict PPI, we have compared the algorithm with IMF and AMF. We have divided the adjacency matrix prepared in Section 2.4, into training (70%) and testing (30%) data set. Thereafter, randomly some 1's, i.e. edges are eliminated from testing data set. Then IMF, AMF, and the proposed method are trained using training data set and applied to test data set to examine how many eliminated 1's, i.e. how many artificially removed links can be predicted. We have repeated the procedure for five times by eliminating 10, 20, 30, 40, and 50% edges from the testing matrix. Five-fold cross-validation scheme is employed and the performance metric of each case and their average is reported in Table 4. The performance metric (M) can be calculated as

$$M = \frac{\text{The number of eliminated edges detected}}{\text{The total number of eliminated edges}} \times 100\% . \quad (1)$$

The algorithm of above procedure is discussed in Fig. 3. It can be seen from the table that Biclustering-based ARM can detect more removed links (77.2%) compared to the IMF (69.6%) and AMF (74.4%). Therefore, we have predicted a set of unknown cancer PPI using Bimax-based ARM technique and these interactions can be further validated by experimental results.

3 Predicting new interactions

Bimax algorithm calculates a total of 31 FCIs. Using Apriori algorithm of ARM, 294 rules describing the association among the proteins are generated from these FCIs considering minimum support (*minsup*) 0.03 and minimum confidence (*minconf*) 0.5. The adjacency matrix that we have formed is a sparse matrix. So, the probability of the same protein_annotationtype_direction appearance in the rule sets is very less. Also, the minimum support value should be low enough to obtain sufficient numbers of rules. Considering min_support value 0.01, only 11 rules are produced, which is very small compared to our database size. After several trials, we have considered the minimum support (*minsup*) is 0.03 and minimum confidence (*minconf*) is 0.5 so that a reasonable number of unique high throughput interactions can be predicted. This criterion generates 294 rules. The procedure of generating rules from FCIs can be explained by the following example.

Consider a FCI consisting of annotated human proteins as follows: $P1\_a1\_d1$, $P2\_a2\_d2$, $P3\_a3\_d3$, and $P4\_a4\_d4$, where P_i denotes interacting proteins, each a_i denotes the interaction type and d_i denotes direction tagged with each of these proteins. A list of some possible rules constructed from those proteins is as follows:

(i) $P1\_a1\_d1, P2\_a2\_d2, P3\_a3\_d3 \Rightarrow P4\_a4\_d4$
(ii) $P1\_a1\_d1, P2\_a2\_d2 \Rightarrow P3\_a3\_d3, P4\_a4\_d4$
(iii) $P1\_a1\_d1, P2\_a2\_d2, P4\_a4\_d4 \Rightarrow P3\_a3\_d3$ etc.

A two-step filtering process is applied to these rules. First, redundant rules are eliminated. Then the rules having confidence 1 are also removed as they signify the interactions that already exist in the data set. A total of 88 rules from these 294 rules are found to be unique after the elimination step. All these rules are analysed in the following manner to predict new PPI. Let us consider, a predicted rule is,

Require: Performance Metric of each algorithm.
Ensure: Adjacency matrix (A_{ij}), where i= no. of the rows and j= no. of the columns of the matrix.

Step 1: Split the binary adjacency matrix into training (70%) and testing matrix (30%).
Step 2: Train the IMF, AMF and proposed algorithm with the training matrix.
Step 3: Count the number of 1's of the testing matrix and store the indices of the 1's in an array.
Step 4: Randomly generate 10%, 20%, 30%, 40% and 50% index positions of total number of indices and eliminate them. Update the testing matrix accordingly.
Step 5: Apply the IMF, AMF and the proposed method on the updated testing matrix to check how many artificially removed links can be detected.
Step 6: Use the performance metric formula (equation 1) to check the efficiency of each algorithm and report the values.

Fig. 3 *Algorithm 2: algorithm for performance metric calculation*

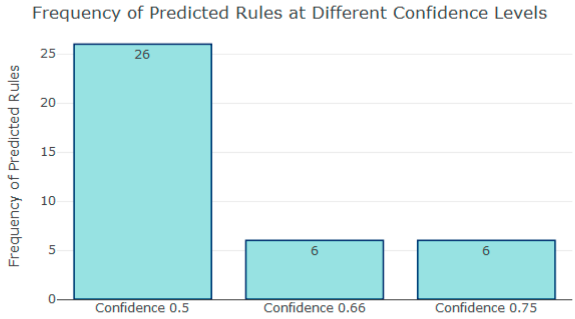

Fig. 4 *Frequency of the number of predicted interactions with annotation and direction at different confidence levels. It can be noted from the figure that at confidence level 0.5, maximum number (26) of predicted interactions are found and at confidence level 0.75 and 0.66, only six interactions are found at each level*

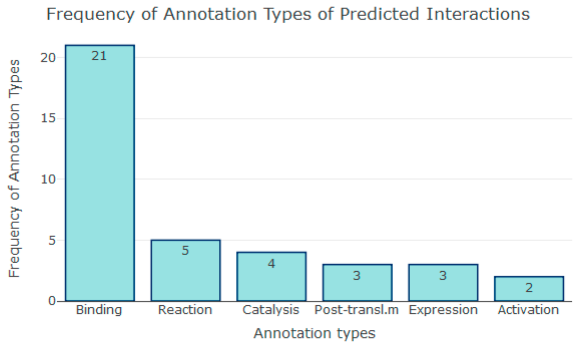

Fig. 5 *Distribution of frequency of the annotation type of the predicted interactions. It can be seen from the figure that most of the predicted interactions (21) have annotation type binding, whereas, only two PPIs have annotation type activation*

CXCR4_Activation_1,STAT3_Post-trans.m_1 ⇒ *CXCR4_Expression_1*.

The above rule can be understood in the following manner: if CXCR4 activates any protein with direction 1 and signal transducers and activators of transcription 3 (STAT3) post-translates that protein with direction 1 then there is a high chance of interaction between that protein and CXCR4 with annotation type 'Expression' and direction '1'. In this way, we can predict a list of interactions using the rules generated from the above procedure.

We have searched our database and found that the above two proteins (CXCR4 and STAT3) interact with CSF3 and CXCL12 with above annotation type and direction. There is another protein ERBB2 which is activated by CXCR4 and post-translated by STAT3 but not expressed by CXCR4. So, we can predict that CXCR4 will interact with ERBB2 with annotation type 'Expression' and direction '1'. The analysis can be depicted as follows:

From database we got the rules

(i) ERBB2 ⇒ CXCR4_Activation_1 and
(ii) ERBB2 ⇒ STAT3_Post-trans.m_1

Now, from the above rules, we can predict the interaction

ERBB2 ⇒ CXCR4_Expression_1

Employing this procedure and eliminating redundant entries we have predicted a total of 38 unique interactions, not present in the cancer database (File D4). The circulation of the confidence levels of these predicted associations varies from 0.5 to 0.75. Fig. 4 demonstrates the dissemination of the number of predicted rules at various confidence levels. If annotation type and direction of the interactions are not considered, then applying the same methodology and threshold, 55 new interactions can be predicted (File D7). The frequency of each interaction with the annotation type is shown in Fig. 5. It can be seen that most of the predicted interactions (21) have annotation type binding, whereas, only two PPIs have annotation type activation. These frequencies of annotation types also follow the ratio of the annotation types in actual cancer PPI database.

4 Statistical analysis of PPI database

In this section, the whole cancer PPI network has been modelled as a graph in which nodes represent the proteins, and edges represent the interactions between the proteins. Then the network is analysed using Cytoscape [33] network analyser. The statistical analysis of cancer PPI network exhibits the scale-free property of degree distribution of the nodes as only a few proteins interact with a large number of proteins and a majority of proteins participate in a few interactions. The degrees of all the 355 unique proteins present in the database are given in Appendix (File D6). Comparing degrees of all nodes we got eight hub proteins. They are TP53, UBC, BRCA1, EGFR, PTPN11, MUC1, HDAC1, and RUNX1. Hub proteins are very important in cancer research as they are highly expressed in diseased cells. It has been noticed that these hub proteins are present in our predicted interactions. Among 38 interactions, PTPN11 is involved in 12 cases. So, they can be used as potential drug targets. Before investigating the network, some parameters ought to be known, are clarified beneath.

Network density: The network density indicates how thickly the system is populated with edges, disregarding self-circles and copied edges.

Let G be the graph with the set of edges E and the set of vertices V. For undirected graph, the density can be defined as

$$d = \frac{2|E|}{|V|(|V|-1)}, \tag{2}$$

and for directed graph, the density can be defined as

$$d = \frac{|E|}{|V|(|V|-1)}. \tag{3}$$

The density ranges between 0 and 1. A system that contains no edges and exclusively segregated hubs has a density of 0.

Clustering coefficient: In a graph, the clustering coefficient of a node is defined as the ratio of number of edges among the neighbours of the node to the maximum number of edges that could potentially exist among the neighbours. The total network clustering coefficient can be calculated as the average of clustering coefficients of all the nodes in the network. It is basically the measure of the number of triangles in the graph and can be defined as

$$C_i = \frac{\text{Number of triangles connected to node } i}{\text{Number of triples centered around node } i}. \tag{4}$$

The clustering coefficient for the whole graph can be expressed as

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i, \tag{5}$$

where n = total number of vertices in the graph and $i \in \{1, 2, \ldots, n\}$.

Using Cytoscape we got the density of our network as 0.008 and clustering coefficient as 0.291. It is evident that the network is very sparse in nature as density is near to 0. The clustering coefficient of 0.291 signifies that every node has two to three neighbour nodes. Hence, it also supports the fact that the proteins responsible for a similar type of disease, like cancer, tend to interact with each other. The connectivity for each hub protein is presented in Fig. 6.

5 Discussion

We have predicted 38 high confident interactions, not present in the cancer PPI database, by applying biclustering based ARM on our PPI database. All these interactions are analysed and some evidence of these predicted PPI are collected from published works of literature. We have additionally looked PUBMED for exploiting some current study recognising the predicted interactions. The references to those articles can be served as a proof of our predictions. Among the 38 predicted interactions, 28 interactions are observed to be experimentally substantial. Most of the cases we got proofs of interaction type as well as direction also. Table 5 shows some predicted interactions with annotation type and direction which are validated by literature with PUBMED ID and references.

STAT3 protein are excited by various cytokines and oncogenes. Mizowaki *et al.* and Iyer *et al.* analysed the connection between STAT3 and IL10, which is a cytokine with powerful anti-inflammatory characteristics in [43, 44]. In T cells, STAT1, and STAT3 are basic for IL10 gene expression, whereas STAT3 is essential for IL-6-mediated IL-10 creation. Mouse Double Minute 2 (MDM2) participates in protein synthesis and folding and it is a proteostasis hub protein. Nicholson J *et al.* [35] showed that 8-plex iTRAQ (nanoLC-MS/MS) of MCF7 cells are excited with the MDM2-binding ligand Nutlin-3. This may help to recognise the most bounteous cell protein changes over early time focuses. Using this process 1323 unique proteins are identified and among which NPM1 is one protein having a steady-state interaction level with MDM2. Another paper [45] shows that NPM1 binds with MDM2 to prevent proteasomal degradation of p53. This supports our prediction that MDM2 binds with NPM1 and direction is 0. S100A4 is a small calcium-binding protein that is usually overexpressed with different tumor types like tumor suppressor p53. Orre *et al.* examined that S100A4 and p53 collaborate in complex tests and the collaboration increments after inhibition of MDM2-subordinate p53 degradation [37]. p53 binds with S100A4 and MDM2 and so from transitive dependency the binding relation between MDM2 and S100A4 can be established. Recently, an isoform of BAG1 protein (i.e. RAP46) has been announced to bind several steroid hormone proteins, including AR [38], which agrees our prediction result. In [42] the interaction between CXCR4 and ERBB2 is established using interstitial fluid flow technique through the tissue matrix. This procedure shows that ERBB2 expresses breast cancer cells, depends on epithelial-to-mesenchymal transition (EMT) and it behaves through a CXCR4-PI3 K pathway. So, CXCR4 expresses ERBB2 and thus we get the direction and interaction type of this interaction. PTPN11, SOS1, RAF1, and KRAS are the four responsible proteins behind the Noonan syndrome (NS) and autosomal dominant disorder. Different types of interaction between PTPN11 and SOS1 are established by some experiments and are presented in [46]. Wang

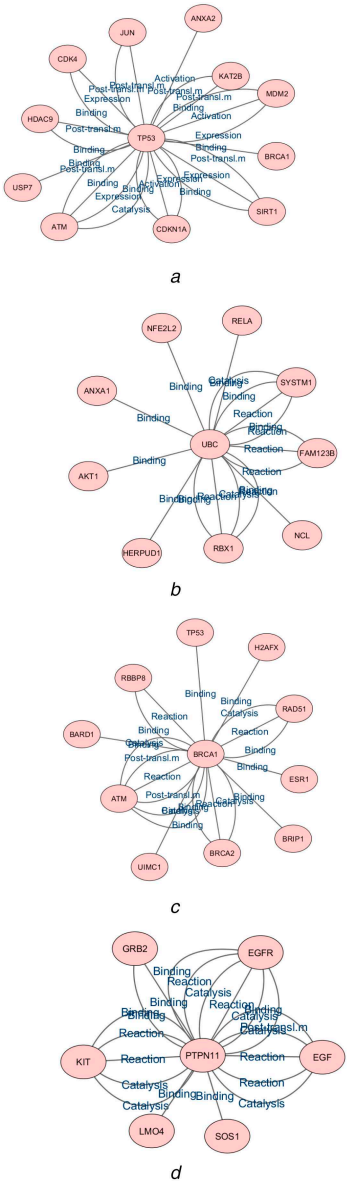

Fig. 6 *Major hub proteins of cancer database with annotation type and direction. Four major hub proteins, namely, TP53, UBC, BRCA1, and PTPN11 are shown in the figure*
(a) TP53, ***(b)*** UBC, ***(c)*** BRCA1, ***(d)*** PTPN11

Table 5 Evidences of some predicted interactions with predicted and experimentally proven annotation type, direction, PUBMED ID, and References

Protein	Interacting protein	Predicted annotation type	Predicted direction	Experimental annotation type	Experimental direction	PUBMED ID	Reference
STUB1	HSPA8	binding	0	binding	0	23880665	[34]
MDM2	NPM1	binding	0	binding	0	23039052	[35]
CREBBP	SIN3A	binding	0	binding	0	12392082	[36]
MDM2	S100A4	binding	0	binding	0	23752197	[37]
AR	HSPA8	binding	0	binding	0	23828170,9565586	[38, 39]
AKT1	IL10	post-transl.m	1	suppresses	1	21255011	[40]
AKT1	ERBB2	post-transl.m	1	activates/cascades	1	26645663	[41]
CXCR4	ERBB2	expression	1	expression	1	25566992	[42]

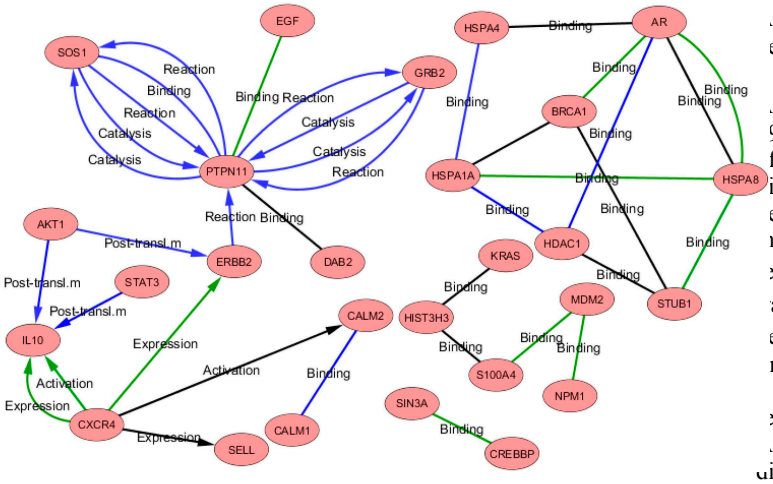

Fig. 7 *Predicted PPIs in this study. Nodes represent proteins and edges represent predicted annotation type and direction. Green edges represent the predicted interactions that are experimentally validated in the published literatures with annotation type and direction. Blue edges represent the interactions that are experimentally validated but the predicted annotation type and direction are not validated by the literatures*

et al. [47] proved that CXCR4 expression is upregulated by cytokine IL10. So, we can conclude that most of the predicted interactions are supported by recent literature along with predicted annotation type and direction. Rest of the proofs of interactions are given in File D5 with predicted and experimentally proven annotation type, direction, PUBMED ID, paper name and author name.

According to our evaluation, 28 numbers of interactions among 38 predicted interactions are found to be experimentally validated. In most of the cases (11), the predicted annotation types and directions are also supported by existing research documents. The remaining ten interactions are not found to be validated by any of the current studies. However, the probability of these interactions is very high as they are predicted from the rules having high confidence. We have seen in statistical analysis section that BRCA1, HDAC1, and RUNX1 are hub proteins. So, they interact with large number of proteins and all these interactions may not experimentally analysed. For example, we have predicted BRCA1 binds to STUB1 and HSPA1A. However, no proof of these interactions are found in published literatures. In [48], Buckley *et al.* showed that BRCA1 interacts with different proteins of HSP protein family like HSP90, HSP70 etc. (HSPA1A also belongs to HSP family). So, there is high chance that BRCA1 interacts with HSPA1A. In another study [49], we have seen that ITGA4, SELL, CXCR3, and CXCR4 act as homing receptors in blood lymphocyte subsets of normal pregnant women. So, CXCR4 and SELL are expressed in same manner under certain scenario and they may interact with each other. Hence we propose that these ten predicted interactions, which do not have any direct evidence from literature, are candidates for possible experimental validation. They are likely to be involved in various cancer types as per the outcome of the proposed computational approach.

All the predicted interactions are shown in Fig. 7. Analysing with published literatures against our predictions, we found that 28 interactions are validated by experimental results. Out of 28, 11 interactions' predicted annotation types and directions are supported by published literatures (File D5). All network diagrams are created in Cytoscape.

6 Comparison between proposed method and classical ARM

To evaluate the performance of the proposed algorithm with that of the ARM, we have analysed the time complexity and execution time for both methods. Although, both the algorithms can generate the same set of rules, the complexity of classical ARM is high compared to Bimax-based ARM. If the transactional data set contains a d number of unique items, then the time complexity of .priori-based ARM is $O(2^d)$. In [49], the authors have emonstrated that if the data set contains a d number of items, then total of $3^d - 2^{d+1} + 1$ number of rules are generated using .priori. So, for a small data set, with say $d = 7$, it will generate 932 rules. The performance of ARM degrades when the number f items in the data set increases. However, most of these rules are iscarded after applying *minsup* and *minconf* as the algorithm enerates frequent itemsets first and then calculates rules from ıese frequent itemsets. The frequent itemset generation step equires $O(t * 2^{k-1})$ time complexity, where t is the number of ansactions and 2^{k-1} is the number of candidate itemsets. The rule eneration step extracts association rules and it takes $O(2^k - 2)$ me complexity for each frequent k-itemset.

This approach is very expensive due to the huge space and time equirement. To minimise these exponential time complexity of .RM frequent itemset mining, biclustering technique is used to irectly mine the closed frequent itemsets reducing the time and space complexity required for mining all frequent itemsets. The number of FCIs generated by Bimax is much less than the number of frequent itemsets produced by Apriori-based ARM, and they contain non-redundant information. Bimax biclustering algorithm takes $O(m * n)$ time and space complexity to obtain the maximal closed frequent itemsets, where m is the number of rows and n is the number of columns of the adjacency matrix. The number of total biclusters could be huge. So, Bimax algorithm reduces the search space by choosing only maximal biclusters. Then rule generation from these maximal biclusters using ARM takes $O(2^k)$ time, where k is the number of maximal frequent itemsets. So, the total time complexity of the biclustering based ARM is $O(m * n) + O(2^k)$.

We have done the experiments on a PC with 4 GB RAM and 2.4 GHz processor running Windows 7 using python programming language. On the same cancer data set, considering same confidence and support, the traditional ARM takes 2.1 min, whereas biclustering-based ARM takes only 30 seconds to generate the rules.

7 Conclusion

In this article, we have given a detailed explanation of association rules to find the specific interactions of the cancer disease that are biologically significant. For this, we have prepared the cancer-human PPI database with annotation type and direction to identify new interactions among several human proteins responsible for various cancer disease. A total of 38 new interactions that are not present in the cancer PPI database (i.e. in String database) are discovered in this study. The algorithm is giving 74% precision as out of 38, 28 interactions have been found to be supported by some recent literature. Among these 28 interactions, 11 interactions' predicted annotation types and directions are also supported by various literature. These agree with the fact that the predicted interactions which are not present in the database, exist actually. So, it reduces the task of biologists to some extent by investigating only the predicted interaction with annotation type and direction. They don't need to check all possible combinations of interaction and annotation type of PPI.

Different studies have been used in different methodologies and different data sets to explore cancer PPI. Therefore, it is not justified to compare all these methods considering only the predicted sets of interaction. This approach based on biclustering-based ARM has a clear advantage over conventional methods because of no loss of relevant information and also no need for negative interaction data set. However, in this study, we have considered a non-weighted PPI network with annotation type and direction. A weight field can be added with each PPI to represent the number of evidence of each interaction found from different databases. This field can be used to build a weighted PPI network which can be mined to predict new interactions. In addition, we plan to improve this approach by incorporating additional features like domain, gene ontology, amino acid composition, etc. of these interacting proteins.

Here, we have analysed the proposed methodology to predict the proteins responsible for cancer disease only. The same procedure may also be implemented for proteins related to other diseases like ebola, zika, dengue, HIV, etc. to predict the PPI.

8 Acknowledgment

A.M. acknowledges the support received from the research project (Memo No: 355(Sanc.)/ST/P/S&T/6G-10/2018 dt. 08/03/2019) of Department of Science & Technology and Biotechnology, Government of West Bengal, India at University of Kalyani.

9 References

[1] Jansen, R., Yu, H., Greenbaum, D., *et al.*: 'A Bayesian networks approach for predicting protein-protein interactions from genomic data', *Science*, 2003, **302**, (5644), pp. 449–453

[2] Ben-Hur, A., Noble, W.S.: 'Kernel methods for predicting protein–protein interactions', *Bioinformatics*, 2005, **21**, (suppl 1), pp. i38–i46

[3] Zhang, Q.C., Petrey, D., Deng, L., *et al.*: 'Structure-based prediction of protein-protein interactions on a genome-wide scale', *Nature*, 2012, **490**, (7421), pp. 556–560

[4] Tomczak, K., Czerwiska, P., Wiznerowicz, M.: 'The cancer genome atlas (tcga): an immeasurable source of knowledge', *Contemp. Oncol.*, 2015, **19**, (1A), p. A68

[5] Zhang, J., Baran, J., Cros, A., *et al.*: 'International cancer genome consortium data portal—a one-stop shop for cancer genomics data', *Database*, 2011, **2011**, pg. bar026

[6] Berk, A., Zipursky, S., Lodish, H.: '*Molecular cell biology*' (W. H. Freeman, New York, 2000, 4th Edn.)

[7] Hasegawa, H.: 'Kernel methods for predicting protein-protein interactions', 2008

[8] Kar, G., Gursoy, A., Keskin, O.: 'Human cancer protein-protein interaction network: a structural perspective', *PLoS Comput. Biol.*, 2009, **5**, (12), p. e1000601

[9] Li, Z., Ivanov, A.A., Su, R., *et al.*: 'The oncoppi network of cancer-focused protein–protein interactions to inform biological insights and therapeutic strategies', *Nat. Commun.*, 2017, **8**, p. 14356

[10] Mani, K.M., Lefebvre, C., Wang, K., *et al.*: 'A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas', *Mol. Syst. Biol.*, 2008, **4**, (1), p. 169

[11] Guda, P., Chittur, S.V., Guda, C.: 'Comparative analysis of protein-protein interactions in cancer-associated genes', *Genomics Proteomics Bioinformatics*, 2009, 7, (1-2), pp. 25–36

[12] Wu, G., Feng, X., Stein, L.: 'A human functional protein interaction network and its application to cancer data analysis', *Genome Biol.*, 2010, **11**, (5), p. R53

[13] Taylor, I.W., Linding, R., Warde-Farley, D., *et al.*: 'Dynamic modularity in protein interaction networks predicts breast cancer outcome', *Nat. Biotechnol.*, 2009, **27**, (2), p. 199

[14] Szklarczyk, D., Franceschini, A., Kuhn, M., *et al.*: 'The string database in 2011: functional interaction networks of proteins, globally integrated and scored', *Nucleic Acids Res.*, 2011, **39**, (suppl 1), pp. D561–D568

[15] Von Mering, C., Huynen, M., Jaeggi, D., *et al.*: 'String: a database of predicted functional associations between proteins', *Nucleic Acids Res.*, 2003, **31**, (1), pp. 258–261

[16] Chatr-Aryamontri, A., Oughtred, R., Boucher, L., *et al.*: 'The biogrid interaction database: 2017 update', *Nucleic Acids Res.*, 2017, **45**, (D1), pp. D369–D379

[17] Xenarios, I., Salwinski, L., Duan, X. J., *et al.*: 'Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions', *Nucleic Acids Res.*, 2002, **30**, (1), pp. 303–305

[18] Maulik, U., Bhattacharyya, M., Mukhopadhyay, A., *et al.*: 'Identifying the immunodeficiency gateway proteins in humans and their involvement in microrna regulation', *Mol. BioSyst.*, 2011, **7**, (6), pp. 1842–1851

[19] Mondal, K.C., Pasquier, N., Mukhopadhyay, A., *et al.*: 'A new approach for association rule mining and bi-clustering using formal concept analysis'. Int. Workshop on Machine Learning and Data Mining in Pattern Recognition, Berlin, Germany, 2012, pp. 86–101

[20] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: 'A novel biclustering approach to association rule mining for predicting hiv-1–human protein interactions', *PLOS One*, 2012, **7**, (4), p. e32289

[21] Ben-Hur, A., Noble, W.S.: 'Choosing negative examples for the prediction of protein-protein interactions'. BMC Bioinformatics, Whistler, British Columbia, Canada, 2006, Vol. 7, p. S2, BioMed Central

[22] Barman, R.K., Saha, S., Das, S.: 'Prediction of interactions between viral and host proteins using supervised machine learning methods', *PloS One*, 2014, **9**, (11), p. e112034

[23] Liberona, J.L., Powell, J.A., Shenoi, S., *et al.*: 'Differences in both inositol 1, 4, 5-trisphosphate mass and inositol 1, 4, 5-trisphosphate receptors between normal and dystrophic skeletal muscle cell lines', *Muscle Nerve*, 1998, **21**, (7), pp. 902–909

[24] Yim, S., Yu, H., Jang, D., *et al.*: 'Annotating activation/inhibition relationships to protein-protein interactions using gene ontology relations', *BMC Syst. Biol.*, 2018, **12**, (1), p. 9

[25] Eom, J.-H., Zhang, B.-T.: 'Prediction of protein interaction with neural network-based feature association rule mining', *Neural Inf. Process.*, 2006, **4234**, pp. 30–39

[26] Sahoo, S.S., Swarnkar, T.: 'A theoretical approach for augmenting association rule mining to predict protein-protein interaction', *Exp. Tech.*, 2011, **2**, (5), p. 8

[27] Gyenesei, A., Wagner, U., Barkow-Oesterreicher, S., *et al.*: 'Mining co-regulated gene profiles for the detection of functional associations in gene expression data', *Bioinformatics*, 2007, **23**, (15), pp. 1927–1935

[28] Mukhopadhyay, A., Ray, S., Maulik, U.: 'Incorporating the type and direction information in predicting novel regulatory interactions between hiv-1 and human proteins using a biclustering approach', *BMC Bioinformatics*, 2014, **15**, (1), p. 26

[29] Voggenreiter, O., Bleuler, S., Gruissem, W.: 'Exact biclustering algorithm for the analysis of large gene expression data sets', *BMC Bioinformatics*, 2012, **13**, (Suppl 18), p. A10

[30] Barkow, S., Bleuler, S., Preli, A., *et al.*: 'BicAT: a Biclustering Analysis Toolbox', *Bioinformatics*, 2006, **22**, (10), pp. 1282–1283

[31] Acar, E., Dunlavy, D.M., Kolda, T.G.: 'Link prediction on evolving data using matrix and tensor factorizations'. 2009 IEEE Int. Conf. on data mining workshops, Miami, Florida, USA, 2009, pp. 262–269

[32] Yokoi, S., Kajino, H., Kashima, H.: 'Link prediction in sparse networks by incidence matrix factorization', *J. Inf. Process.*, 2017, **25**, pp. 477–485

[33] Shannon, P., Markiel, A., Ozier, O., *et al.*: 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Res.*, 2003, **13**, (11), pp. 2498–2504

[34] Ferreira, J.V., Fôfo, H., Bejarano, E., *et al.*: 'Stub1/chip is required for hif1a degradation by chaperone-mediated autophagy', *Autophagy*, 2013, **9**, (9), pp. 1349–1366

[35] Nicholson, J., Neelagandan, K., Huart, A.-S., *et al.*: 'An itraq proteomics screen reveals the effects of the mdm2 binding ligand nutlin-3 on cellular proteostasis', *J. Proteome Res.*, 2012, **11**, (11), pp. 5464–5478

[36] Lee, M.-O., Kang, H.-J.: 'Role of coactivators and corepressors in the induction of the rar. beta. Gene in human colon cancer cells', *Biol. Pharm. Bull.*, 2002, **25**, (10), pp. 1298–1302

[37] Orre, L., Panizza, E., Kaminskyy, V., *et al.*: 'S100a4 interacts with p53 in the nucleus and promotes p53 degradation', *Oncogene*, 2013, **32**, (49), pp. 5531–5540

[38] Froesch, B.A., Takayama, S., Reed, J.C.: 'Bag-1 l protein enhances androgen receptor function', *J. Biol. Chem.*, 1998, **273**, (19), pp. 11660–11666

[39] Kwon, J. Y., Weon, J.-I., Koedrith, P., *et al.*: 'Identification of molecular candidates and interaction networks via integrative toxicogenomic analysis in a human cell line following low-dose exposure to the carcinogenic metals cadmium and nickel', *Oncol. Rep.*, 2013, **30**, (3), pp. 1185–1194

[40] Antoniv, T.T., Ivashkiv, L.B.: 'Interleukin-10-induced gene expression and suppressive function are selectively modulated by the pi3k-akt-gsk3 pathway', *Immunology*, 2011, **132**, (4), pp. 567–577

[41] Carmona, F.J., Montemurro, F., Kannan, S., *et al.*: 'Akt signaling in erbb2-amplified breast cancer', *Pharm. Therapeutics*, 2016, **158**, pp. 63–70

[42] Tchafa, A.M., Ta, M., Reginato, M.J., *et al.*: 'EMT transition alters interstitial fluid flow-induced signaling in erbb2-positive breast cancer cells', *Mol. Cancer Res.*, 2015, **13**, (4), pp. 755–764

[43] Mizowaki, T., Sasayama, T., Tanaka, K., *et al.*: 'STAT3 activation is associated with cerebrospinal fluid interleukin-10 (il-10) in primary central nervous system diffuse large b cell lymphoma', *J. Neuro-Oncol.*, 2015, **124**, pp. 1–10

[44] Iyer, S.S., Cheng, G.: 'Role of interleukin 10 transcriptional regulation in inflammation and autoimmune disease', *Crit. Rev.Immunol*, 2012, **32**, (1), pp. 23–63

[45] Box, J.K., Paquet, N., Adams, M.N., *et al.*: 'Nucleophosmin: from structure and function to disease development', *BMC Mol. Biol.*, 2016, **17**, (1), p. 19

[46] Brasil, A.S., Malaquias, A.C., Wanderley, L.T., *et al.*: 'Co-occurring ptpn11 and sos1 gene mutations in Noonan syndrome: does this predict a more severe phenotype?', *Arquivos Brasileiros de Endocrinologia Metabologia*, 2010, **54**, (8), pp. 717–722

[47] Wang, J., Guan, E., Roderiquez, G., *et al.*: 'Role of tyrosine phosphorylation in ligand-independent sequestration of cxcr4 in human primary monocytes-macrophages', *J. Biol. Chem.*, 2001, **276**, pp. 49236–49243

[48] Buckley, N., D'costa, Z., Kaminska, M., *et al.*: 'S100a2 is a brca1/p63 coregulated tumour suppressor gene with roles in the regulation of mutant p53 stability', *Cell Death Dis.*, 2014, **5**, (2), p. e1070

[49] Burke, S.D., Seaward, A.V., Ramshaw, H., *et al.*: 'Homing receptor expression is deviated on cd56 + blood lymphocytes during pregnancy in type 1 diabetic women', *PloS One*, 2015, **10**, (3), p. e0119526

10 Appendix

10.1 File D1

Excel file containing different cancer diseases and the proteins responsible for the particular types of cancer (XLSX).

10.2 File D2

Excel file containing human cancer related PPIs with annotation type and direction (XLSX).

10.3 File D3

Excel file containing adjacency matrix with annotation type and direction of cancer related PPI network (XLSX).

10.4 File D4

Excel file containing predicted PPI with annotation type and direction (XLSX).

10.5 File D5

Excel file containing predicted PPI's evidence with PUBMED ID (XLSX).

10.6 File D6

Excel file containing degree distribution of all cancer-focused proteins.

10.7 File D7

Excel file containing predicted PPI without considering annotation type and direction of interactions.

10.8 File D8

Codes and flowchart of the proposed methodology. Source: https://sites.google.com/site/biclusteringbasedarm/biclustering-code