Preprocessing of Colorectal Cancer Protein Expressions Using Correlation Co-Efficient Factor and Disease Score

B Madhav Rao[1*], M Sampath Kumar[2]

[1] *Research Scholar, Department of CS&SE, A.U College of Engineering(A), Andhra University, Visakhapatnam, India*
[2] *Professor, Department of CS&SE, A.U College of Engineering(A), Andhra University, Visakhapatnam, India*
[*]*Corresponding Author:* madhavraob@gmail.com

***Abstract*—**Preprocessing is one of the important techniques in identifying the molecular sequences. Identifying the Colorectal Cancer protein sequences is one of the challenging task. So, an approach is needed to discover and identify the related proteins as well as targeted treatment from the large data set, so a preprocessing technique called as seed proteins co expressions Protein-Protein Interaction (PPI) system is implemented. In this algorithm first construction of protein network, then consideration of the disease score and fix the threshold value based on that filter the network. Finally consider the properties of protein sequences and correlate the properties based on co-efficient co relation algorithm, based on the result the network is splitted into sub-networks. This method filters the top proteins which normally effects the identification in colorectal cancer. There are proteins which effects in 33 regions of Colorectal Cancer such as EGF, PLAU, HRAS and TP53 by using the data set of 2000 proteins. This method purifies and identifies the most effected protein expressions for the clinical outcome to understand the functional mechanism of essential proteins. Some computational methods and topological features with biological data has been taken to explore the purification of protein sequences. These functional mechanisms can be applied for invention of Colorectal cancer drugs, so this approach attracts more since it reduces the cost factor.

***Keywords*—**Colorectal Cancer, Protein-Protein Interaction, Seed Protein, Sub Network, Coefficient Correlation

I. INTRODUCTION

Colorectal Cancer(CRC) is the third most commonly occurring cancer in men and the second most commonly occurring cancer in women and account for 9.4% of all cancer deaths. There were over 1.9 million new cases in 2018 [2]. In developing nations, the ascent of colorectal malignant growth can be because of the rising maturing populace, horrible present day eating designs, and an expansion in danger factors, for example, smoking, low active work, and weight, so distinguishing proof at the beginning phase is one of the significant components. The occurrence rates are geographically different, with more than half of the cases of CRC occurring in developing countries. Be that as it may, mortality is higher in the less evolved nations who have restricted assets and insufficient wellbeing foundation. Death rates have been diminishing in numerous Western nations because of a blend of different components like early recognition because of screening and improved treatment of CRC [1].

India has a low commonness of CRC—assessed five-year predominance is 87 for every 100,000 populaces. Contrasts in dietary examples and ways of life are believed to be liable for the low occurrence of CRC in the creating scene. Additionally, pervasiveness of weight which is a danger factor for CRC contrasts in the created and the creating scene. Another conceivable explanation behind low occurrence can be a more youthful populace—CRC is more normal in the old. It ought to be noticed that the populace vaults in India cover just 7.45% of the populace, while overall malignant growth libraries cover 21% of the populace; thus, some measure of under revealing might be conceivable in India. Nonetheless, concentrates on Indian foreigners from nations with a high predominance of CRC like the USA and Singapore show that CRC frequency is lower in Indians than in the local populace however higher than that saw from the Indian registries [4]. This shows in spite of the fact that there are probably going to be some hereditary components associated with the lower occurrence of CRC, ecological factors additionally have a task to carry out. As of late, with the advancement of high-throughput biotechnologies, a lot of organic information has been created. For example, yeast two-half and half frameworks, protein complex and quality articulation profiles, and so forth. These information are valuable assets for concluding and understanding quality capacities [5]. Up until now, the Protein-Protein Interaction (PPI) information has been broadly utilized for quality capacity forecast with the supposition that connecting proteins share the equivalent or have comparable capacities and henceforth might be engaged with a similar pathway. This "liable by affiliation" rule was first proposed by Nabieva et al. what's more, can likewise be utilized to recognize malignant growth related qualities [6].

The advancement of CRC is normally seen with the guide of the initiation of the KRAS and BRAF qualities and the restraint of the p53 tumor silencer quality articulation; changes in those qualities are identified with adjustments in the number and state of chromosomes. The fundamental driver of CRC incorporates dietary and natural components, just as hereditary transformations. The concurrent methylation of the CpG site and transformation of the BRAF quality are

additionally significant components in the turn of events and movement of CRC [3]. Inability to identify CRC early might be one reason for helpless anticipation in patients. Subsequently, there is a critical requirement for effective indicative and restorative techniques. The forecast of CRC is poor because of an absence of successful analytic techniques at a beginning phase. In this way, a successful arrangement must be accommodated resulting analysis and therapy by better understanding the quality articulation of CRC during its event and advancement and distinguishing the qualities that might be associated with the event and movement of disease [3].

PPI networks contains the worth data to comprehend the organic which are the regular highlights the development of PPI networks save the comparable natural capacities. There are numerous online devices for recovering the data of the PPI information, for example, Uniport, STRING which contains an immediate and circuitous relationship of natural information. Recognize the transformations which give the development of the Colorectal malignancy[7]. In CRC epigenetic and change level data functions as significant part to consider the protein co-articulation and PPI examination [8].

II. PPI DETECTION METHODS

Yeast Two-Hybrid and Co-Immunoprecipitation methods are used for identifying the protein interaction. By using these two methods, the proposed approach identifies the physical binding. Co-immunoprecipitation (Co-IP) is a mainstream strategy to distinguish physiologically pertinent protein–protein communications by utilizing objective protein-explicit antibodies. These protein buildings would then be able to be investigated to distinguish new restricting accomplices, restricting affinities, the energy of official and the capacity of the objective protein. Along these lines, Co-IP is viewed as one of the standard strategies for recognizing or affirming the event of protein-protein association occasions in vivo. Co-IP examinations can recognize proteins by means of immediate or backhanded associations or in a protein complex[26].Yeast Two-Hybrid method is used for detecting pair wise proteins which consist of directly binding to DNA and Activating Domain(AD) [7][11]. This method activates physical association of the binding domain.T hese proteins are called hybrid proteins when they fused. The reported proteins can be used for measuring and detection of PPI proteins [7][9]. In this method, to increase the quality of experiments, take correlation between the interacting proteins and it can be represented as a matrix. Further it is divided into different categories using MIPS. The abovementioned model considers all protein interactions for purification. In this juncture, there is a chance to get false negative results with non-interaction proteins. The matrix model contains a complex pair wise interactions and pair wise with all positive interactions.

Different Computational approaches are needed for analyzing the complex networks for identifying the complex proteins. Many approaches have been proposed such as geometric approach, least square regression approach, co-expression networks and so on. These approaches will not purify accurately. So, a new method need to develop which can be more robust molecular predict to identify novel proteins [14][10]. The prediction of complex networks and principle of cellular organisms are key functions. So far, many algorithms have been proposed to purify the complex network such as MCODE, CMC, MCL. Random walk algorithm is applied for simulating PPI networks to identify protein complexes [15] [16][17]. These methods only focus on static networks since most of the networks are dynamic. So, identification of protein complexes is a critical task. So, injected sequence expression is used for detecting complex networks. Dynamic hierarchical clustering is used for presenting complex sequences dynamically along with DPC algorithm [18][19][20]. Core part of a protein complex is used for similarity measures sequence ontology. To find the structure of the sequence weights between proteins are used. All these approaches still lack in purification of best sequences [21][22][23][24].In this study extracting the Colorectal cancer data from the STRING database and then construct the network and sub network with protein co-expressions using PPI system. The network is denoted as seed proteins and candidate proteins. By using this approach, the common proteins can be found. Later, consider the parameters disease score and co-efficient correlation then analyze the network to filter the data and utilize the identified proteins to calculate the risk factors.

III. IMPLEMENTATION PROCESS

In this proposed method collect the data from STRING database and use the seed proteins to construct the biological network using Cytoscape software. Later construct the sub network to identify the common proteins which are related to common proteins collected from TCGA[12]. Then the co-expression network will be constructed using Python programming which contains the information about the co-expression and PPI information of a biological network. This represents the number of nodes and edges and PPI network of 2000 seed proteins collected from Colorectal cancer data base. By using this preprocessing technique 179 proteins can be obtained. To get the above protein expressions accurately disease score of greater than 0.75 and correlation coefficient factor 0.00004 parameters are used. With 4 sub networks, 179 common proteins are identified as key candidate proteins of Colorectal cancer. In this technique a pair of queries has been given to online protein data base to retrieve information about each protein with the annotation information and also gather's information about partially connected proteins using dynamic parsing which allows a greater number of proteins from the neighbours information.

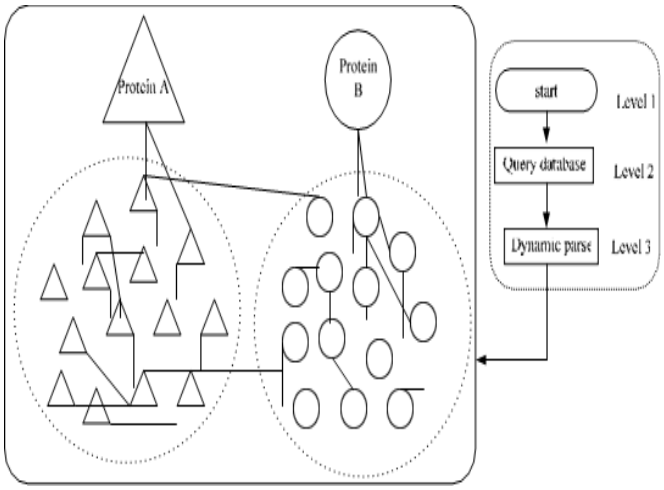

Figure 1: dynamic parsing of query base PPI sequence

In this coefficient correlation method where protein is represented as a node and edges which represents the interaction between to corresponding proteins. Given a PPN query PN = (PV, PE) and query protein pairs PP(a) and PP(b) included in the PV, it uses the correlation similarity algorithm to obtain PN for sub-networks. After the network construction, neighbor proteins network partition method is used which represents as a node.

3.1. Algorithm

Step 1: Importing Colorectal Cancer causing protein.

Step 2: Compute the disease score of each node.

Step 3: Remove the nodes from thePPI network which are less than the threshold value of0.75.

Step 4: Construct the vertex basing on disease score for a degree of a graph.

Step 5: Select the uncollected edges to find the maximal number of network paths which cross the edge.

Step 6: Repeat the step 3 and4 to find the sub network of PP(a) and PP(b).

Step 7: Construct the sub network and subsets.

Step 8: Apply the co-efficient correlation algorithm to preprocess and filter the top sequences with threshold value of 0.00004.

Step9: Visualize the resultant network.

3.2. Construction of PPI Network Based on Disease Score Filtration

Filter the network by considering parameter disease score and taken threshold value 0.75 and filter the network. Only consider the protein nodes which are greater than the threshold value. This proposed method gives 879 proteins with 5681 edges from the given 4000 proteins nodes and its associate edges.

3.3. Correlation Coefficient Calculation Method

Correlation Coefficient (CC) algorithm is used for finding linear co-efficient correlation between the given two sequences by considering the similarities between two sequence variables to calculate the similarity measure the sequence is denoted $s1=(s1_1,s1_{2,}\text{----------},s1_m)$ and $s2=(s2_1,s2_{2,}\text{----------},s2_m)$ and the properties between two proteins can be calculated as

$$CC = \frac{\sum(s1_i - \overline{s1})(s2_i - \overline{s2})}{\sqrt{\sum(s1_i - \overline{s1})^2 \sum(s2_i - \overline{s2})^2}} \quad (1)$$

Finally, co-efficient similarity between two proteins is achieved using identity coefficient as in equation-2 that is sum of two sequence similarity co-efficient.

$$ICC(CC_i) = \sum CC(s1_i, s2_j) \quad (2)$$

With this preprocessing technique, extract the sub networks and obtain 176 seeds of the overlapping proteins which have a strong relationship and found that these proteins place a vital role with other related proteins. The protein encodes hydrolyses family it binds to the proteins and converts to hydroids which enables the growth of tumor cells [17].The family of glucagon proteins which simulates the growth from the gland and it is cleaved to form somatoliberin. Proteinrate the sub network of PPI protein expressions a standalone platform and dependent software has been used for proteinrating the sub network. The input data of the seed proteins and common proteins are used by using weightage and Coefficient Correlation are taken into consideration to construct the network and identify the most common proteins which causes the Colorectal Cancer[13].By applying the Correlation Coefficient parameter with threshold value of 0.00004 to preprocess the network. The Network is pruned with the top most proteins which cause Colorectal cancer.

3.4. Identifying Neighbour Proteins in the PPI network

In the proposed method examine the co expression network with the similar feature after importing seed proteins which represents protein co expressions and PPI information [l1].With the weightage algorithm to calculate the edge weightage of each protein and filter the data and apply matrix method to construct the purification of interacting system [25].In this method the candidate proteins are considered to obtain the sub networks and represent in vector format. This method uses 176 proteins which are preprocessed and used to find the neighbour proteins which are closely associate with them. It results 852 link proteins and 176 base proteins total number of 1028proteins form a network by using Cytoscape software and analyze it.

IV. RESULTS

In this method proteins related to Colorectal Cancer are extracted from STRING database by using Python script and form a network by using Cytoscape software. The result is a network before preprocessing as shown in Figure 2.

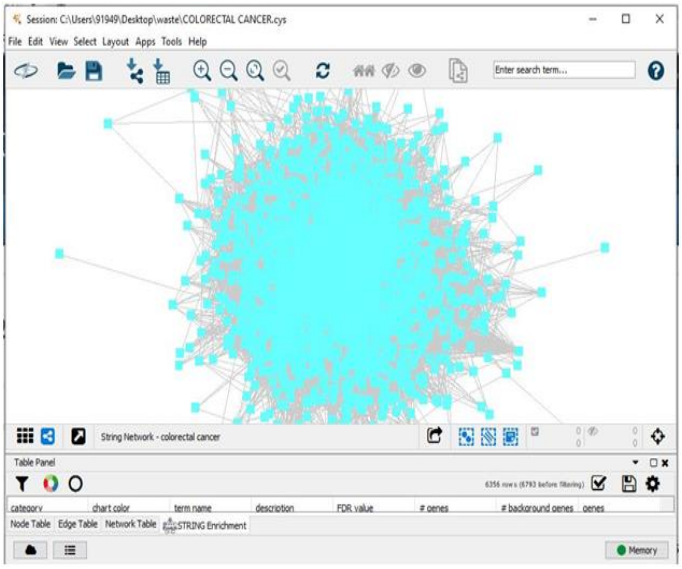

Figure 2: Protein ProteinInteraction network with 2000 nodes

After constructing the unprocessed protein network, the first step is to filter the network with disease score parameter with the threshold value of 0.75.After performing this filtration, the result is1768 proteins out of 2000 proteins. The network visualization of 1768 proteins with associated edges as shown in Figure 3.

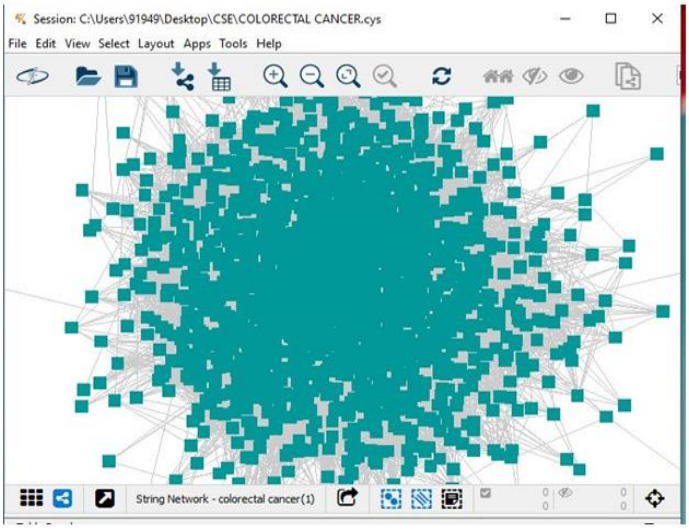

Figure 3: Protein-ProteinInteraction network with 879nodes filtered with disease score>=0.75.

In second step the partially pruned network is filtered by using Coefficient correlation factor with the threshold value of 0.00004. If the taken value is less than the threshold value, we get only very few proteins with that it is not possible to analyze the network. After applying this cutoff value, the result network is formed with 176 proteins. These are the top most proteins which are common in majority of the databases like mint, Bio GRID, Uniport. The resultant network with 176 pruned proteins as shown in Figure 4.

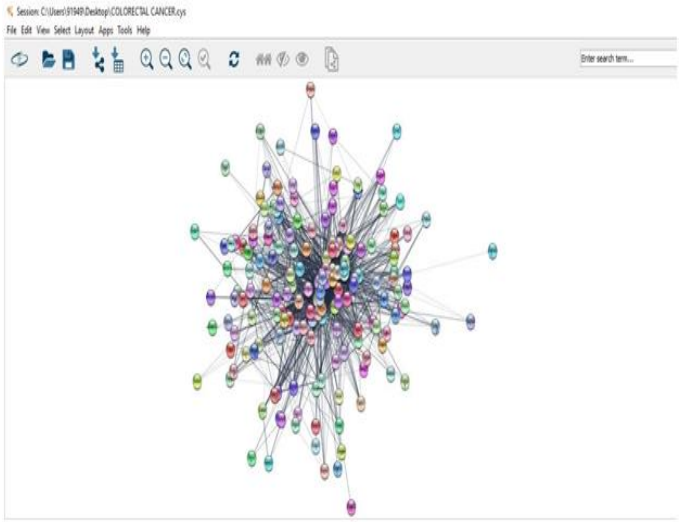

Figure4: Filtered data with default Coefficient-correlation=" .000004" parameter.

Figure 5 shows the parameters which are available for analysis of the proteins. In protein node table some of the available parameters are disease score, name, sequence, confidence score, correlation value, display nameetc,.

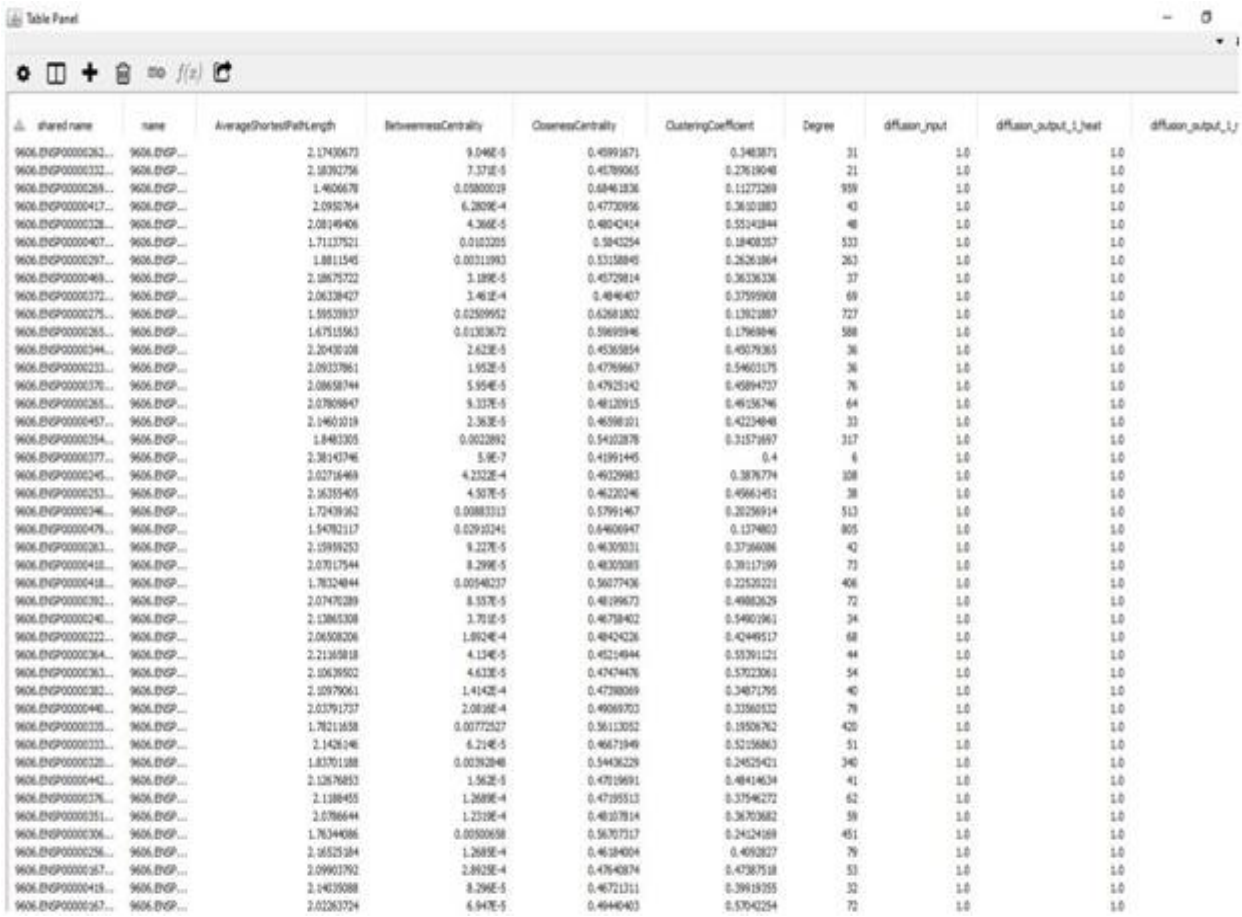

Figure 5: The network details of the seed and protein data.

The number of nodes and edges of 176 top most protein sequences extracted after applying disease score and correlation coefficient factor Are shown in Table-1.

Table-1: The number of nodes and edges of 176 protein sequences extracted after preprocessing

	PPI Network Filter with Disease score	PPI Network with Neighboring proteins	PPI Network Filter with Correlation Coefficient
Proteins	1768	1028	176
Proteins Associated Edges	89544	2400	1842

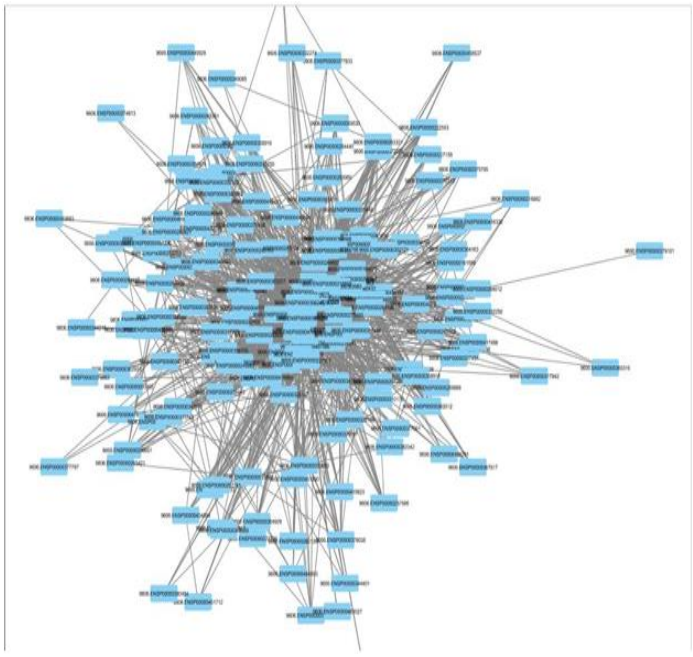

Figure6: The Sub Network proteinrated from seed and protein weightage.

Based on 176 pruned proteins results after the filtration by using disease score and correlation coefficient. This method

finds the neighbouring proteins which are linked with the pruned proteins as shown in Figure 6.

5. CONCLUSION

With this approach, preprocess the protein sequence data by filter the network based on disease score and correlation coefficient parameters to get the top most 176 Colorectal Cancer proteins. The resultant network is visualized with the help of Cytoscape software. This pruned protein sequences are common in various databases like UniProt, MINT and IntAct. In this protein network the proposed method implements the micro array analysis of co expressed proteins across the samples to identify the related proteins which are frequently conserved, transcript modules of protein sequence data in which proteinrate multi-dimensional mapping. By using Python script construct the co-relation network by taking the consideration of disease score and evaluate the distance between the pair of proteins using adjacent function. This method has assessed their individual contribution to produce the best performance results in a number of validation tests.

REFERENCES

[1]. Prachi S.Patil, Avanish Saklani, Pravir Gambhire, Shaesta Mehta, Reena Engineer,AshwinDe'Souza,Supriya Chopra and Munita Bal "Colorectal Cancer in India: An Audit from a Tertiary Center in a Low Prevalence Area ", Dec; 8(4): 484–490 ,doi: 10.1007/s13193-017-0655-0.

[2]. Bray F, Ferlay J, Soerjomataram I, et al. "Global cancer statistics 2018: GLOBOCAN

Estimatesof incidence and mortality worldwide for 36 cancers in 185 countries". CA Cancer Journal, Clin. 2018; 68, pp:394–424.

[3]. ZhihuaChen, Yilin Lin, JiGao, SuyongLin, YanZheng, YisuLiu, Shao Qin Chen," Identification of key candidate genes for colorectal cancer by bioinformatics analysis", Oncology Letters, December-2019 Volume 18 Issue 6, October 17, 2019

https://doi.org/10.3892/ol.2019.10996,pp: 6583-6593.

[4]. Prachi S. Patil, AvanishSaklani,PravirGambhire,Shaesta Mehta,Reena Engineer, Ashwin De'Souza,Supriya Chopraand Munita Bal"Colorectal Cancer in India: An Audit from a Tertiary Center in a Low Prevalence Area", Indian Journal of Surgical Oncology. 2017 Dec; 8(4): 484–490. Published online 2017 Apr 22.

doi: 10.1007/s13193-017-0655-0.

[5]. Li, Bi-Qing, et al. "Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network." PLoS ONE, vol. 7, no. 4, 2012, p. e33393. Gale OneFile: Health and Medicine.

[6]. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps", Bioinformatics 21, June,2005, Suppl 1: i302–10.

[7]. Sparano JA, Paik S "Development of the 21-protein assay and its application in clinical practice and clinical trials". Journal of Clinical Oncology, Volume 26 Issue 5 February 10, 2008, pp:721-728.

[8]. Von der HeydeS, Wagner S, Czerny A, Nietert M, LudewigF, Salinas-RiesterG, Arlt D, BeiBbarthT. mRNA profiling reveals determinants of trastuzumab efficiency in HER2- positive breast cancer. PLoSOne. 2015; 10: e0117818.

[9]. Zhu L, You Z-H, Huang D-S, Wang B. t-LSE: a novel robust geometric approach for modeling protein- protein interaction networks. PLoSOne. 2013;8:e58368 .

[10]. Huang DS, Zhang L, Han K, Deng S, Yang K, Zhang H, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression",Curr Protein Pept Sci. 2014;15(6):553-60. doi: 10.2174/1389203715666140724084019.

[11]. Deng S-P, Zhu L, Huang D-S,"Predicting Hub Proteins Associated with Cervical Cancer through Protein Co- expression Networks", IEEE/ACM Trans ComputBiolBioinform. 2015; 13, pp:27-35.

[12]. FranceschiniA, Szklarczyk D, Frankild S, Kuhn M, SimonovicM, Roth A, Lin J, MinguezP, Bork P, von MeringC. STRING v9. 1: Protein-Protein Interaction networks, with increased coverage and integration . Nucleic Acids Res. 2013;41:D808-D815 .

[13]. Mancino M, Ametller E, Gasc6nP, Almendro V, "The neuronal influence on tumor progression",BiochimBiophysActa . 2011; 1816:105-118.

[14]. PospisilP, IyerLK, AdelsteinSJ, KassisAI.,"A combined approach to data mining of textual and structured data to identifycancer-relatedtargets", BMC Bioinform atics.2006;7:1.

[15]. MMACl/TEP,"Tumor suppressor protein decreases cell growth and induces apoptosis and anoikis in breast cancer cells", Oncoprotein.1999; 18:7034-7045.

[16]. Bader, G.D.; Hogue, C.W.V.An automated method for finding molecular complexes in large protein interaction networks .BMC Bioinform.2003, 4, 2.

[17]. Liu, G.; Wong, L.; Chua, H.N, "Complex discovery from weighted PPI networks", .Bioinform. 2009, 25, 1891-1897.

[18]. Srihari, S.; Ning, K.; Leong, H.W. MCL-CAw: :A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure", BMC Bioinform.2010, 11, 504.

[19]. Nepusz, T.; Yu, H.; Paccanaro, A, "Detecting overlapping protein complexes in protein-protein interaction networks", Nat. Methods 2012, 9, 471-472 .

[20]. Chen, B.; Fan, W.; Liu, J.; Wu, F.X, "Identifying protein complexes and functional modules- from static PPI networks to dynamic PPI networks",Brief .Bioinform. 2014, 15, 177-194.

[21]. Wang, J.; Peng, X.; Li, M.; Pan, Y, "Construction and application of dynamic protein interaction network based on time course protein expression data", Proteomics 2013, 13, 301-312.

[22]. Park, Y.; Bader, J, "How networks change with time", Bioinformatics 2012, 28, 40-48.

[23]. Ou-Yang, L.; Dai, D.Q.; Li, X.L.; Wu, M.; Zhang, X.F.; Yang, P, "Detecting temporal protein complexes from dynamic Protein-Protein Interaction networks", BMC Bioinform. 2014, 15, 335- 348.

[24]. Li, M.; Chen, W.; Wang, J.; Wu, F.X.; Pan, Y, "Identifying dynamic protein complexes based on protein expression profiles and PPI networks", BioMed Res. Int. 2014.

[25]. Glas AM, Floore A, DelahayeLJ, WitteveenAT, PoverRC, BakxN, Lahti-Domenici JS, Bruinsma TJ, WarmoesMO, BernardsR, "Converting a breast cancer microarray signature into a high-throughput diagnostic test", BMC Genomics. 2006;7:1.

[26]. Lin JS., Lai EM. (2017) Protein–Protein Interactions: Co-Immunoprecipitation. In: Journet L., Cascales E. (eds) Bacterial Protein Secretion Systems. Methods in Molecular Biology, vol 1615. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-7033-9_17.