Topologically significant directed random walk with applied walker network in cancer environment

Choon Sen Seah[1], Shahreen Kasim[1], Rd Rohmat Saedudin[3], Mohd Farhan Md Fudzee[1], Mohd Saberi Mohamad[4], Rohayanti Hassan[5] and Mohd Arfian Ismail[6]

[1]Soft Computing and Data Mining Centre, Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Johor, Malaysia
[2]Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Karung Berkunci, Bachok, Kelantan, Malaysia
[3]School of Industrial Engineering, Telkom University, Bandung, West Java, Indonesia
[4]Faculty of Bioengineering and Technology, Universiti Malaysia Kelantan, Jeli Campus, Lock Bag, Jeli, Kelantan, Malaysia
[5]Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
[6]Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Pahang, Malaysia

Abstract: Numerous cancer studies have combined different datasets for the prognosis of patients. This study incorporated four networks for significant directed random walk (sDRW) to predict cancerous genes and risk pathways. The study investigated the feasibility of cancer prediction via different networks. In this study, multiple micro array data were analysed and used in the experiment. Six gene expression datasets were applied in four networks to study the effectiveness of the networks in sDRW in terms of cancer prediction. The experimental results showed that one of the proposed networks is outstanding compared to other networks. The network is then proposed to be implemented in sDRW as a walker network. This study provides a foundation for further studies and research on other networks. We hope these finding will improve the prognostic methods of cancer patients.

Keywords: Significant directed random walk, cancer classification, gene expression dataset, walker network.

INTRODUCTION

High throughput technologies, such as microarray, tandem affinity purification, have enabled the production of a large amount of dataset. In biological context, these datasets contain a lot of information about organisms. Gene expression dataset, pathway dataset, and protein-protein interaction (PPI) dataset are the examples of a huge biological dataset that contains a lot of biological information about an organism (Seah *et al*., 2018). These data can be modelled by networks, where nodes in the networks represent gene/protein and edges represent the relationship between the nodes. These networks, with other high throughput data, can be used to offer unprecedented opportunities for both biological and computational researchers to study the cell at the system level. Many researchers are using micro array data in their experiments (Kasim *et al*., 2016; Angelin-Bonnet *et al*., 2018). Micro array data is adopted to profile gene expression dataset by determining the significant expression level, which is the core criteria in cancer classification.

Pathway networks are used in a biased random walk to allow the walker to apply assisted bias on cancer classification process. Two of the most common pathway networks used are: the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway network and the Protein-Protein Interaction network. The KEGG Pathway network has been used to cross-check the difference between normal genes and cancerous genes (Young & Craft, 2016), while the Protein-Protein Interaction network is applied to predict the gene functions, functional pathways and protein complexes (Huang *et al*., 2016). With the right materials and classification tools, the sensitivity of cancer prediction and accuracy of classification can be increased. Furthermore, much effort had been devoted recently towards incorporating between biological datasets to obtain a better mechanistic understanding of cancerous diseases and to improve the diagnosis and treatment of the diseases such as genetic algorithm (Odeh, 2017), pathway-based cancer classification (Graudenzi, 2017), and random walk (Yang *et al*., 2017).

Random walk algorithm has been used by several researchers (Liu *et al*., 2013; Chen *et al*., 2016; Choon Sen *et al*., 2017; Yang *et al*., 2017; Suki & Frey, 2017) to study the most efficient method in cancer classification. In 2016, Chen introduced random walk with restart for lncRNA-disease association prediction (Chen *et al*., 2016). Lan proposed a bi-random walk to identify potential miRNA environmental factor interaction (Lan *et al*. 2016). Random walk with restart probability, proposed by Matteo in 2017, is used to rank the cancerous gene with respect to cancer modules (Matteo & Giorgio, 2017). In 2017, Suki implemented random walk in his human growth study (Suki & Frey, 2017). In the same year, Seah had introduced significant directed random walk (sDRW) which is able to predict and identify the cancerous gene from gene expression datasets (Seah *et al*., 2017 c). Seah

**Corresponding author:* e-mail: shahreen@uthm.edu.my

had derived a biased random walk equation with different tuning parameters which aim to increase the sensitivity of cancer prediction (Seah *et al*., 2017b).

Significant directed random walk (sDRW) algorithm is designed to predict and classify the cancerous genes from gene expression datasets. In significant directed random walk, KEGG pathway is applied to study the relationship among genes and identify the cancerous genes. The walker in sDRW will walk on the pathway network and cross-check the weight of gene expression dataset to identify the cancerous genes. This model was proved to be successful in classifying cancerous genes by cross-checking the mutual information between gene expression datasets and KEGG pathway datasets.

A network is one of the mandatory data in sDRW. By applying different networks, we can extract different outcomes. In this study, the authors planned to improve the usage of the network in sDRW. Based on the identified feature genes, protein-protein interaction network is constructed which is replaced or integrated with KEGG pathway network in the framework of sDRW. The remaining key functions in sDRW such as tuning parameter selection, weight as a parameter, as well as K-fold cross-validation and classifier remained the same (Liu *et al*., 2013). Six datasets have been applied in this study. The purpose of this study is to identify the different results from the network that run through the sDRW. With these six datasets used as the benchmark datasets, more cancerous datasets could be applied in sDRW. The results of the different applied networks are compared with each other. The validation of this study was evaluated by the number of identified cancerous genes which also determines the sensitivity of cancer prediction and accuracy of cancer classification. The contribution of this approach is listed as below:

- Report statistically significant result by comparison with the four proposed methods.
- Improve significant directed random walk by implementing a network that is more sensitive.

In the next section, we present the applied dataset in this study and the details of methodology of the proposed approach. In section 3, we present the results and discussion of cancer prediction with the number of identified genes. Lastly, we provide the conclusion of the study in section 4.

MATERIALS AND METHODS

We applied multiple datasets to assess the sensitivity of cancer prediction and classification of sDRW with different networks. Gene expression datasets are applied as input dataset. In sDRW, Seah implemented KEGG network as an applied network (Seah *et al*., 2017 c). In this study, PPI was added to study the performance of sDRW against different networks. Fig. 1 illustrates the classification of datasets that are used in this study, while fig. 2 shows the type of datasets employed in this study.

Microarray data

Six gene expression datasets were obtained from Gene Expression Omnibus database (GEO) of National Centre for Biotechnology Information (NCBI). There are two criteria that had been set as a requirement of datasets chosen from NCBI. First, the selected dataset must consist of cancerous genes and normal genes. Second, the dataset should consist of at least 50 samples per dataset. After screening, six cancerous gene expression datasets were chosen to be implemented in this study. Table 1 shows the details of the selected datasets.

Fig. 1: Illustrations of dataset classification

Table 1: Adjacency matrix of KEGG / PPI Network

Genes	A	B	C	2	D	E
A	0	1	0	0	0	0
B	0	0	1	1	0	0
C	0	0	0	1	0	0
2	0	0	0	0	1	0
D	0	0	0	0	0	1
E	0	0	0	0	0	0

For lung cancer dataset, GSE10072 (Landi *et al*., 2008) was selected in this study. GSE10072 contained 107 samples, in which 58 were cancer, while 49 were normal tissue samples. Overall, these samples were collected from 20 non-smokers, 26 former smokers and 28 current smokers.

For liver cancer dataset, GSE17856 (Tsuchiya *et al*., 2010) was selected in this study. GSE17856 contained 87 samples, in which 43 were cancer, while 44 were normal samples. In overall, these 87 samples were hepatocellular carcinoma tissue samples, while the remaining 8 samples were metastatic liver cancer samples.

For thyroid cancer dataset, GSE5364 (Yu *et al*., 2008) was selected in this study. GSE5364 contained 341 samples. Out of 341 samples, 51 were thyroid based dataset, in which 35 were cancer samples, and 16 were normal samples.

Table 1: Information of gene expression datasets

Cancerous Type	Platform ID	Gene Expression Dataset ID	Sample ID	Number of Cancerous Samples	Number of Normal Samples
Lung	GPL96	GSE10072	GSM254625 - GSM254731	58	49
Liver	GPL6480	GSE17856	GSM446165 - GSM446251	43	44
Thyroid	GPL96	GSE5364	GSM121979 - GSM122029	35	16
Stomach	GPL570	GSE13911	GSM350411 - GSM350479	38	31
Kidney	GPL9101	GSE17895	GSM444445 - GSM444610	138	22
Breast	GPL96	GSE1456	GSM107072 - GSM107231	22	130

Table 2: Number of gene detected by sDRW across different walker networks

Datasets	Significant Directed Random Walk, sDRW	Restart Probabilities, *r*								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Lung, GSE10072	KEGG	268	160	118	49	112	118	118	118	49
	PPI	58	49	76	42	84	94	108	59	75
	KEGG-PPI	287	168	134	57	143	130	120	131	89
	PPI-KEGG	274	161	120	50	143	120	120	118	78
Liver, GSE17856	KEGG	21	170	61	73	40	67	136	109	21
	PPI	31	150	69	35	83	95	158	69	62
	KEGG-PPI	46	189	73	84	68	89	163	147	49
	PPI-KEGG	38	163	69	57	102	118	184	94	43
Thyroid, GSE5364	KEGG	23	29	39	33	98	52	13	76	51
	PPI	32	37	31	36	70	37	19	53	85
	KEGG-PPI	50	48	58	43	118	73	21	110	75
	PPI-KEGG	41	41	53	47	120	78	32	105	70
Stomach, GSE13911	KEGG	41	53	109	65	70	108	24	89	41
	PPI	35	63	119	57	69	123	52	69	31
	KEGG-PPI	64	79	154	70	104	154	48	102	64
	PPI-KEGG	75	70	149	68	97	149	58	89	48
Kidney, GSE17895	KEGG	73	39	175	34	53	94	73	19	161
	PPI	57	28	150	42	47	79	53	28	142
	KEGG-PPI	96	48	193	68	75	105	107	32	198
	PPI-KEGG	85	50	185	37	64	95	94	29	184
Breast, GSE1456	KEGG	19	12	19	44	35	21	19	23	26
	PPI	16	10	7	47	28	31	20	27	16
	KEGG-PPI	28	18	25	56	53	47	29	47	37
	PPI-KEGG	26	16	26	53	59	42	24	42	31

For stomach cancer dataset, GSE13911 (D'Errico *et al*., 2009) was selected in this study. GSE13911 was built up by 69 samples, in which 38 were cancer samples and the rest were normal samples.

For kidney cancer dataset, GSE17895 (Dalgliesh *et al*., 2010) was selected in this study. GSE17895 consisted of 138 cancer samples and 22 normal samples.

Breast cancer dataset GSE1456 (Pawitan *et al*., 2005) was selected in this study. GSE1456 consisted of 152 samples, in which 22 were poor samples and 130 were good samples. Breast cancer patients who died within 5 years were considered poor samples while patients who managed to survive more than 5 years without any additional reported events were considered as good samples.

Kyoto encyclopedia of genes and genomes (KEGG) pathway dataset

One of the networks used was obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Database. The global pathway network consists of 150 metabolic and 150 non-metabolic pathways. These pathways are relevant to protein generation. When mutation happens, the proteins continuously generate, resulting in tumor generation.

KEGG pathways are then applied to plot the directed graph by using Sub path way Miner, one of the dependencies package in R programming (Li *et al*., 2009). Overall, the directed graph covers 4113 nodes (genes) and 40875 directed edges. The directed edges show the connection between each gene. fig. 3 displays one of the

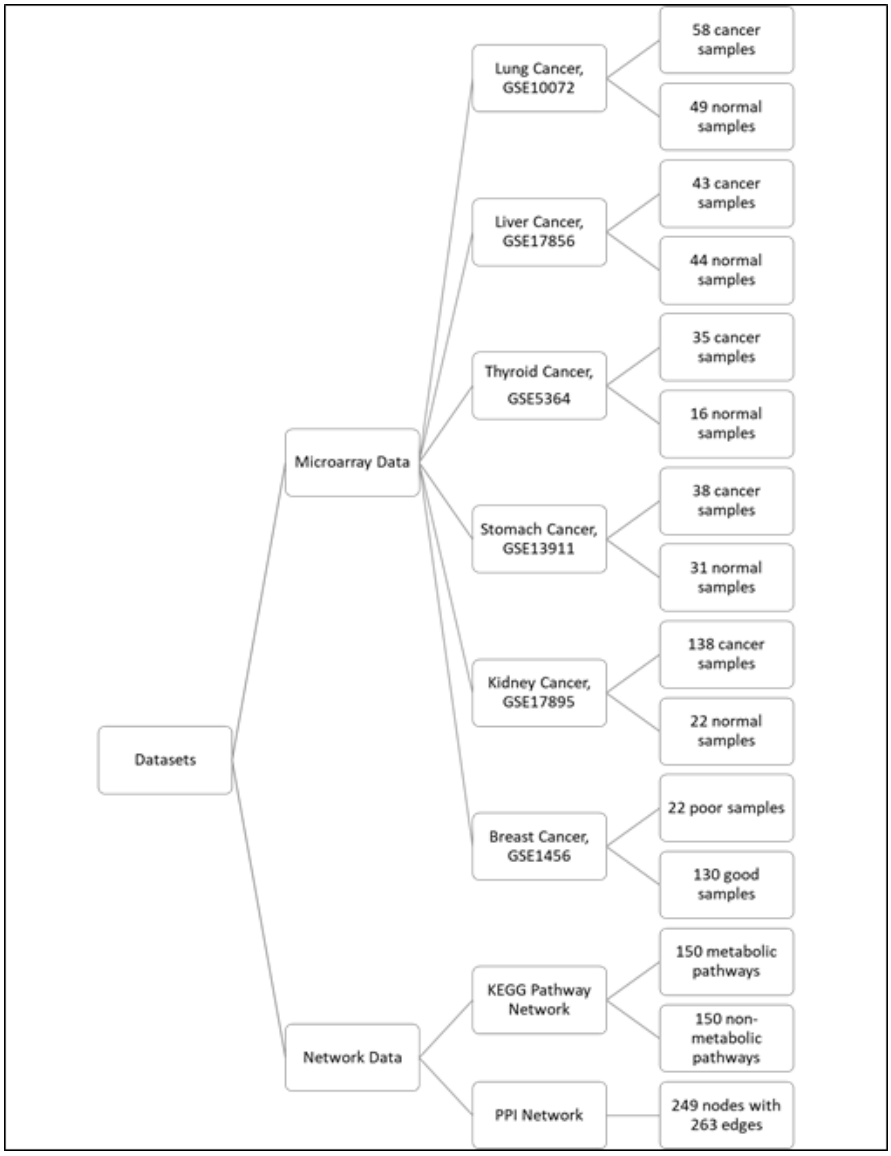

Fig. 2: Type of different datasets used in the proposed approach

biological pathways used in KEGG pathway network, ECM-Receptor Interaction.

Protein-protein interaction (PPI) network

Another network that is applied in this study is PPI network. PPI network is known as one of the important approaches to identify functional modules where the functional modules are set of proteins that are connected (Ren & Liu, 2012). The PPI pathway was downloaded from Human Protein Reference Database (HPRD). The PPI pathway was mapped with identified feature genes to form PPI network. The PPI network was plotted with Path2PPI; one of the dependencies package in R programming (Philipp *et al*., 2016). The constructed PPI network consists of 249 nodes and 263 edges. 249 nodes can be further classified as feature gene nodes (58) and extension gene (direct interaction with more than five feature genes) nodes (191). fig. 4 shows the simple illustration of plotted PPI pathway. Fig. 5 displays the plotted PPI network that is applied in this study.

Methodology

This section defines the approach to construct a significant directed random walk with better walker networks. The first part describes the background of random walk, while the second part introduces the proposed approach.

The random walk

Karl Pearson had introduced random walk to predict the infestation of mosquito in the forest (Pearson, 1905). Random walk can be described as the movement of a particle in a certain state space under the random action (Aldous & Fill, 2014). The state space is usually a dimensional Euclidean space or the integral lattice. Furthermore, random walk was then drawn to the subject

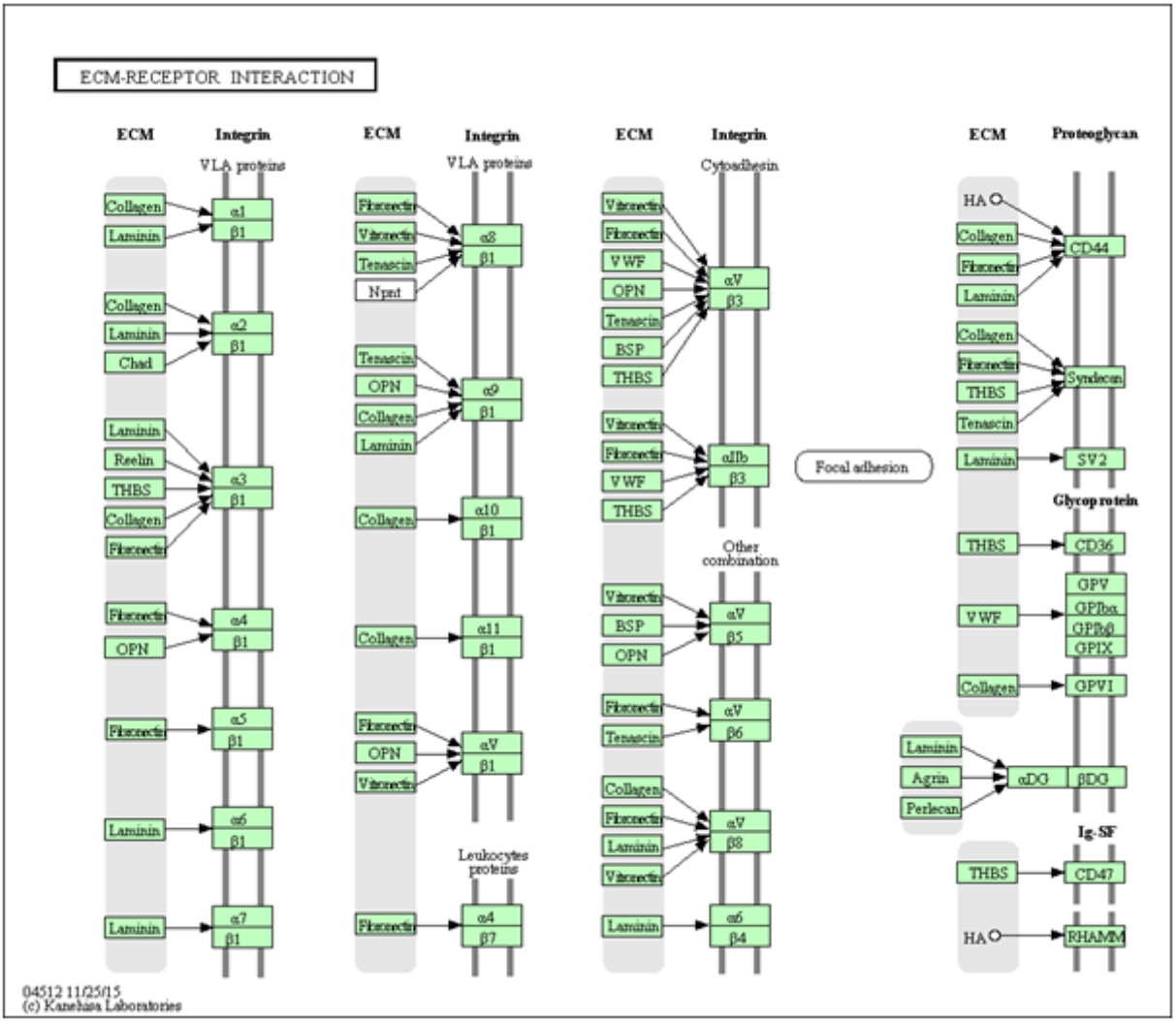

Fig. 3: ECM-Receptor Interaction Pathway (KEGG Pathway, 2017)

and many important fields, such as random processes, random noise, spectral analysis and stochastic equations (Chen *et al*., 2016). And today, many researchers have discovered the pattern or additional information that can influence the final outcome of random walk.

$$X_t = k + \delta_1 + \delta_2 + \ldots + \delta_t, \quad (1)$$

δ_i is every walk at time i. Random walk was enhanced with biased properties (Suki & Frey, 2017) and one of the biased random walks which is famous in cancer classification is directed random walk. Directed random walk (DRW) has been proposed by Liu in 2013 (Liu *et al*., 2013). DRW is aimed to mine the topological information of the protein-protein interaction network. DRW is applied to evaluate the topological importance of genes based on the topological information in a directed graph, and this method is performed on a merged global pathway network build by KEGG pathway (Li *et al*., 2009). DRW will restart the stimulation of random walker that starts on an appointed initial source node s (Seah *et al*., 2017 a). The walker transits from its current node to a random neighbour node or goes back to the source node s with probability, r. Formally, DRW is written as:

$$W_{t+1} = (1 - r)M^T W_1 + rW_0 \quad (2)$$

W_t is a vector where the z-th element holds the probability of being at node z at time t, and M is the row-normalized adjacency matrix of the graph G.

The initial probability, W_0 was constructed by assigning it to each node by using their t-test score, after normalizing to a unit vector. The restart probability, r was set as 0.7. Due to the use of t-test scores as the initial probability, the magnitude of the t-test scores also contributed to weight adjustments (Liu *et al*., 2013). Thus, genes which are both topologically important and significantly differentially expressed will obtain higher weights. This approach was then further enhanced to improve the sensitivity of cancer prediction and accuracy of cancer classification.

Proposed enhanced random walk approach

In this study, Significant directed random walk (sDRW) has been developed from directed random walk with the implementation of walker network which was built by two different sources. By introducing the walker network, the prediction of cancerous genes becomes more sensitive and the accuracy also increases. Properties of random walk allowed the vector to walk randomly on the network, while significant directed random walk (sDRW) uses the vector to walk on the network based on the adjacency matrix of walker networks (Seah *et al*., 2017 c). Significant directed random walk improved the cancerous gene prediction and classification by implementing tuning parameter selection (Nawi *et al*., 2017) and weight as a parameter (Seah *et al*., 2017 c). Tuning parameter selection provides a range of restart probabilities in different cancer datasets. The optimum accuracy of different datasets will be reflected in the study with different restart probabilities. Weight as a parameter in sDRW introduced the weight of genes relating to the next connected genes. The weight vector of genes is different which is dictated by influences of the previous genes (Seah *et al*., 2017 c). Formally, sDRW is defined as:

$$W_{t+1} = (1-r)(M)(\frac{N_1+N_2}{2}) + rW_t \quad (3)$$

Where, W_{t+1} is the vector, determining the cost of travelling towards the next gene, while r is the restart probability with a range of 0.1 until 0.9. M is an adjacency matrix developed from the original directed graph. For single network, M is obtained from the corresponding network. For cross-network, such as KEGG-PPI or PPI-KEGG, M will be M_{KEGG} + M_{PPI}. As weight is one of the parameters playing an important role in determining the connectivity between genes, weights of two connected genes, N_1 and N_2 are used as an average of both the genes to obtain a stable connectivity. W_t is a vector of N node which is transmitted from N-1 node (Seah *et al*., 2017 a).

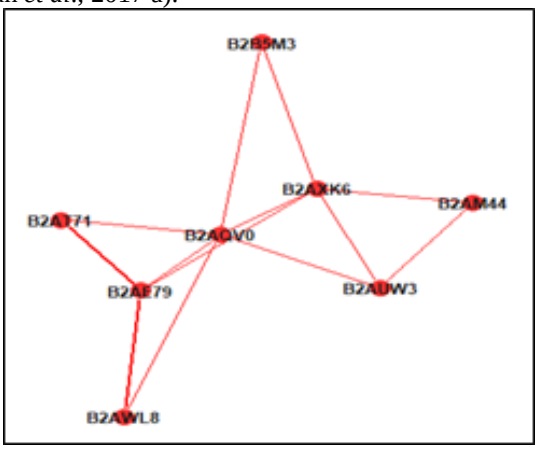

Fig. 4: Simple illustration of PPI network

First, the gene expression datasets were imported into statistical program R. To avoid the effects from variation in the technology rather than from biological differences between the samples, normalization is needed to adjuvant the microarray data. The expression variables were normalized and computed using the function in DRWP Class with program R (Liu *et al*., 2013). Other than pre-processing the microarray data, references datasets applied in this experiment are also needed to transform data-form into network-form followed by adjacency matrix-form. KEGG and PPI networks were downloaded, and four experiments had been conducted to study the role of reference datasets in cancer classification. Normally, reference datasets are used as tertiary datasets to cross-check the result of cancer classification, but in significant directed random walk, reference datasets are used as a network to lead the walker along the experiment, which is also introduced as a walker network. These walker networks are used as a guideline to guide the walker to walk according to the rules. Hence, four walking methods are applied in these four networks. These four experiments were proposed to study the effectiveness of those datasets which affected the results after these networks were applied. The methods are proposed as below:

- KEGG as walker network
- PPI as walker network
- KEGG-PPI as walker network
- PPI-KEGG as walker network

The frameworks which applied these four networks are shown in fig. 6. By going through sDRW, the walker will study the vector and p-value of each gene from the pathway. Those genes that have p-value of less than 0.05 will be used in this study. This is because the p-value will determine the significant cancer mutation. In this study, the previous methods in sDRW such as tuning parameter selection with a range of 0.1 to 0.9 and weight as a parameter were applied in this study. The chosen restart probability for the cancerous datasets is identified as, lung cancer (0.1), liver cancer (0.4), thyroid cancer (0.5), stomach cancer (0.8), kidney cancer (0.6), and breast cancer (0.4) (Seah *et al*., 2017 c).

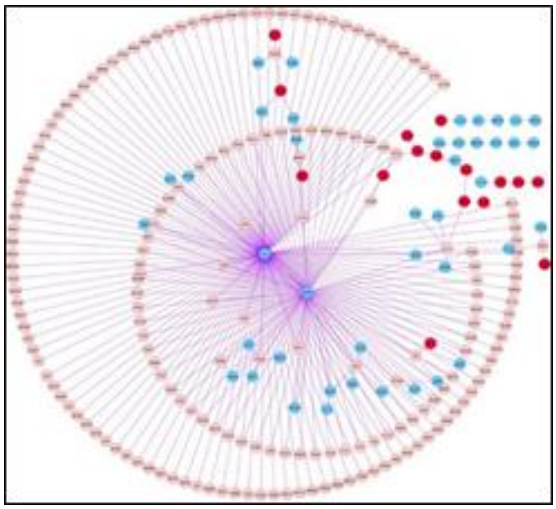

Fig. 5: PPI network for this study

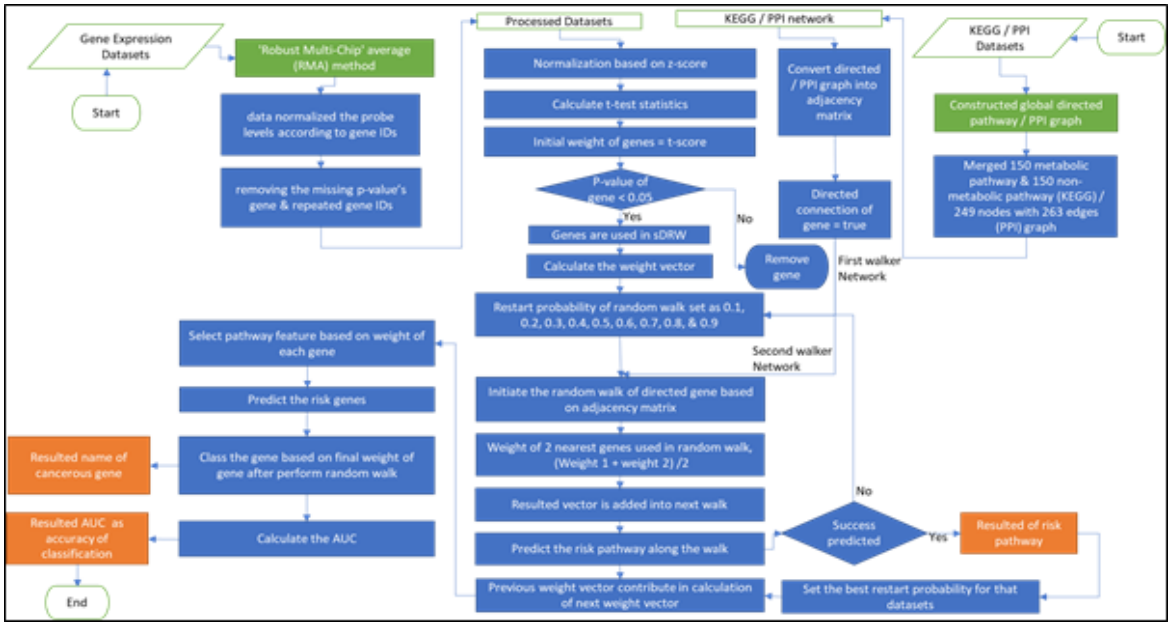

Fig. 6: Framework of sDRW with four proposed walker networks

In this study, four pathway networks are proposed as walker networks to evaluate the impact of the walker networks in cancer classification. The first method illustrated the use of KEGG network as a walker network and allowed sDRW to cross-check the genes weight along with the biological pathway in the directed graph. The second method describes the usage of PPI network as a walker network and allows sDRW to cross-check the sequences of genes which play roles in the formation of protein. The third method illustrates the combination of KEGG and PPI networks (KEGG-PPI). KEGG-PPI applied KEGG as the first walker network while PPI as the second walker network. The genes that are significant in KEGG network will be selected and the process will take place again with the second network, PPI network. Only those genes that are significant in both the networks will be selected for the next process. The fourth method is similar to the third method, but the sequences of the applied network are upside down. PPI network is applied as the first walker network while KEGG as the second walker network.

Usage of walker network

Previously in SDRW, KEGG is the only data source which is implemented as the walker network. Throughout the proposed methods, data source for walker network is expanded into two and the implementation methods increased to four.

KEGG and PPI are implemented as a walker network in SDRW. The connectivity within genes is the reflection of the relationship among genes. With this; it shows "1" as an adjacency matrix, while "0" adjacency matrix is shown if no connection appears. Fig. 7 shows the simple illustration of KEGG/PPI pathway, while table 1 shows the adjacency matrix after conversion from fig. 7. Adjacency matrix in table 1 plays an important role to lead the walker to prejudice towards connection within the genes.

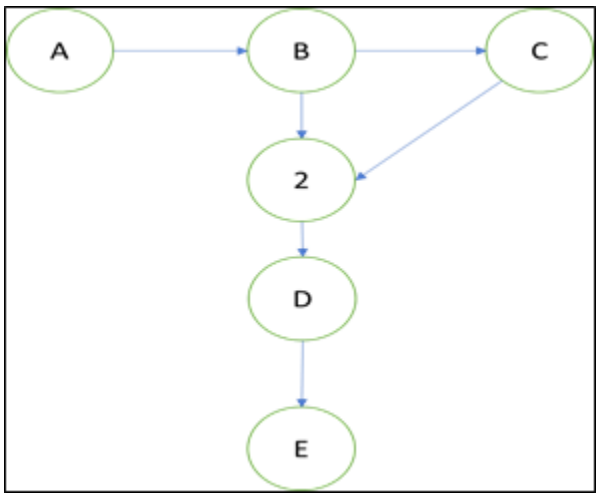

Fig. 7: Simple illustration of KEGG / PPI Network

While for KEGG-PPI & PPI-KEGG walker network, the method to combine both walker networks is hybridization. In KEGG-PPI walker network, KEGG is used as a fundamental network and the PPI network is added into it based on the genes identity (ID) and the connection within the genes. The same method is applied in PPI-KEGG which uses PPI as the fundamental network, and adjacency matrix in KEGG is added into it.

By producing these four walker networks, the experiment can be studied across the results after implementation of SDRW on the networks.

Table 3: Name of risk pathway that predicted by sDRW across different walker networks

Datasets	Restart Probability, *r*	Significant Directed Random Walk, sDRW			
		KEGG Network	PPI Network	KEGG-PPI Network	PPI-KEGG Netowrk
Lung, GSE10072	0.1	Endocytosis, Tight junction, Focal adhesion [268]	Tight junction [58]	Endocytosis, Tight junction, Focal adhesion [287]	Endocytosis, Tight junction [274]
	0.2	Pancreatic selection, Regulation of actin cytoskeleton [160]	ECM-receptor interaction [49]	Pancreatic selection, Regulation of actin cytoskeleton [168]	Pancreatic selection, Regulation of actin cytoskeleton [161]
	0.3	Focal adhesion [118]	ECM-receptor interaction [76]	Focal adhesion [134]	Focal adhesion [120]
	0.4	ECM-receptor interaction [49]	ECM-receptor interaction [42]	ECM-receptor interaction [57]	ECM-receptor interaction [50]
	0.5	Leukocyte transendothelial migration, ECM-receptor interaction [112]	Focal adhesion 84	Leukocyte transendothelial migration, ECM-receptor interaction [143]	Leukocyte transendothelial migration, Focal adhesion [143]
	0.6	Focal adhesion [118]	Leukocyte transendothelial migration [94]	Focal adhesion [130]	Focal adhesion [120]
	0.7	Focal adhesion [118]	Focal adhesion [108]	Focal adhesion [120]	Focal adhesion [120]
	0.8	Pancreatic secretion, Focal adhesion [118]	Regulation of actin cytoskeleton [59]	Pancreatic secretion, Regulation of actin cytoskeleton [131]	Pancreatic secretion, Focal adhesion [118]
	0.9	ECM-receptor interaction [49]	Pancreatic secretion [75]	ECM-receptor interaction, Pancreatic secretion [89]	ECM-receptor interaction [78]
Liver, GSE17856	0.1	Sphingolipid metabolism [21]	Tight junction [31]	Sphingolipid metabolism, Tight junction [46]	Sphingolipid metabolism [38]
	0.2	Focal adhesion, Tight junction [170]	Focal adhesion [150]	Focal adhesion, Tight junction [189]	Focal adhesion [163]
	0.3	Tight junction [61]	Tight junction [69]	Tight junction [73]	Tight junction [69]
	0.4	Sphingolipid metabolism, Glycerolipid metabolism, Lysosome [73]	Glycerolipid metabolism, Lysosome [35]	Sphingolipid metabolism, Glycerolipid metabolism, Lysosome [84]	Sphingolipid metabolism, Lysosome [57]
	0.5	Bacterial invasion of epithelial cells [40]	Sphingolipid metabolism [83]	Bacterial invasion of epithelial cells, Focal adhesion [68]	Bacterial invasion of epithelial cells [102]
	0.6	Glycerolipid metabolism, Bacterial invasion of epithelial cells [67]	Sphingolipid metabolism [95]	Glycerolipid metabolism, Bacterial invasion of epithelial cells, Sphingolipid metabolism [89]	Glycerolipid metabolism, Bacterial invasion of epithelial cells [118]
	0.7	Focal adhesion, Glycerolipid metabolism [136]	Glycerolipid metabolism [158]	Focal adhesion, Glycerolipid metabolism [163]	Focal adhesion [184]
	0.8	Glycerolipid metabolism [109]	Glycerolipid metabolism [69]	Glycerolipid metabolism [147]	Glycerolipid metabolism [94]
	0.9	Sphingolipid metabolism [21]	Bacterial invasion of epithelial cells [62]	Sphingolipid metabolism [49]	Sphingolipid metabolism [43]

Continue…

Thyroid, GSE5364	0.1	Tight junction [23]	Tight junction [32]	Tight junction [50]	Tight junction [41]
	0.2	Cell adhesion molecules (CAMs), Fatty acid metabolism [29]	Cell adhesion molecules (CAMs), [37]	Cell adhesion molecules (CAMs), Fatty acid metabolism, Tight junction, [48]	Cell adhesion molecules (CAMs), Fatty acid metabolism [41]
	0.3	Tight junction, Cell adhesion molecules (CAMs) [39]	Tight junction, Cell adhesion molecules (CAMs) [31]	Tight junction, Cell adhesion molecules (CAMs) [58]	Tight junction, Cell adhesion molecules (CAMs) [53]
	0.4	Fatty acid metabolism, Fc gamma R-mediated phagocytosis [33]	Fatty acid metabolism, [36]	Fatty acid metabolism, Fc gamma R-mediated phagocytosis [43]	Fc gamma R-mediated phagocytosis [47]
	0.5	Regulation of actin cytoskeleton, Wnt signaling pathway, Fc gamma R-mediated phagocytosis, Fatty acid metabolism [98]	Regulation of actin cytoskeleton, Fatty acid metabolism [70]	Regulation of actin cytoskeleton, Wnt signaling pathway, Fc gamma R-mediated phagocytosis, Fatty acid metabolism [118]	Regulation of actin cytoskeleton, Wnt signaling pathway, Fc gamma R-mediated phagocytosis, [120]
	0.6	Wnt signaling pathway, Cell adhesion molecules (CAMs) [52]	Wnt signaling pathway, Cell adhesion molecules (CAMs) [37]	Wnt signaling pathway, Cell adhesion molecules (CAMs) [73]	Wnt signaling pathway, Cell adhesion molecules (CAMs) [78]
	0.7	Fatty acid metabolism [13]	Fatty acid metabolism [19]	Fatty acid metabolism [21]	Fatty acid metabolism [32]
	0.8	MAPK signaling pathway, Fatty acid metabolism [76]	MAPK signaling pathway [53]	MAPK signaling pathway, Fatty acid metabolism [110]	MAPK signaling pathway, Fatty acid metabolism [105]
	0.9	Focal adhesion [51]	Focal adhesion [85]	Focal adhesion [75]	Fatty acid metabolism [70]
Stomach, GSE13911	0.1	TGF-beta signaling pathway [41]	TGF-beta signaling pathway [35]	TGF-beta signaling pathway, Notch signaling pathway [64]	TGF-beta signaling pathway, Notch signaling pathway [75]
	0.2	Hedgehog signaling pathway, Notch signaling pathway [53]	Notch signaling pathway [63]	Hedgehog signaling pathway, Notch signaling pathway [79]	Hedgehog signaling pathway, Notch signaling pathway [70]
	0.3	Wnt signaling pathway, Notch signaling pathway [109]	Wnt signaling pathway, Notch signaling pathway [119]	Wnt signaling pathway, Notch signaling pathway, TGF-beta signaling pathway [154]	Wnt signaling pathway, Notch signaling pathway, TGF-beta signaling pathway [149]
	0.4	Hedgehog signaling pathway, TGF-beta signaling pathway [65]	Hedgehog signaling pathway, TGF-beta signaling pathway [57]	Hedgehog signaling pathway, TGF-beta signaling pathway [70]	Hedgehog signaling pathway, TGF-beta signaling pathway [68]
	0.5	Notch signaling pathway, TGF-beta signaling pathway [70]	Notch signaling pathway, TGF-beta signaling pathway [69]	Notch signaling pathway, TGF-beta signaling pathway [104]	Notch signaling pathway, TGF-beta signaling pathway [97]

Continue…

Stomach, GSE13911	0.6	Regulation of actin cytoskeleton [108]	Regulation of actin cytoskeleton, TGF-beta signaling pathway [123]	Regulation of actin cytoskeleton, TGF-beta signaling pathway [154]	Regulation of actin cytoskeleton, TGF-beta signaling pathway [149]
	0.7	Hedgehog signaling pathway [24]	Hedgehog signaling pathway, Regulation of actin cytoskeleton [52]	Hedgehog signaling pathway, Regulation of actin cytoskeleton [48]	Hedgehog signaling pathway, Regulation of actin cytoskeleton [58]
	0.8	Alanine, Aspartate and glutamate metabolism, Shigellosis, TGF-beta signaling pathway [89]	Alanine, Aspartate and glutamate metabolism, Shigellosis [69]	Alanine, Aspartate and glutamate metabolism, Shigellosis, TGF-beta signaling pathway [102]	Alanine, Aspartate and glutamate metabolism, Shigellosis, TGF-beta signaling pathway [89]
	0.9	TGF-beta signaling pathway [41]	TGF-beta signaling pathway [31]	TGF-beta signaling pathway, Alanine [64]	TGF-beta signaling pathway [48]
Kidney, GSE17895	0.1	Endocytosis, Regulation of actin cytoskeleton [73]	Endocytosis, Regulation of actin cytoskeleton [57]	Endocytosis, Regulation of actin cytoskeleton, Calcium signaling pathway [96]	Endocytosis, Regulation of actin cytoskeleton, Calcium signaling pathway [85]
	0.2	Regulation of actin cytoskeleton [39]	Regulation of actin cytoskeleton [28]	Regulation of actin cytoskeleton [48]	Regulation of actin cytoskeleton [50]
	0.3	Calcium signaling pathway, Phosphatidylinositol signaling system [175]	Calcium signaling pathway, Phosphatidylinositol signaling system [150]	Calcium signaling pathway, Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [193]	Calcium signaling pathway, Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [185]
	0.4	Endocytosis [34]	Endocytosis [42]	Endocytosis, Regulation of actin cytoskeleton [68]	Endocytosis [37]
	0.5	Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [53]	Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [47]	Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [75]	Phosphatidylinositol signaling system, Regulation of actin cytoskeleton [64]
	0.6	Protein processing in endoplasmic reticulum, PPAR signaling pathway, Regulation of actin cytoskeleton [94]	Protein processing in endoplasmic reticulum, PPAR signaling pathway [79]	Protein processing in endoplasmic reticulum, PPAR signaling pathway, Regulation of actin cytoskeleton [105]	Protein processing in endoplasmic reticulum, PPAR signaling pathway, Regulation of actin cytoskeleton [95]
	0.7	Endocytosis, Regulation of actin cytoskeleton [73]	Endocytosis, Regulation of actin cytoskeleton [53]	Endocytosis, Regulation of actin cytoskeleton, PPAR signaling pathway [107]	Endocytosis, Regulation of actin cytoskeleton, PPAR signaling pathway [94]
	0.8	PPAR signaling pathway [19]	PPAR signaling pathway [28]	PPAR signaling pathway [32]	PPAR signaling pathway [29]
	0.9	Calcium signaling pathway [161]	Endocytosis, Regulation of actin cytoskeleton [142]	Calcium signaling pathway, Endocytosis, Regulation of actin cytoskeleton [198]	Calcium signaling pathway, Endocytosis, Regulation of actin cytoskeleton [184]
Breast, GSE1456	0.1	Neuroactive ligand-receptor interaction [19]	Neuroactive ligand-receptor interaction [16]	Neuroactive ligand-receptor interaction [28]	Neuroactive ligand-receptor interaction [26]
	0.2	Glycerophospholipid metabolism [12]	Glycerophospholipid metabolism [10]	Glycerophospholipid metabolism, Neuroactive ligand-receptor interaction [18]	Glycerophospholipid metabolism [16]

Continue…

Breast, GSE1456	0.3	Neuroactive ligand-receptor interaction [19]	Neuroactive ligand-receptor interaction [7]	Neuroactive ligand-receptor interaction [25]	Neuroactive ligand-receptor interaction [26]
	0.4	Adipocytokine signaling pathway, Fatty acid metabolism, Jak-STAT signaling pathway [44]	Adipocytokine signaling pathway, Fatty acid metabolism, Jak-STAT signaling pathway [47]	Adipocytokine signaling pathway, Fatty acid metabolism, Jak-STAT signaling pathway [56]	Adipocytokine signaling pathway, Fatty acid metabolism, Jak-STAT signaling pathway [53]
	0.5	Cytokine-cytokine receptor interaction, Fatty acid metabolism [35]	Cytokine-cytokine receptor interaction [28]	Cytokine-cytokine receptor interaction, Fatty acid metabolism [53]	Cytokine-cytokine receptor interaction, Fatty acid metabolism [59]
	0.6	Jak-STAT signaling pathway [21]	Fatty acid metabolism [31]	Jak-STAT signaling pathway, Fatty acid metabolism [47]	Jak-STAT signaling pathway, Fatty acid metabolism [42]
	0.7	Neuroactive ligand-receptor interaction [19]	Neuroactive ligand-receptor interaction [20]	Neuroactive ligand-receptor interaction [29]	Neuroactive ligand-receptor interaction [24]
	0.8	Chemokine signaling pathway [23]	Chemokine signaling pathway [27]	Chemokine signaling pathway, Fatty acid metabolism [47]	Chemokine signaling pathway, Fatty acid metabolism [42]
	0.9	Adipocytokine signaling pathway, Glycerophospholipid metabolism [26]	Adipocytokine signaling pathway, Fatty acid metabolism [16]	Adipocytokine signaling pathway, Glycerophospholipid metabolism, Fatty acid metabolism [37]	Adipocytokine signaling pathway, Glycerophospholipid metabolism, Fatty acid metabolism [31]

The bold **r** is the optimum result among the restart probability in sDRW.
[ ] is the number of cancerous genes detected.

RESULTS

This section illustrates the result of sDRW with four different methods which applied different walker networks. We evaluated the results based on the impact of the selected genes and risk pathways. We tested different walker networks (KEGG, PPI, KEGG-PPI & PPI-KEGG) with six gene expression datasets (GSE10072, GSE7856, GSE5364, GSE13911, GSE17895, GSE1456) to assess the comprehensiveness of the contribution of networks in sDRW. The main aim of this study is to analyse the effect of using a different combination of walker networks.

Evaluation of the impact of selected significant genes

Cancer prediction is used to predict the significant genes that are relevant to cancerous genes (Khan *et al*., 2017). Cancer prediction is applied to identify the significant genes across four walker networks. Genes expression datasets are being implemented and run on walker networks with sDRW. Previously, in sDRW, tuning parameter and weight as a parameter play important roles to evaluate and identify the significant genes. However, since the key method in this experiment is the implementation of different walker networks, the validation of the walker networks is evaluated by the sensitivity of cancer prediction, which is the number of detected genes. Table 2 shows the number of significant genes that are identified by sDRW through different walker networks. The bolded number refers to the best restart probability based on the accuracy of cancer classification.

Evaluation of the impact of selected risk pathway

Other than the significant genes, we evaluated the impact of the selected risk pathway. Generally, biological pathway contains genes to produce proteins. We named those pathways that contain cancerous genes as risk pathways. By identifying more risk pathways, we can identify more significant genes. Table 3 displays the risk pathways that are identified by sDRW across different walker networks. The selected risk pathways are the potential pathways that contain significant genes.

DISCUSSION

By comparing the optimum restart probability result among the four methods, KEGG-PPI has the 5 best results compared to others, while PPI-KEGG has 1 best result

which falls under liver cancer dataset. Hence, the results show that the best performance among walker networks is of KEGG-PPI. The second highest rank falls under PPI-KEGG. The reason for the best performance of these two walker networks is because they are hybridized walker network from the two original walker networks, KEGG and PPI network. The order of hybridization of walker network brought difference in the results.

The performance of KEGG-PPI was observed to be slightly higher compared to PPI-KEGG, due to the evaluation of effectiveness of walker network by the number of identified genes. Hence, KEGG-PPI would be nominated as the new walker network for sDRW.

In the analyses of the effectiveness of risk pathway, the results obtained are almost the same for those walker networks that applied KEGG network. This is because those risk pathways are protein formation pathways, also known as one of the functions in KEGG pathway (Hongmeng *et al.,* 2016). In general, the total number of identified risk pathways by KEGG-PPI was higher compared to the other networks. Hence, the results proved that the KEGG-PPI network produced the best results.

Generally, KEGG-PPI networks could be used to identify more risk pathways and more significant genes compared to other walker networks. This is rational because the benefit of the biological pathway in KEGG combined with the strength of protein generation sequences in PPI network resulted in the making of the newly selected method, KEGG-PPI network. But the hybridization of PPI-KEGG network doesn't have the same ability level as KEGG-PPI because of the difference in the hybridization order. Different hybridization order slightly changed the adjacency matrix and this could lead to the difference within the genes connection.

CONCLUSION

In this study, KEGG and PPI networks were applied in four methods to study the effectiveness of those networks against the sensitivity of sDRW in significant identified genes and risk pathways. The four networks applied in this study are KEGG, PPI, KEGG-PPI, and PPI-KEGG. The results of this study show that cancer prediction is feasible using cross-network with KEGG-PPI. The outcome shows that the method of KEGG network followed by PPI (KEGG-PPI) network has high-performance results compared to the other methods. Previous studies have used KEGG network as a walker network to lead the walker along the network to identify significant genes and risk pathways. The comparison between the four methods was done by comparing the sensitivity of cancer prediction. The sensitivity of cancer prediction was then further classified as the number of selected significant genes and the number of identified risk pathways. The result demonstrated that the KEGG-PPI network is more effective, and more sensitive compared to other methods. For further enhancement of this research, hybridization of walker network could be done to identify the optimum result from the walker network. Besides that, clustering technique could also be applied to predict cancer genes (Remli *et al*., 2017).

ACKNOWLEDGEMENTS

We would like to thank to Centre for Graduate Studies, Universiti Tun Hussein Onn Malaysia and Mybrain15, Ministry of Education, Malaysia for supporting this research. We would also like to thank UTHM for supporting this research under the Contract Grant Vot number W004.

REFERENCES

Aldous and Fill (2014). Reversible Markov Chains and Random Walks on Graphs. [cited 2 March 2018]. Available from: http://www.stat.berkeley.edu/~aldous/RWG/book.html.

Angelin-Bonnet O, Biggs PJ and Vignes M (2018). Gene Regulatory Networks: A Primer in Biological Processes and Statistical Modelling. Methods in Molecular Biology Gene Regulatory Networks. School of Veterinary Science Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand, pp.347-383.

Chen X, You Z, Yan G and Gong D (2016). IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction. *Oncotarget.,* **7**(36): 57919-57931.

Choon Sen S, Kasim S, Fudzee M, Abdullah R and Atan R (2017). Random walk from different perspective. *Acta electron. Malays.,* **1**(2): 26-27.

D'Errico M, de Rinaldis E, Blasi MF, Viti V, Falchetti M, Calcagnile A, Sera F, Saieva C, Ottini L, Palli D, Palombo F, Giuliani A and Dogliotti E (2009). Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. *Eur. J. Cancer*, **45**(3): 461-469.

Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, Teague J, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Forbes S, Jia M, Jones D, Knott H, Kok CY, Lau KW, Leroy C, Lin ML, McBride DJ, Maddison M, Maguire S, McLay K, Menzies A, Mironenko T, Mulderrig L, Mudie L, O'Meara S, Pleasance E, Rajasingham A, Shepherd R, Smith R, Stebbings L, Stephens P, Tang G, Tarpey PS, Turrell K, Dykema KJ, Khoo SK, Petillo D, Wondergem B, Anema J, Kahnoski RJ, Teh BT, Stratton MR and Futreal PA (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, **463**(7279): 360-363.

Graudenzi A (2017). Pathway-based classification of breast cancer subtypes. *Front Biosci.,* **22**(10): 1697-1712.

Hongmeng X, Hua H, Wei H, Zhihong F, Cong L and Ashraf MA (2016). Research on *in vitro* release of Isoniazid (INH) super paramagnetic microspheres in different magnetic fields. *Pak. J. Pharm. Sci*., **29**(6): 2207-2212

Huang L, Liao L and Wu CH (2016). Inference of protein-protein interaction networks from multiple heterogeneous data. *EURASIP J. Bioinformatics Syst. Biol.*, **1**: 8. eCollection 2016 Dec.

Kasim S, Fudzee MF, Salamat MA, Ramli AA, Mahdin H and Abdullah MH (2016). An improved computational framework using one stage filtration by incorporating knowledge in gene expression clustering. Proceedings of the International Conference on Artificial Intelligence and Robotics and the International Conference on Automation, Control and Robotics Engineering - ICAIR-CACRE 16.

KEGG PATHWAY: ECM-Receptor Interaction Pathway - Homo sapiens (human) [Internet]. Genome.jp. 2017 [cited 2 March 2018]. Available from: http://www.genome.jp/kegg-bin/show_pathway?hsa04512

Khan S, Mohd Nawi N, Shahzad A, Ullah A, Mushtaq M, Mir J and Aamir M (2017). Comparative Analysis for Heart Disease Prediction. *JOIV*, **1**(4-2): 227.

Lan W, Wang J, Li M, Lu C and Wu F *et al*. (2016). Predicting microRNA-environmental factor interactions based on bi-random walk and multi-label learning. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.27-32.

Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, Murphy SE, Yang P, Pesatori AC, Consonni D, Bertazzi PA, Wacholder S, Shih JH, Caporaso NE and Jen J (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE*. **3**(2): e1651.

Li C, Li X, Miao Y, Wang Q, Jiang W and Xu C, Li J, Han J, Zhang F, Gong B and Xu L (2009). Subpathway Miner: A software package for flexible identification of pathways. *Nucleic Acids Res.*, **37**(19): e131-e131.

Liu W, Li C, Xu Y, Yang H, Yao Q, Han J, Shang D, Zhang C, Su F, Li X, Xiao Y, Zhang F, Dai M and Li X (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics,* **29**(17): 2169-2177.

Matteo R and Giorgio V (2017). Random Walking on Functional Interaction Networks to Rank Genes Involved in Cancer. 8th International Conference on Artificial Intelligence Applications and Innovations (AIAI), pp.66-75.

Nawi N, Zaidi N, Hamid N, Rehman M, Ramli A and Kasim S (2017). Optimal Parameter Selection Using Three-term back propagation algorithm for data classification. *Int. J. Adv. Sci. Eng. Inf. Technol.,* **7**(4-2): 1528.

Odeh A (2017). Novel genetic algorithm for early prediction and detection of lung cancer. *Journal of Cancer Treatment and Research*, **5**(2): 15.

Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S and Bergh J (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Res.*, **7**(6): R953-964.

Pearson K (1905). The problem of the random walk. *Nature*, **72**(1865): 294.

Philipp O, Osiewacz H and Koch I (2016). Path2PPI: An R package to predict protein protein interaction networks for a set of proteins. *Bioinformatics*, **32**(9): 1427-1429.

Remli MA, Daud KM, Nies HW, Mohamad MS, Deris S, Omatu S and Sulong G (2017). K-Means Clustering with Infinite Feature Selection for Classification Tasks in Gene Expression Data. Advances in Intelligent Systems and Computing 11th International Conference on Practical Applications of Computational Biology & Bioinformatics, pp.50-57.

Ren G and Liu Z (2012). NetCAD: A network analysis tool for coronary artery disease-associated PPI network. *Bioinformatics*, **29**(2): 279-280.

Seah C, Kasim S, Fudzee M and Mohamad M (2017a). A direct proof of significant directed random walk. IOP Conference Series: Materials Science and Engineering. **235**: 012004.

Seah C, Kasim S, Mohamad M (2017b). Specific Tuning Parameter for Directed Random Walk Algorithm Cancer Classification. *International Journal on Advanced Science, Engineering and Information Technology*, **7**(1): 176.

Seah C, Kasim S, Fudzee M, Law Tze Ping J, Mohamad M, Saedudin R and Ismail M (2017c). An enhanced topologically significant directed random walk in cancer classification using gene expression datasets. *Saudi J. Biol. Sci.*, **24**(8): 1828-1841.

Seah CS, Kasim S, Fudzee MF, Mohamad MS, Saedudin RR, Hassan R and Atan R (2018). An effective pre-processing phase for gene expression classification. *Indonesian Journal of Electrical Engineering and Computer Science*, **11**(3): 1223.

Suki B and Frey U (2017). A time-varying biased random walk approach to human growth. *Scientific Reports*, 7(1): 7805.

Tsuchiya M, Parker J, Kono H, Matsuda M, Fujii H and Rusyn I (2010). Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma. *Mol. Cancer,* **9**(1): 74.

Yang W, Zhide C, Mengjie T, Haixia Z, Xiaolu Y, Ashraf MA, Shuting M and Jing W (2017). Expression of LDH-C (sperm-specific lactate dehydrogenase gene) in skeletal muscle of plateau pika, Ochotona curzoniae, and its effect on anaerobic glycolysis. *Pak. J. Zool*, **49**(3): 905-913

Young MR and Craft DL (2016). Pathway-Informed Classification System (PICS) for cancer analysis using gene expression data. *Cancer Informatics*, 15.

Yu K, Ganesan K, Tan L, Laban M, Wu J and Zhao X *et al*. (2008). A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genetics*, **4**(7): e1000129.