Under consideration for publication in Network Science

Symmetry in cancer networks identified : Proposal for multi-cancer biomarkers

Pramod Shinde
Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Indore 453552
(*e-mail*: pramodshinde119@gmail.com)

Loïc Marrec
Sorbonne Universit, CNRS, Laboratoire Jean Perrin (UMR 8237), Paris F-75005 France
(*e-mail*: loic.marrec@sorbonne-universite.fr)

Aparna Rai
Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Indore 453552
(*e-mail*: raiaparna13@gmail.com)

Rajesh Kumar
Discipline of Physics, Indian Institute of Technology Indore, Indore 453552
(*e-mail*: rajeshkumar@iiti.ac.in)

Alok Yadav
Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Indore 453552
(*e-mail*: physicistalok@gmail.com)

Alexey Zaikin
Department of Mathematics and Institute for Womens Health, University College London, London, WC1E 6BT, UK
(*e-mail*: alexey.zaikin@ucl.ac.uk)

Sarika Jalan*
Complex Systems Lab, Discipline of Physics, Indian Institute of Technology Indore, Indore 453552
Discipline of Biosciences and Biomedical Engineering, Indian Institute of Technology Indore, Indore 453552
and
Lobachevsky University, Gagarin avenue 23, Nizhny Novgorod, 603950, Russia.
(*e-mail*: sarikajalan9@gmail.com)

Abstract

One of the most challenging problems in biomedicine and genomics is the identification of disease biomarkers. In this study, proteomics data from seven major cancers were used to construct two weighted protein-protein interaction (PPI) networks *i.e.,* one for the normal and another for the cancer conditions. We developed rigorous, yet mathematically simple, methodology based on the degeneracy at -1 eigenvalues to identify network patterns or structural symmetry in network. Utilising eigenvectors corresponding to degenerate eigenvalues in the weighted adjacency matrix, we identified structural symmetry in underlying weighted PPI networks constructed using seven cancer data. Functional assessment of proteins forming these structural symmetry exhibited the property of cancer hallmarks. Survival analysis refined further this protein list proposing *BMI*, *MAPK11*, *DDIT4*, *CDKN2A*, and *FYN* as putative multi-cancer biomarkers. The combined framework of networks and spectral graph theory developed here can be applied to identify symmetrical patterns in other disease networks to predict proteins as potential disease biomarkers.
keywords: Cancer networks, Eigenvalue analysis, Graph symmetry, Biomarkers.

Contents

1 Introduction

Each cancer tissue comprises a heterogeneous and multi-factorial milieu that varies in cytology, physiology, signaling mechanisms, cell regulation, control mechanisms and response to therapy. Both the existence of genetic diversity among tumors of same cancer and the surprising amount of similarity among different cancers have been reported (Stratton et al., 2009). Interestingly, similarities among different cancers have widely been observed in cell proliferation rate, cell-cell interactions, metastatic potential and sensitivity to therapy. Therefore, the abundance of similarity in different cancers has allowed us to consider cancer as a single system in the present study. Furthermore, many biological processes can

be modeled as graphs composed of interactions among numerous cellular and molecular components (Barabási & Otavi, 2004; Shen R. et al., 2006). Probing a complex system in network or graph theory framework allows understanding a phenomenon or system's behavior by collecting information of all its constituents rather than focusing to a smaller part with apparent relation with a phenomenon. Network studies have been providing global understanding to corresponding biological processes and functional interactions (Xu et al., 2006; Folador et al., 2009; Shinde & Jalan, 2015; Shinde et al., 2018). Important outcomes were that different types of biological networks had exhibited network features such as complexity, robustness, hierarchical and scale-free behavior (Barabási & Otavi, 2004). Cancer network based studies helped to predict protein function, genotype-phenotype relationships between cancer proteins, a combined effect of DNA, RNA, protein modifications on overall cancer development and impact of mutations in altering molecular pathways (Yixuan et al., 2010; Lage et al., 2007). These cancer network studies have been successful in developing drug strategies as well as in finding important cancer pathways, e.g., mTOR signaling, p53 pathway, MAPK, and PI3K signaling pathways (Ahn et al., 2011; Chiang & Abraham, 2007). However, these investigations have focused mainly on structural positions of proteins or pathways in underlying networks.

Moreover, biological networks have been found to possess abundant symmetrical patterns (Wang et al., 2012). These symmetrical structures have been heavily investigated for their relevance of biological processes (Ocone & Sanguinetti, 2011; Cheng et al., 2016) and functional failure of such local structures can have substantial global impacts (Milo et al., 2004). For instance, a group of tumor suppressor genes forming onco-modules recently identified whereas oncogenic mutations in these modules altered the pan-cancer metabolic landscape (Cubuk et al., 2018). In this work, we focused on spectral (eigenvalues) properties of the network adjacency matrix for unraveling symmetrical patterns and corresponding proteins in the underlying network. Degeneracy in graph spectra has contributed significantly in our understanding of structural and dynamical properties of corresponding graphs (Van Mieghem, 2010). The driving force behind the investigation of origin and implication of degenerate eigenvalues is that the biological networks constructed using empirical data show very high degeneracy, particularly at 0 and -1 eigenvalues, than corresponding random networks (Shinde et al., 2015; Marrec & Jalan, 2017; Rai A. et al., 2018). **Indeed, these degenerate eigenvalues have been shown to exist due to an outcome of the complete and the partial node duplication (Yadav & Jalan, 2015) which is akin of the fundamental process in the evolution-related with gene duplication and diversification process (Shinde et al., 2015; Teichmann & Babu, 2004). As part of the cell cycle, particularly during replication of the genome, seldom another copy of a gene is synthesized. Immediately after this gene duplication event, both the original gene and the new identical copy of the gene have the same DNA sequence, so both interact with the same set of molecular partners. Consequently, as these genes are guided for their particular functions, each of the molecular partners that interacted with the ancestor gains a new interaction (Teichmann & Babu, 2004). Similarly in cancer genomes, clonal duplication and proliferation are achieved by DNA mutations (Furlong, 2013), mainly using somatic copy number alterations (Zack et al, 2013).** The gene duplication and diversification process play a crucial role in the growth, adaption, evolution, and subsistence of the biological system (Teichmann & Babu, 2004). Though

degeneracy at −1 eigenvalue can be related to specific structures in a network, the origin and implication of such structural patterns are not that obvious. Herein, we focused on symmetrical patterns corresponding to −1 degenerate eigenvalues and devised methodology to identify such essential network symmetrical structures.

In this work, we first provided a methodology to identify an origin and implications of eigenvalue degeneracy in weighted networks. Second, we applied this technique to find structural patterns corresponding to degenerate eigenvalue in weighted multi-cancer PPI network. Network structures linking to −1 degeneracy provided a framework for identification of proteins corresponding to underlying local patterns. The functional assessment further deduced that these proteins corresponding to −1 eigenvalue degeneracy have the property of cancer hallmarks. With survival analysis, we predicted cancer proteins *i.e., BMI*, *MAPK11*, *DDIT4*, *CDKN2A*, and *FYN* as putative multi-cancer proteins.

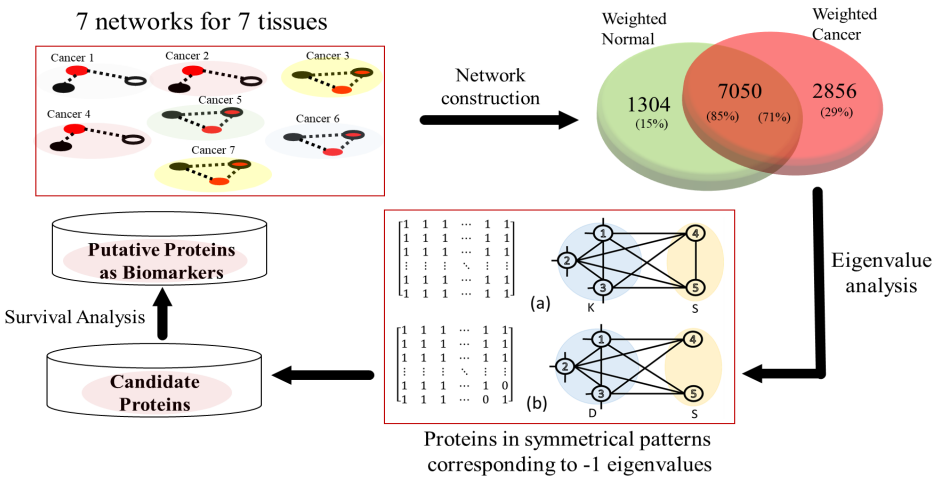


Fig. 1. **Work-flow diagram depicting network construction, eigenvalue analysis, and identification and characterisation of multi-cancer biomarkers.**

2 Data and methods

2.1 Dataset sources

We constituted our multi-cancer PPI network using proteomics data from morphologically seven different cancers such as Breast, Cervical, Colon, Lung, Oral, Ovarian and Prostate. **For straight-forward comparison of the proteome in a diseased state, we also retrieved PPIs in corresponding healthy tissue states.** We termed healthy tissues as 'normal' and cancer tissues as 'disease'. **In this way, we have 14 datasets** *viz,* **seven for healthy tissues and seven for disease tissues. PPIs in a healthy and the corresponding disease tissues were designated on the basis of their occurrence in the normal or the diseased tissue, respectively. PPI data mining was broadly divided into two steps,** *i.e.,* **(I) retrieval of protein names pertaining to a particular tissue, and then (II) retrieval of PPI of corresponding to proteins identified in step I.**

In the first step, protein-name data mining was independently performed on each tissue. Here, it is to be noted that we used the text-mining approach to map proteins for a particular dataset. Similarly, other approaches such as proteins corresponding to highly expressed genes can be utilized to map proteins. Also, it should be noted that our protein-name data mining was entirely based on the information available in secondary bioinformatics databases which are already curated and largely followed data sources *viz,* UniProtKB and GeneBank databases. **Protein-name data mining was performed using different search words, and accordingly, protein-names were destined to a particular dataset.** For example, if a protein entry in the UniProtKB database has been related to the information of oral cancer tissue, we marked that protein entry as a member of oral cancer dataset. The details of search words (queries) used for protein-name mining from these databases is given in Supplementary Materials (SM). Additionally, we explored other resources to enrich our protein-name collection. Swiss-2DPage (https: //world- 2dpage. expasy.org/swiss -2dpage/) and Cervical cancer database (CCDB) (http:// crdd.osdd. net/raghava /ccdb/) for cervical tissues, ACTREC Oral Cancer Database (http:// www. actrec. gov.in/ OCDB/ index.htm) and Head and Neck Oral Cancer Database (http:// gyanxet.com/ hno.htm) for oral cancer, ATCC cell line database (https: //www. atcc.org/) and Cancer Cell Line Encyclopedia (https://portals. broadinstitute. org/ccle) for all considered cancers. After diligent and enormous efforts of mining literature and database text, we collected the list of proteins in the healthy tissues and the corresponding cancer tissues from various literature and data archives. **In the second step,** once all the proteins for seven different tissues for the normal and disease states were collected, leading to fourteen datasets, the interacting partners of these proteins were retrieved from the STRING database version 9.189 (Szklarczyk et al., 2014). We used the default parameters in STRING database while retrieving PPI's. An interaction between a pair of proteins was considered if there exists a direct (*i.e.,* physical), indirect (*i.e.,* functional) or both relation between them.

In this way, we have seven datasets for the normal and seven datasets for the corresponding disease states. The detailed information of these fourteen datasets representing PPIs among all the fourteen tissues can be found at (FigShare).

2.2 Weighted multi-cancer PPI network construction

In a PPI network, vertices represent proteins and edges represent interactions between the proteins. We overlaid PPIs derived from seven tissues in two datasets *i.e.,* (1) normal and (2) disease, separately to construct two weighted PPI networks. Weights were assigned based on edge overlapping *viz,* the number of times an interaction is found in a set of cancers (schematic is provided in Fig. 4A). For instance, if an interaction between two nodes k and l found in colon and breast cancer, and was absent in other cancers that would yield a weight, $w_{kl} = 2$. Consequently, each element in the adjacency matrix has value ranging from 1 (min) to 7 (max). If an interaction existed in all seven cancers, the corresponding weight entry in the adjacency matrix would be 7 and if an interaction existed in only one cancer in the adjacency matrix, the weight entry would be 1. The weighted

adjacency matrix can be given as:

$$W_{\mathrm{ij}} = \begin{cases} w_{\mathrm{ij}} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where w_{ij} was the number of times i interacted with j. In such manner, we considered two interaction matrices, one for healthy state and another one for the disease state.

2.3 Various structural measures of a network

The most basic structural parameter of a network would be the degree of a node (k_i), which can be defined as a number of weighted edges connected to the node i ($k_i = \sum_j w_{ij}$). Further, the clustering coefficient (C) can be defined as a ratio of the number of interactions a neighbor of a particular node is having and the possible number of connections the neighbors can have among themselves. For a weighted network, C can be defined as the geometric average of the subgraph edge weights, $C_i = \frac{1}{k_i k_{i-1}} \sum_{j,k} (\hat{w_{ij}} \hat{w_{jk}} \hat{w_{ik}})^{\frac{1}{3}}$ (Saramki et al., 2007). The edge weights $\hat{w}$ were normalized by the maximum weight in the network $\hat{w} = \frac{w}{max(w)}$. Betweenness centrality (Brandes, 2008) of a node i defined as the sum of the fraction of all-pairs shortest paths that were passing through i, such that $\beta c(i) = \sum_{s,t \in V} \frac{\sigma(s,t|i)}{\sigma(s,t)}$ where V was the set of nodes, $\sigma(s,t)$ was the number of shortest (s,t)-paths, and $\sigma(s,t|i)$ was the number of those paths passing through some node i other than s,t.

2.4 Theoretical framework: relating structural symmetry and degenerate eigenvalues in weighted networks

We considered finite undirected and weighted graphs defined by $G = \{V,E\}$ with V is the node set, and E is the edge set such as $|V| = N$ and $|E| = N_c$. Eigenvalues $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$ were obtained by computing the roots of the characteristic polynomial of $\mathbf{W}$. Note that the eigenvalues were real because $\mathbf{W}$ was symmetric. The associated eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$,..., $\mathbf{v}_N$ satisfied the eigen-equation $\mathbf{W}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ with $i = 1,2,...,N$.

All the origin and implications of 0 eigenvalue degeneracy in networks spectra has been already well characterized (Yadav & Jalan, 2015). Briefly, the spectrum of a matrix of size N and rank r should encompass 0 eigenvalue with multiplicity $N - r$ (Cragg & Donald, 1997). There exist three conditions which would lead to the lowering of the rank of a matrix: (i) $R_i = (00\ldots0)$ a row with only zero-entries. (ii) $R_i = R_j$ at least two rows are equal. (iii) $\sum_i a_i R_i = \sum_j b_j R_j$ with $a_i, b_j \in \mathbb{R}$ two or more rows together are equal to some other rows. We would not consider the condition (i) which was related to the isolated nodes in $\mathbf{W}$. Additional information regarding 0 degeneracy can be found in SM.

As precribed in (Marrec & Jalan, 2017), it was possible to reduce the computation of x-eigenvalue of $\mathbf{W}$ matrix to the 0-eigenvalue of $(\mathbf{W} - x\mathbf{I})$ matrix. Now, lets understand the occurrence of x-eigenvalue degeneracy in weighted networks and see when should

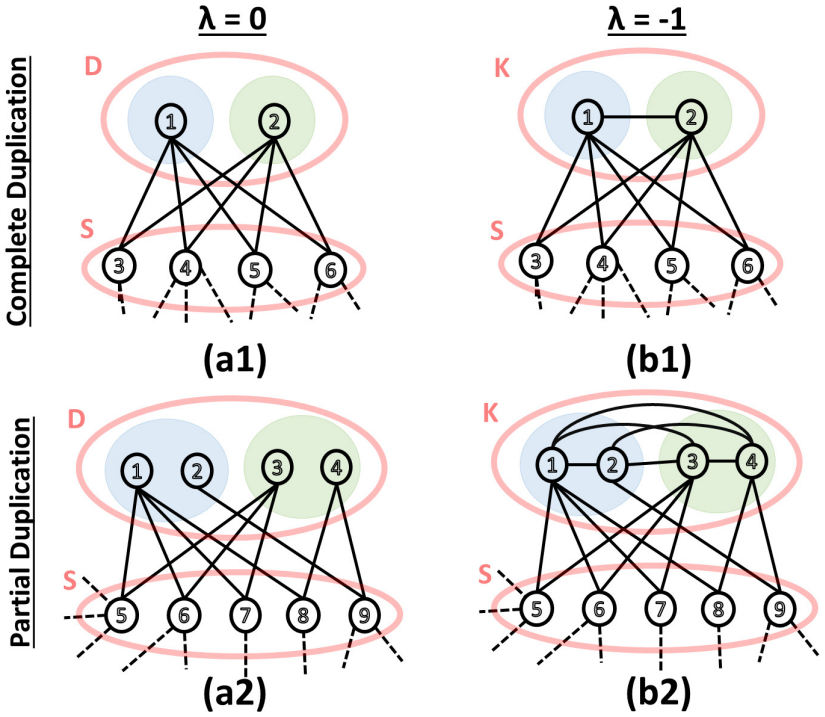


Fig. 2. **Structures leading to** 0 **and** -1 **eigenvalue degeneracy in weighted cancer PPI network.** (a) for $\lambda = 0$ and (b) for $\lambda = -1$. Note that to fulfil partial duplication condition may require more number of nodes. It can be that there is only one difference between the structures corresponding to 0 and -1 degeneracy. In the case of 0 degeneracy, there is no interaction between the nodes (*i.e.,* 1 and 2) whereas the interaction exists between them in the case of -1 degeneracy. Similar is true for partial duplication.

conditions (ii) and (iii) get fulfilled in $(\mathbf{W} - x\mathbf{I})$ which was written as follows:

$$\mathbf{W} - x\mathbf{I} = \begin{pmatrix} -x & w_{12} & \cdots & w_{1N} \\ w_{12} & -x & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1N} & w_{2N} & \cdots & -x \end{pmatrix} \tag{2}$$

Note that we considered self-loop less graphs. The condition (ii) can met if and only if: $w_{ik} = w_{jk}$ with $k = 1,2,...,N$. In the particular case $R_1 = R_2$, the previous equation becomes:

$$\begin{cases} w_{12} = -x \\ w_{1k} = w_{2k} \text{ with } k = 3,4,...,N \end{cases} \tag{3}$$

Specifically, the case of -1 eigenvalue can be related to $K * S$ structures (Figure 2). In these structures, all the nodes of K were interlinked with the same weight w_K. In addition, all the nodes of K are connected to the same set of neighbours, S having identical weight, w_{Si} (Figure 2 (c)). For $(\mathbf{W} - x\mathbf{I})$ matrix, we would get the relation $w_K = -x$. By this way, $D * S$ structure can be seen as a particular case of $K * S$ with $w_K = 0$ (Figure 2 (d)). This

has highlighted one of the most interesting aspects of degeneracy in weighted graphs. The condition (ii) would always give a $K*S$ sub-graph and the only difference was based on the weight of edges. More particularly, the weight of edges in K was directly related to the eigenvalue to which it contributes. However, because of this supplementary constraint, we would expect a lower degeneracy resulting from the condition (ii) in weighted networks as compared to unweighted networks. Indeed, it was sure that most of the $K*S$ structures observed in unweighted networks would not fulfill this constraint if the weights were taken into account.

Further, to simplify condition (iii), we considered the particular case $R_1+R_2=R_3+R_4$, which gave:

$$w_{1k}+w_{2k}=w_{3k}+w_{4k} \text{ with } k=1,2,...,N \tag{4}$$

The last equation can be developed as a system:

$$\begin{cases} w_{12}-x=w_{13}+w_{14}=w_{23}+w_{24} \\ w_{34}-x=w_{13}+w_{23}=w_{14}+w_{24} \\ w_{1k}+w_{2k}=w_{3k}+w_{4k} \text{ with } k=5,6,...,N \end{cases} \tag{5}$$

Contrary to the condition (ii), the condition (iii) did not shed light on a typical structure. Indeed, since w_{ij} can take any real value, the number of possible solutions were high and so it should be difficult to find a general solution to the previous equation. This was due to the fact that condition (iii) can result from a linear combination of rows. By this way, we would expect a degeneracy resulting from the condition (iii) at more eigenvalues in weighted graphs than in the case of unweighted graphs and so, contrary to the case of condition (ii), we would observe more different structures which were not brought in to light by degeneracy in unweighted networks. Here we would limit ourselves to provide an example of graph that could verify $R_1=R_2+R_3$ in W and $(\mathbf{W}+\mathbf{I})$ (see Figure 2 (d)).

So far, we focused on finding structures behind occurrence of eigenvalue degeneracy. The next question was: *could we identify the nodes involved in such structures ?* The answer was yes since it had been shown that it's possible by using the eigenvectors associated to degenerate eigenvalues (Marrec & Jalan, 2017). More particularly, the components of these eigenvectors verify the following relation:

$$\begin{cases} \sum_{i\in K_p} v_i=0 \text{ with } v_i\neq 0 \text{ and } p=1,2,...,n_{K*S} \\ v_{j\in V\setminus\{K_1\cup K_2\cup...\cup K_{n_{K*S}}\}}=0 \end{cases} \tag{6}$$

for nodes belonging to $K*S$ structures, where n_{K*S} denoted the number of such sub-graphs in the whole network.

Similarly, for the nodes which has belonging to a sub-graph verifying the condition (iii) in $\mathbf{W}-x\mathbf{I}$, one has the relation:

$$\begin{cases} \sum_{i\in(L.C)_p} v_i=0 \text{ with } v_i\neq 0 \text{ and } p=1,2,...,n_{L.C} \\ v_{j\in V\setminus\{(L.C)_1\cup(L.C)_2\cup...\cup(L.C)_{n_{L.C}}\}}=0 \end{cases} \tag{7}$$

where $L.C$ and $n_{L.C}$ were the linear combinations and the number of linear combinations, respectively. Thanks to these relations, one could identify easily the nodes contributing to degenerate eigenvalue. One way to handle this issue was to consider the matrix $(\mathbf{W}-x\mathbf{I})$ and to search for each $R_i=R_j$, which was computationally doable. Then, we could consider

one of the eigenvectors associated to x eigenvalue and identified all the non-null entries. These should not obey $R_i = R_j$ to belong necessarily to a linear combination.

2.5 Gene enrichment and survival analysis

We used genes from significant signatures *i.e,* corresponding to -1 eigenvalue degeneracy as an input into STRING (Szklarczyk et al., 2014), Panther (Mi et al., 2005), and MSigDB (Liberzon et al., 2011) gene ontology platforms. Further, we measured the correlation between each gene activity and patient survival outcomes using Cox proportional risks group hazards models available with SurvExpress biomarker validation tool (Aguirre-Gamboa et al., 2013) for TCGA cancer gene expression data (SM).

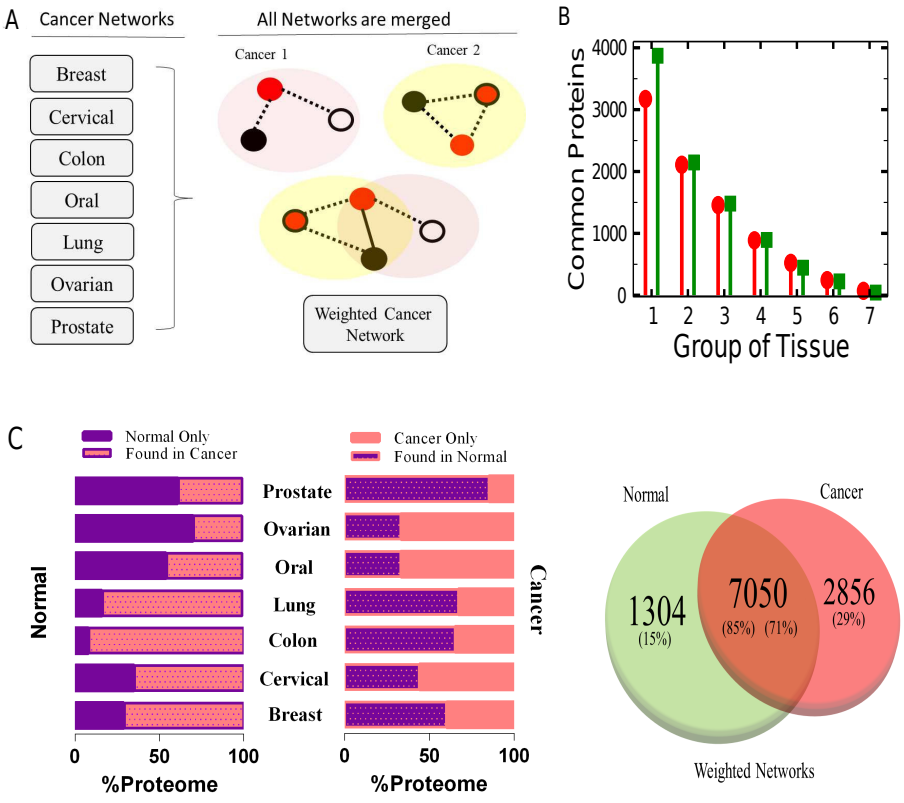


Fig. 3. **Construction and analysis of weighted multi-cancer PPI network.** (A) Schematic diagram illustrates the construction of weighted network where edges from two cancer (in actual seven) networks constitute one weighted network. Dotted lines are unweighted edges (having weight 1), and the solid line is weighted edge (weight 1 to 7). (B) shows the number of times a protein can be found in a particular group of tissue. For instance, if a protein is present in breast and oral cancer, it is said to present in a group of two tissues. (C) Venn diagram shows the number of proteins found in both normal and cancer tissues.

Table 1. **Properties of (un-) weighted normal and disease PPI networks.** *Here N, N_C, $\langle K \rangle$, k_{max}, λ_{-1}, $\lambda_{-1}^{nodes}(ii)$ and $\lambda_{-1}^{nodes}(iii)$ represent number of nodes, connections, average degree, degree of the hub node, number of minus one eigenvalues, number of proteins linked to degeneracy corresponding to λ_{-1} with condition ii and iii, respectively.* ^{w} *stands for weighted network and* ^{unw} *stands for unweighted network.*

Network	N	N_c	$\langle k^{unw} \rangle$	$\langle k^{w} \rangle$	k_{max}^{unw}	k_{max}^{w}	λ_{-1}	$\lambda_{-1}^{nodes}(ii)$	$\lambda_{-1}^{nodes}(iii)$
Normal	9946	105491	21	32	636	1565	23	39	0
Disease	8354	102701	25	35	877	1409	20	35	4

3 Results

3.1 Analysis of normal and cancer datasets

Before we would present results based on the analysis of weighted networks, we outlined few observations about number of proteins in the healthy and cancer tissues by considering all the cancers as a single unit. Existence of common proteins in both the normal and corresponding cancer states have suggested that their common aetiology and common functions which are essential for cell survival and growth. We found that more than 65% ($\pm$ 23%) proteins in the normal tissues were found in corresponding cancer tissues when we considered each tissue separately (Figure 3(C)). However, there were as much as 85% of proteins in normal tissues were found in cancer tissues when we took all the normal tissues as a single unit (Figure 3(D)). It suggested that a large portion of proteome of healthy tissues have role in some or other cancer related activities. Similarly, more than 51% ($\pm$16%) proteins from individual cancer tissues were found in normal tissues when we considered each tissue separately (Figure 3(C)). Interestingly, there were as much as 71% of cancer proteins were also present in normal tissues when we considered all cancer tissues as a single unit (Figure 3(D)). It would be institutive to have a higher proteome overlap when different tissues were considered as one unit but it was interesting to note that cumulative cancer tissue proteome has less overlap than cumulative normal tissue proteome.

3.2 Importance of nodes based on structural properties of networks

We constructed two types of networks for both the disease and the normal datasets: (1) unweighted networks which was constructed based on the presence and the absence of interactions between proteins, and (2) weighted networks where weights were assigned to an interaction based on the number of times an interaction was repeated in the combined list (Figure 3(A)). First, we examined the structural properties of these networks. We found that $\langle k \rangle$ was higher in the disease networks than that of normal networks (for both the weighted and the unweighted cases) suggesting that cancer proteins have more affinity to interact among themselves. Further, we found that the highest degree nodes (k_{max}) in the unweighted and the weighted multi-cancer networks were different (SM). The hub protein

in unweighted cancer network was *UBC* ($k = 877$) whose pathway function is translation regulation whereas the hub protein in weighted multi-cancer network was *CACNB2* ($k = 1565$). Additionally, the top 10 degree proteins in weighted multi-cancer network were also among pathway regulators (SM). This observation lie in accordance of known fact that the regulatory proteins were high degree proteins in PPI networks (Fox et al., 2011). Second, weighted multi-cancer network has *CACNB2* and *BRD7* (k = 1524) as two high degree proteins in which *CACNB2* has the role among CCR5 pathway in macrophages and PEDF induced signaling (http://www.proteinatlas.org/ENSG00000165995-CACNB2/cancer) and *BRD7* has TP53 activity (Yu et al., 2016). It was interesting to note that though *CACNB2* and *BRD7* perform essential cancer activities (Yu et al., 2016), they are yet to get thoroughly investigated for drug related activities in cancer. Nevertheless, weighting scheme have provided identification of another set of nodes which were vital for cellular processes in cancers under investigation.

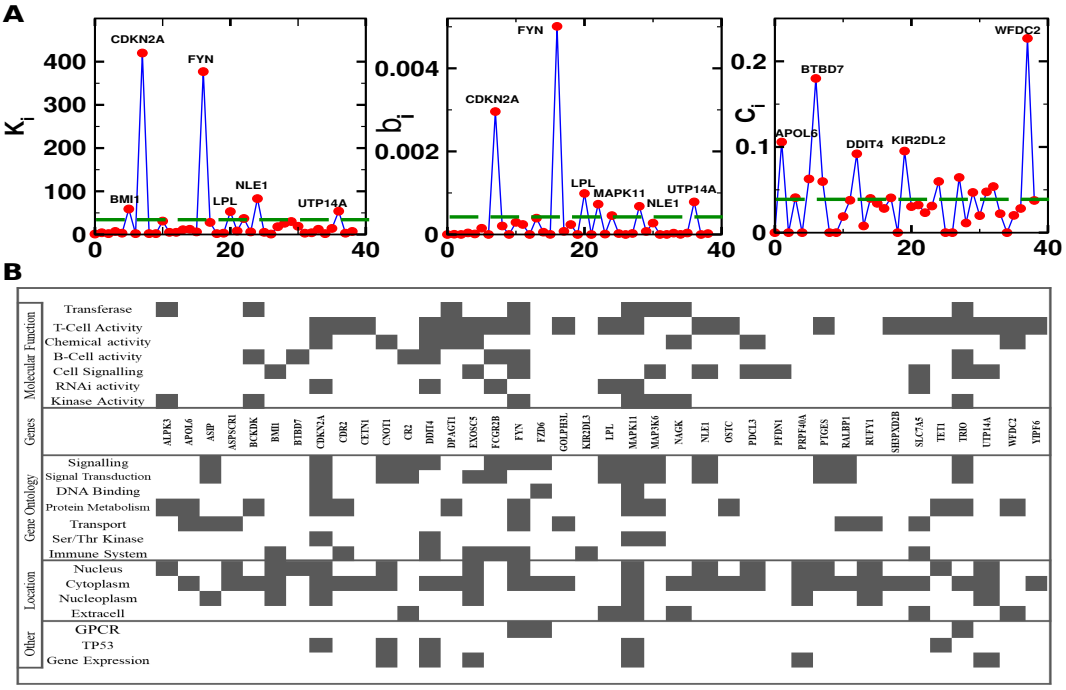


Fig. 4. **Network structural properties and functional assessment of 39 cancer proteins identified using network symmetry.** (A) Network structural properties such as degree (k), betweeness centrality (β_C) and clustering coefficient (C) for 39 proteins are displayed where proteins showing high value of particular property are highlighted. Horizontal line (green) shows the average value.(B) Gene functional assessment is categories into four groups *i.e.,* molecular function, gene ontology, location and other.

3.3 Importance of nodes corresponding to degeneracy in weighted cancer network

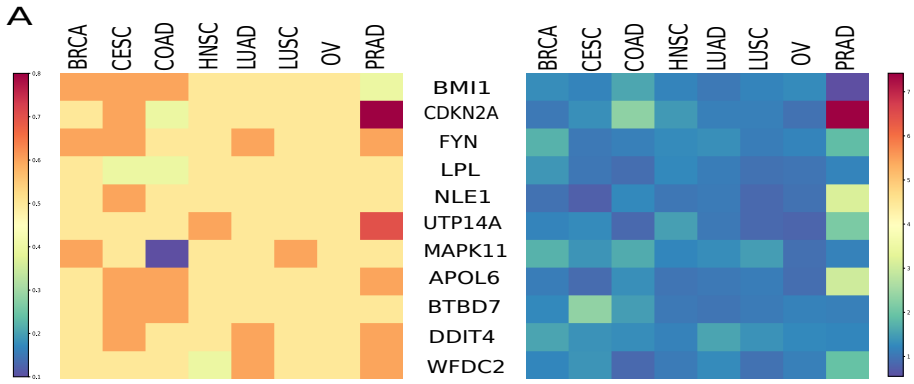
Next, we discussed to the prime focus of the current study by screening nodes forming structural patterns corresponding to -1 eigenvalues (λ_{-1}) in weighted multi-cancer PPI net-

work and further noted down their biological significances and network properties. There were 39 proteins corresponding to -1 eigenvalues (summarized in SM Table 1). These proteins, except *APOL6* and *BTBD7*, have been reported to be related to more than one tumors. Each of 39 proteins posses one or more property of cancer's hallmarks (Hanahan & Weinberg, 2011) as they have participated in cell signaling, signal transduction, transport etc (Figure 4(B)). Additionally, they exhibited essential bio-physical and bio-chemical activities such as enzymatic (kinase, transferase), immunological (B-cell, T-cell) and molecular (RNAi, signalling) activities. Few of them were also found to perform activities at multiple cellular locations such as cell nucleus, cytoplasm, nucleoplasm and extracellular matrices which was biologically more relevant in performing specific biological activities related to cellular communications (Figure 4(B)). It was interesting to report that these 39 proteins did not take any significant structural position in global-level weighted multi-cancer PPI network. Therefore, they were not detectable at global-level network using various measures such as node degree, clustering coefficient and betweenness centrality (Figure 4 and SM) available in network literature to identify them as structurally important nodes.

Since it was known that disease biomarkers would tend to have higher degree and connectivity in comparison to non-disease genes because of higher values of gene expressions (Winter et al., 2014), among these 39 proteins we first focused on those proteins which have degree higher than average degree of 39 nodes. Second, we focused on proteins having β_c higher than $\langle\beta_c\rangle$ of 39 proteins. It's known that high degree nodes have high β_c value. However, there were interesting reports where moderate degree nodes have high β_c and these nodes were proposed to be selected as effective cancer targets (Barh et al., 2014). Lastly, we noted down proteins with higher C values than $\langle C\rangle$ of 39 proteins. Clustering coefficient demonstrates cluster forming ability of nodes or how well a node is connected among its direct neighbours (Albert & Barabási, 2002). Interestingly, above five nodes were among top 10 nodes with high C value in our weighted multi-cancer PPI network. Overall, the short-listed 12 candidate proteins showed reasonably high k, C and β_c in the pathway network and selecting them for drug targeting would be reasonable.

3.4 Survival Analysis

It's essential to study function of a protein with its role in patient survival to devise cancer biomarker (Brockmoller et al., 2011). To achieve this, we assessed whether the selected 12 candidate proteins were also associated with the overall survival (OS) in different cancers. First, we performed OS with multi-protein (12 proteins) to understand the role of degeneracy in eigenvalues which arises due to underlying symmetry in interaction in each cancer. Risk hazard ratio (HR) was found be more than 1.5 for each cohort in which Cervical squamous cell carcinoma (CESC) and Prostate adenocarcinoma (PRAD) displayed significantly high value (Figure 5). Secondly, we carried out the single-protein analysis which identified most significantly associated proteins with OS in each cancer, independently. We identified five proteins as putative multi-cancer biomarkers *i.e., BMI*, *MAPK11*, *DDIT4*, *CDKN2A*, and *FYN*. These proteins have HR value more than average HR value for atleast 3 cohorts as well as they occurred in atleast 3 cohorts in multi-protein analysis.



B

Cohort	HR [95% CI; p-value]	Cox interesting proteins	C-Index
BRCA	2.1 [1 - 2.1; 0.04]	BMI1, LPL, MAPK11, DDIT4	0.6
CESC	4.8 [2.3 - 10.1; 2.888e-05]	LPL, BTBD7, WFDC2	0.7
COAD	2.3 [1.4 - 3.6; 0.001]	CDKN2A, NLE1	0.7
HNSC	1.9 [1.4 - 2.5; 4.915e-06]	CDKN2A, FYN, LPL, UTP14A, MAPK11	0.6
LUAD	1.8 [1.3 - 2.4; 0.0002]	FYN, MAPK11, BTBD7, DDIT4	0.6
LUSC	1.6 [1.1 - 2.5; 0.03]	BMI1	0.6
OV	1.8 [1.3 - 2.5; 0.0004]	BMI1, DDIT4	0.6
PRAD	6.7 [0.8 - 53.2; 0.07]	CDKN2A, UTP14A	0.7

Fig. 5. **Comparison of biomarkers using overall survival analysis.** The listed 12 proteins are distinguished proteins among 39 proteins. (A) Single protein analysis. C-index and HR values of 12 candidate proteins in cancer cohorts are shown. It is to note that we select two cancer cohorts for lung cancer *i.e.,* LUAD and LUSC. (B) Multi-protein analysis where group of 12 candidate proteins are analysed against each cancer cohort.

In particular, we identified *BMI1* which was epigenetic regulator and it promoted oncogenesis with DNA damage response (Nacerddine et al., 2012). Our OS analysis found that *BMI1* has increased levels of gene expression in high-risk groups in seven cohorts (expect in HNSC) (SM). Interestingly, we found that *CDKN2A* was absent in Breast cancer adenocarcinoma (BRCA) cohort and it has decreased expression in high risk patients (except Colon adenocarcinoma (COAD) and PRAD). We also found that *CDKN2A* showed significant HR value in COAD (HR 2.3 [95% 1.2 - 4.4), p=0.01]) and PRAD (HR 7.4 [95% 0.9 - 59.1, p=0.06]) cohorts. Further, we found the increased level of *MAPK11* expression in high-risk patients of BRCA (SM) which was supported by the fact that *MAPK11* was highly expressed in the metastatic breast cancer (He et al., 2014). In a way, it was possible to correlate increased *MAPK11* expression in high risks patients with high HR value. Interestingly, we found that *DDIT4* has high HR values in all fast growing cancers (SM). *DDIT4* is considered to be a driver in the aggressiveness of cancer cells because of its apoptotic activity. *DDIT4* is induced by a variety of stress conditions and inhibit mTORC1 pathway (Pinto et al., 2017). *FYN* is known to be up-regulated in human prostate cancer and has role in cancer progression and metastasis (Elias & Ditzel, 2015). Overall, our survival analysis predicted *i.e., BMI*, *MAPK11*, *DDIT4*, *CDKN2A*, and *FYN* as putative multi-cancer proteins which could effectively stratify low and high-risk cancer patients.

4 Discussion

Our analysis focused on weighted PPI network constructed based on the number of times a particular interaction among couple of proteins present in seven most prevalent and morphologically different cancers. Most of the earlier works were typically node-centric where as we adopted a holistic approach excessively exploiting significance of functional interactions among different cancer tissues. In a way, our method provided an ultimate scope for identification of a protein set that would not have over-represented otherwise. The observation of cumulative cancer tissue has lesser protein overlap than normal tissues suggested the possibility of diverse cancer related activities within and across cancer tissues. We also found that hub proteins in unweighted and weighted networks were completely different. In unweighted network, *CACN2* and *BRD7* were turned out as two high degree proteins in which both of them were largely unexplored for their therapeutic use. In other case, *UBC* was identified as hub protein in unweighted network which was very well known for cancer related activities. All this implicated the significance of weighing scheme to identify another set of nodes which were vital.

Significance of this approach laid in the identification of simple and precise, yet fundamental, symmetrical structures of underlying network through -1 eigenvalues. Biological network motifs drive very specific functions depending on the needs of the cell. Though many efforts have been devoted to identify network motifs to capture particular local functionality within a biological network, still scope persist for efficient method development. Our method identified symmetrical structures in the underlying weighted PPI networks and picked up proteins forming these essential network structures as candidate proteins. These symmetrical structures were based on degeneracy in -1 eigenvalues. In general, degeneracy in cancer can be understood in terms of independent adaptation of each cancer gene arising due to natural selection (Hanahan & Weinberg, 2011). In present context, we used network structures corresponding to -1 eigenvalues as a measure of degeneracy in network graph.

Symmetrical structures presented here depict groups of proteins having a topological equivalence in a network. -1 eigenvalue degeneracy (both case ii and iii) essentially detects pairs of nodes which are not only connected to one another forming network motifs, but also connected to exactly the same other nodes. Case (ii) is straight-forward, representing the nodes with the exact node duplication (Figure 2(c)). Here, corresponding nodes of -1 degeneracy are grouped together in set K and their interaction partners, other than in set K, are grouped in set S. Nodes in set K are all-to-all connected and form a complete subgraph motif in both the case (ii) and (iii). In case (ii), all the nodes in K set interacts with all the nodes in set S. Case (iii) is not straight-forward, representing the nodes with the partial node duplication (Figure 2(d)). Here, nodes in K set are grouped in two (or more) subsets and these two (or more) subsets ineract with all nodes in set S, independetly. One specific example of case (iii) is showed in (Figure 2(d)).

This structural phenomenon is very interesting since many proteins are essentially backups for others, and can perform similar functions if one is knocked out or not functional at a particular phase of the cell cycle. For example, when a eukaryotic cell is exposed to ionizing radiation, a group of RAD52 proteins attends as a backup pathway operating independently in place of DNA dependent protein kinase (Perrault et al., 2004). In another

example, the presence of two distinct pathways of glycoproteins and non-glycoproteins exist in mammalian cells for translocation of misfolded proteins from the endoplasm reticulum (ER) to the cytosol (Ushioda et al., 2013). First one is functional in non-stress condition and later is functional in ER stress (Ushioda et al., 2013). One more interesting example is that a significant number of cancerous mutations found to fall at structurally equivalent positions within the proteins catalytic core, particularly in kinases (Dixit et al., 2009). These structurally equivalent positions are also termed as mutational hotspots (Dixit et al., 2009).

The identified 39 proteins corresponding to patterns linked to -1 eigenvalue degeneracy did not take any significant structural position in weighted multi-cancer PPI network and hence they were not detectable using various measures such as node degree, clustering coefficient and betweenness centrality. However, these proteins should have profound effects on information processing in the protein-protein interactions since their position in a network arised due to underlying symmetry among interactions. In addition, the list of 39 proteins showed important biological roles given by gene enrichment analysis. Essentially, because these 39 proteins were not hub proteins, their removal would have little impact on the overall statistics of the network which was essential to rid of false positive outcomes.

Further, we short-listed 12 significant candidate proteins by refining the list of 39 proteins with respect to network properties. These 12 proteins showed reasonably high k, C and β_c values and selecting them for drug targeting would be reasonable. Finally, we convinced with five putative proteins which displayed high HR values in both single- and multi-protein analysis. Also, these five proteins displayed very specific roles in group of cancers in survival analysis. The current study demonstrated that the spectral graph theory framework is a powerful concept and tool for revealing important structural patterns in network. Utilizing networks, cancer biomarkers were identified considering their stands in pathways and cycles instead of mere higher values of network features alone.

5 Conclusion

The current study was focused on the importance of interactions between proteins participating among various cancer tissues. Two main objectives were currently pursued: first, the glance at functional interactions among all cancers as single unit, which permitted us to look at all cancer related processes under one data framework; and second, the use of network theory and spectral graph theory as a means to identify important causative agents for multi-cancer diagnosis and therapy.

Overall, the systems biology and spectral graph theory approach that we adopted in this study allowed us to identify putative proteins those can be termed as multi-cancer biomarkers. In which, some proteins were already known to serve as candidate multi-cancer biomarkers that have confirmed the reliability of our results. Our study has broadened the approach to identify cancer biomarkers using patterns corresponding to -1 eigenvalue degeneracy. The selected five proteins *viz., BMI*, *MAPK11*, *DDIT4*, *CDKN2A*, and *FYN* showed both biological significance and effectiveness in survival analysis. The identification of multi-cancer biomarkers may lead to proposals of novel diagnostic tools and therapeutic schemes. This finding could lead to another predictive angle and biological validation in the future. Furthermore, on technical ground, the article has presented a

method to detect symmetrical patterns in weighted networks. The technique can be used to detect these patterns in any networks generated from other real-world data.

Additional files

File of supplementary material comprises of Supplementary information, figures and tables. It lists information on Hallmarks of Cancer. It also includes table information on (Table S1) Datasets used for survival analysis. All datasets are considered for TCGA database, (Table S2) Datasets of seven cancers and their details, (Table S3) Top 10 degree nodes in weighted multi-cancer PPI network, (Table S4) Gene Expressions by risk groups, (Table S5) Biological functions of proteins.

Availability of data and materials

All data generated or analysed during this study are included in this article and its supplementary information files. The software used in this paper to detect symmetrical patterns based on degenerate eigenvalues is available for download at https://github.com/pramodsshinde/ networkSymmetry. All the codes were written in Matlab.

Acknowledgments

SJ thanks the support by grant of the ministry of education and science of the Russian Federation (Agreement No. 074-02-2018-330) and Department of Science and Technology (DST), Government of India (EMR/2014/000368). P.S. thanks DST for the INSPIRE fellowship (IF150200). A.Y. thanks Council of Scientific and Industrial Research, Government of India for the fellowship. We acknowledge Dr. Hem Chandra Jha for interesting discussions. Authors thank Complex Systems Lab members for timely help and fruitful discussions.

Conflicts of interest

Nothing to disclose.

Abbreviations

PPI : Protein-protein interaction; OS: Overall survival; HR : Hazard ratio; BRCA : Breast cancer adenocarcinoma; CESC : Cervical squamous cell carcinoma; COAD : Colon adenocarcinoma; HNSC: Head and neck squamous cell carcinoma; LOAD : Lung adenocarcinoma; LUSC : Lung squamous cell carcinoma; OV : Ovarian serous cystadenocarcinoma; PRAD : Prostate adenocarcinoma

References

Aguirre-Gamboa R., Gomez-Rueda H., Martnez-Ledesma E., Martnez-Torteya A., Chacolla-Huaringa R., Rodriguez-Barrientos A., Tamez-Pena, J. G., & Trevino V. (2013). SurvExpress:

an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PloS One,* **8(9)**, 74250.

Ahn J., Yoon Y., Park C., Shin E., & Park S. (2011). Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, **27(13)**, 1846-1853.

Albert R., & Barabási AL. (2002). Statistical mechanics of complex networks. *Review of Modern Physics,* **74(1)**, 47.

Barabasi A. L., & Otavi Z. N. (2004). Network biology: understanding the cells functional organization. *Nature Review Genetics,* **5**, 101113.

Barh D., Carpi A., Verma M., & Gunduz M. (2014). Cancer biomarkers: minimal and noninvasive early diagnosis and prognosis. *CRC Press*.

Brandes U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks,* **30(2)**, 136-145.

Brockmoller S. F., Bucher E., Muller B. M., Budczies J., Hilvo M., Griffin J. L., Oresic M., Kallioniemi O., Iljin K., Loibl S., & Darb-Esfahani S. (2011). Integration of Metabolomics and Expression of Glycerol-3-phosphate Acyltransferase (GPAM) in Breast Cancer Link to Patient Survival, Hormone Receptor Status, and Metabolic Profiling. *Journal of Proteome Research,* **11(2)**, 850-860.

Cubuk C., Hidalgo M. R., Amadoz A., Pujana M. A., Mateo F., Herranz C., Carbonell-Caballero J., & Dopazo J. (2018). Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. *Cancer research,* **78(21)**, 6059-72.

Cheng F., Chuang L., Bairong S., & Zhongming Z. (2016). Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. *BMC Systems Biology,* **10(3)**, 65.

Chiang G. G., & Abraham R. T. (2007). Targeting the mTOR signaling network in cancer. *Trends in Molecular Medicine,* **13(10)**, 433-442.

Chung F., Linyuan L., & Van V. (2003). Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences,* **100**, 63136318.

Cragg J. G., & Donald S. G. (1997). Inferring the rank of a matrix. *Journal of Economics,* **76(1)**, 223-250.

Dixit A., Yi L., Gowthaman R., Torkamani A., Schork N.J. & Verkhivker G.M. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PloS one,* **4(10)**, e7485.

Elias D., & Ditzel H. J. (2015). Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacological Research,* **100**, 250-254.

Figshare data: https://dx:doi.org/10.6084/m9.figshare.4193409.v1

Folador E. L., de Carvalho P. V., Silva W. M., Ferreira R. S., Silva A., Gromiha M., Ghosh P., Barh D., Azevedo V., & Rttger R. (2016). In silico identification of essential proteins in Corynebacterium pseudotuberculosis based on protein-protein interaction networks. *BMC Systems Biology,* **10(1)**, 103.

Fox A. D., Hescott B. J., Blumer A. C., & Slonim D. K. (2011). Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics,* **27(8)**, 1135-1142.

Furlong L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, **29(3)**, 150-159.

Hanahan D. & Weinberg R. A. (2011). Hallmarks of cancer: the next generation. *Cell,* **144(5)**, 646-674.

He Z., He J., Liu Z., Xu J., Sofia F. Y., Liu H., & Yang J. (2014). MAPK11 in breast cancer cells enhances osteoclastogenesis and bone resorption. *Biochimie,* **106**, 24-32.

Jalan S., Ung C.Y., Bhojwani J., Li B., Zhang L., Lan S. H., & Gong Z. (2012). Spectral analysis of gene co-expression network of Zebrafish. *EPL,* **99(4)**, 48004.

Kim M. S., Kim J. R., Kim D., Lander A. D., & Cho K. H. (2012). Spatiotemporal network motif reveals the biological traits of developmental gene regulatory networks in Drosophila melanogaster. *BMC Systems Biology,* **6(1)**, 31.

Krause A. E., Frank K. A., Mason D. M., Ulanowicz R. E., & Taylor W. W. (2003). Compartments revealed in food-web structure. *Nature*, **426(6964)**, 282-285.

Lage K., Karlberg E. O., Strling Z. M., Olason P. I., Pedersen A. G., Rigina O., Hinsby A. M., Tmer Z., Pociot F., Tommerup N., & Moreau Y. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, **25(3)**, 309.

Liberzon A., Subramanian A., Pinchback R., Thorvaldsdttir H., Tamayo P., & Mesirov J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics,* **27(12)**, 1739-1740.

Marrec L. & Jalan S. (2017). Analysing degeneracies in networks spectra. *EPL*, **117(4)**, 48001.

Mi H., Lazareva-Ulitsky B., Loo R., Kejariwal A., Vandergriff J., Rabkin S., Guo N., Muruganujan A., Doremieux O., Campbell M. J., & Kitano H. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research,* **33**, D284-D288.

Milo R., Itzkovitz S., Kashtan N., Levitt R., Shen-Orr S., Ayzenshtat I., Sheffer M., & Alon U. (2004). Superfamilies of evolved and designed networks. *Science,* **303(5663)**, 1538-1542.

Nacerddine K., Beaudry J. B., Ginjala V., Westerman B., Mattiroli F., Song J. Y., Van Der Poel H., Ponz O. B., Pritchard C., Cornelissen-Steijger P., & Zevenhoven J. (2012). Akt-mediated phosphorylation of Bmi1 modulates its oncogenic potential, E3 ligase activity, and DNA damage repair activity in mouse prostate cancer. *Journal of Clinical Investigation,* **122(5)**, 1920.

Ocone A. & Sanguinetti G. (2011). Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, **27(20)**, 2873-2879.

Perrault R., Wang H., Wang M., Rosidi B. & Iliakis, G. (2004). Backup pathways of NHEJ are suppressed by DNAPK. *Journal of cellular biochemistry*, **92(4)**, 781-794.

Pinto J. A., Rolfo C., Raez L. E., Prado A., Araujo J. M., Bravo L., Fajardo W., Morante Z. D., Aguilar A., Neciosup S. P., & Mas L. A. (2017). In silico evaluation of DNA Damage Inducible Transcript 4 gene (DDIT4) as prognostic biomarker in several malignancies. *Scientific Reports,* **7**.

Potestio R., Caccioli F., & Vivo P. (2009). Random matrix approach to collective behavior and bulk universality in protein dynamics. *Physical Review Letters,* **103(26)**, 268101.

Rai A., Menon A. V., & Jalan S. (2014). Randomness and preserved patterns in cancer network. *Scientific Reports,* **4**, 6368.

Rai A., Pawar A. K., & Jalan S. (2015). Prognostic interaction patterns in diabetes mellitus II: A random-matrix-theory relation. *Physical Review E,* **92(2)**, 022806.

Rai A., Shinde P. & Jalan S. (2018). Network spectra for drug-target identification in complex diseases: new guns against old foes. *Applied Network Science*, **3(1)**, 51.

Saramki J., Kivel M., Onnela J. P., Kaski K., & Kertsz J. (2007). Generalizations of the clustering coefficient to weighted complex networks, *Physical Review E,* **75**, 027105.

Shen R., Goonesekere N. C., & Guda C. (2012). Mining functional subgraphs from cancer protein-protein interaction networks. *BMC Systems Biology,* **6(3)**, S2.

Shinde P., Yadav A., Rai A., & Jalan S. (2015). Dissortativity and duplications in oral cancer. *European Physical Journal B*, **88(8)**, 197.

Shinde P., & Jalan S. (2015). A multilayer protein-protein interaction network analysis of different life stages in Caenorhabditis elegans. *EPL*, **112(5)**, 58001.

Shinde P., Sarkar C., & Jalan S. (2018). Codon based co-occurrence network motifs in human mitochondria. *Scientific reports*, **8(1)**, 3060.

Stratton M. R., Campbell P. J., & Futreal P.A. (2009). The cancer genome. *Nature,* **458**, 719-724.

Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K. P., & Kuhn M. (2014). STRING v10: proteinprotein interaction networks, integrated over the tree of life. *Nucleic Acids Research,* **43**, D447-D452.

Teichmann S. A., & Babu M. M. (2004). Gene regulatory network growth by duplication. *Nature genetics*, **36(5)**, 492.

Ushioda R., Hoseki J. & Nagata K. (2013). Glycosylation-independent ERAD pathway serves as a backup system under ER stress. *Molecular biology of the cell*, **24(20)**, 3155-3163.

Van Mieghem P. (2010). Graph spectra for complex networks. *Cambridge University Press*.

Wang J., Huang Y., Wu F. X., & Pan Y. (2012). Symmetry compression method for discovering network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* **9(6)**, 1776-1789.

Winter D. L., Erce M. A., & Wilkins M. R. (2014). A web of possibilities: network-based discovery of protein interaction codes. *Journal of Proteome Research,* **13(12)**, 5333-5338.

Yadav A. & Jalan S. (2015). Origin and implications of zero degeneracy in networks spectra. *Chaos: An Interdisciplinary Journal of Nonlinear Science,* **25(4)**, 043110.

Yixuan Y., Kiat L. S., Yee C. L., Huiyin L., Yunhao C., Kuan C. P., Hassan A., Ting W. T., Manuel S. T., Guan Y. K., & Pin L. Y. (2010). Cathepsin S mediates gastric cancer cell migration and invasion via a putative network of metastasis-associated proteins. *Journal of Proteome Research,* **9(9)**, 4767-4778.

Yu X., Li Z., & Shen J. (2016). BRD7: a novel tumor suppressor gene in different cancers. *American Journal of Translational Research,* **8(2)**, 742.

Zack T. I., Schumacher S. E., Carter S. L., Cherniack A. D., Saksena G., Tabak B., Michael S., & Sougnez C. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, **45(10)**, 1134.