International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Pioneering Methods for Enhancing PPI and Phenotype Networks for Candidate Disease Prioritization

M. Renuka Devi, J. Maria Shyla

Abstract: ***The physical contacts of high-specificity between two or more protein molecules constitute Protein-Protein Interactions (PPIs). PPI networks are modeled through graphs where node denotes proteins and edges denote interaction between proteins. The PPI network plays an important role to identify the interesting disease gene candidates. But, the PPI network usually contains false interactions. Many techniques have been proposed to reconstruct PPI network to remove false interactions and improve ranking of candidate disease. Random Walk with Restart on Diffusion profile (RWRDP) and Random Walk on a Reliable Heterogeneous Network (RWRHN) was two among them. In these methods, Gene topological similarity was incorporated with original PPI network to reconstruct new PPI network. Phenotype network was constructed by calculating similarity between gene phenotypes. The reconstructed network and phenotype networks were combined to rank candidate disease genes. However, the PPI reconstruction was fully related with the quality of protein interaction data. In order to enhance the reconstruction of PPI, a Piecewise Linear Regression (PLR) based protein sequence similarity measure and Bat Algorithm based gene expression similarity were proposed with RHN. In this paper, additional measure called Interaction Level Sub cellular Localization Score (ILSLS) is proposed to further reduce the false interaction in the reconstruction of PPI network. ILSLS is the combination of Normalized Sub cellular Localization score (NSL) and Protein Multiple Location Prediction score (PMLP). The proposed work is named as Random Walker on Optimized Trustworthy Heterogeneous Sub Cellular localization aware Network (RW-OTHSN). In order to enhance the ranking of RW-OTHSN, phenotype structure is considered while construction phenotype network to rank the candidate disease genes. The phenotype structure is characterized based on h*-sequence model which identify highly discriminative signatures with only a small number of genes. This proposed work is named as Random Walker on Optimized Trustworthy Heterogeneous Sub Cellular localization and Phenotype structure aware Network (RW-OTHSPN). The efficiency of the proposed methods are evaluated on PPI network database in terms of Average degree, Relative Frequency for PPI reconstruction, Number of successful predictions, precision and recall for candidate disease gene ranking.***

Keywords: ***Candidate disease gene prediction, candidate disease gene prioritization, phenotype structure, random walk, sub-cellular information.***

I. INTRODUCTION

In disease prevention, treatment and drug design, it has become more important to elucidate the basic molecular mechanisms of disease. One of the most important topics in system biology is learning the connection between casual genes and their associated genes. The task of finding new genes as possible candidates for a phenotype or disease is called prioritization of candidate gene [1, 2]. It is considered as an important issue in modern biomedical research.

Various statistical approaches have been proposed to predict and prioritize the candidate genes by combining different types of data from different sources such as sequenced-based features, gene expressing profiles, and functional annotation information. In the disease prediction process, Protein-Protein Interaction (PPI) [3] plays an important role because PPI network offers functional information in a network environment. Phenotypic disease results may also help increase the accuracy of the candidate gene for less tested phenotypes of disease.

Random Walker on the Reliable Heterogeneous Network (RWRHN) [4] was proposed to identify and rank the candidate disease gene. In RWRHN, the PPI network was reconstructed by incorporating topological similarity with original PPI network to achieve an accurate candidate disease gene prediction. Then, a RHN was constructed which composed of phenotype similarity and reconstructed PPI network. A random walk was devised on RHN to prioritize the candidate disease genes. RWRHN based candidate disease gene prediction accuracy was further improved by RW-OTHN [5] where gene expression profile similarity and protein sequence similarity was considered along with topological similarity for PPI reconstruction.

In this paper, candidate disease gene prediction is enhanced by using sub-cellular localization of protein to reconstruct the original PPI network. SC is introduced to calculate sub-cellular localization of protein. But, a protein may appear in multiple sub-cellular location. In order to compute the sub-cellular localization of protein in multiple locations, ILSLS is introduced. It is the summation of confidence score, NSL score and PMLP score. The ILSLS is used along with topological similarity, protein sequence similarity, gene expression similarity and original PPI network to reconstruct the PPI network.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Dr. M. Renuka Devi, HOD, Department of BCA, Sri Krishna Arts & Science College, Coimbatore (Tamil Nadu) India. E-mail: renuka.srk@gmail.com
J. Maria Shyla*, Ph.D Scholar, Bharathiar University, Coimbatore (Tamil Nadu) India. E-mail: mariashylaphd2019@gmail.com

By using multiple measures, the false interaction between the proteins is reduced. A random walker was devised on phenotype similarity and reconstructed PPI network to rank the candidate disease genes. This process is named as RW-OTHSN. The ranking of candidate disease gene is enhanced by calculating gene-phenotype interaction using phenotype structure. Initially, phenotype structure is characterized based on h*-sequence and then the projection divergence measure is utilized to measure the phenotype structure efficiency. The high quality phenotype structure and the reconstructed PPI network are processed by random walker to rank the candidate disease genes. This process is named as RW-OTHSPN.

II. LITERATURE SURVEY

A candidate gene prioritization method [6] was proposed by integrating topological similarity and semantic similarity. The modularity of the infection must be given greater consideration in the future. A disease gene prioritization method called Laplacian normalization based RWRH network (LapRWRH) [7] was proposed. However, the complexity of LapRWRH is high. A disease gene prioritization method [8] was proposed based on Network Diffusing and Rank Concordance (NDRC). However, the uncertainty and noise in the data source may affect the disease gene prioritization method. A HybridRanker [9] was proposed to rank the candidate genes. The core technique of HybridRanker could also be extended to other environments to provide clear evidence that genes that contribute to diseases are detectable. A cluster-based Positive Unlabeled learning method (C-PUGP) [10] was proposed to identify and rank the disease gene. A Support Vector Machine (SVM) was explored to determine and prioritize the candidate genes. But, the selection of proper kernel function is more difficult in SVM which affects the identification and ranking of candidate gene.

To classify the disease gene, an efficient graph node based on the graph decomposition method [11] was suggested. This method will be expanded to issues associated with genetic diseases that manipulate various heterogeneous sources of information. A novel method [12] was proposed which processed multiple data source to rank the disease risk genes. However, the complexity of the novel method is high. A two-step framework [13] was proposed for disease gene prioritization. In the first-step of the framework, a reliable human functional linkage network was constructed using machine learning techniques and sequence information. Then in the second step of the framework, the gene relations were prioritized based on the constructed functional linkage network. However, constructing functional linkage network based on a heterogeneous network model will increase the efficiency of gene prioritization.

III. PROPOSED METHODOLOGY

Here, the RW-OTHSN and RW-OTHSPN for prediction and prioritization of candidate gene is described in detail. Initially, the gene expression similarity, topological similarity and protein sequence similarity are calculated [4, 5] for prediction of candidate genes. Along with the topological similarity, protein sequence similarity and gene expression similarity data, the OTHSPN, incorporates protein's sub-cellular localization score to reconstruct the PPI network. A gene-phenotype interaction is computed using phenotype structure which used to rank the candidate disease genes.

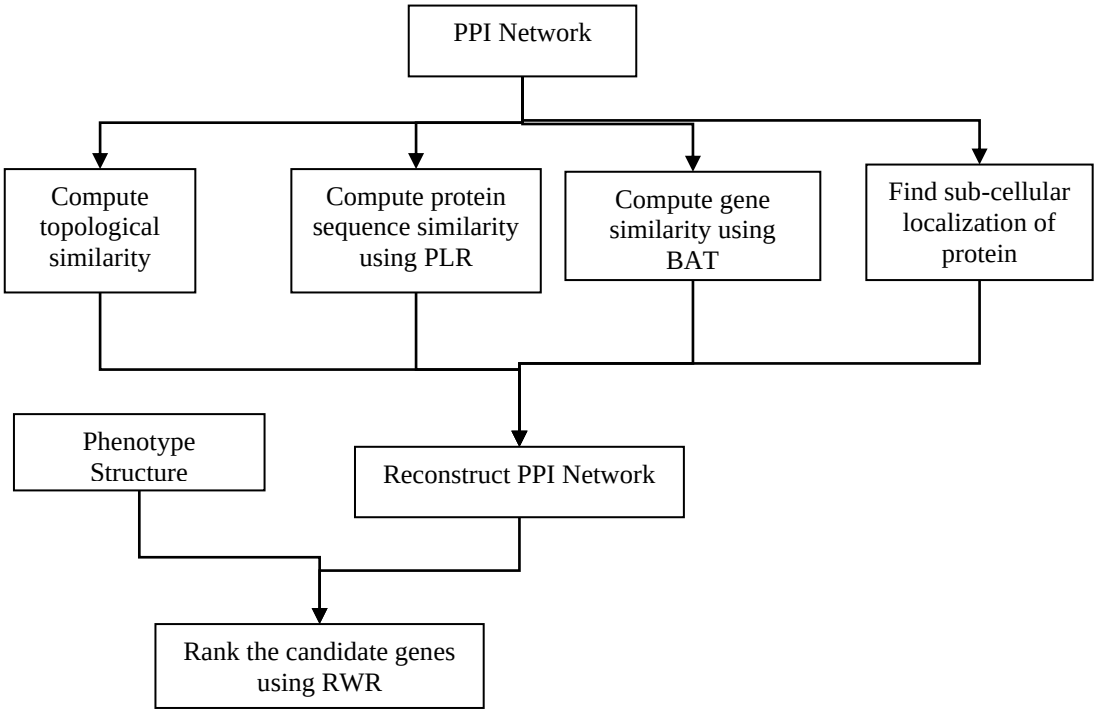

Fig.1. Overall flow of RW-OTHSPN

A. Improving PPI reconstruction using Sub-Cellular localization of proteins

The sub-cellular localization of proteins is calculated as a score function. The cells are elaborately arranged into partitions (i.e., compartments) within membranes such as nucleus or other organelles. All biological functions are conducted in the partitions. As monitored the access and accessible supply of various interacting proteins, micro-environments have an important influence on protein functions. Essentially, the interactions converge strongly among proteins in the same area of the cells. But the degree of concentration widely depends on the partitions. The significance of partition is different in cell activities.

Score of location Compartment

The sub-cellular localization of protein calculates the confidence score between proteins. In order to score the partitions, the number of proteins in each partition is counted. The score of a partition is described as the fraction of the amount of interacted proteins in the partition and the amount of proteins in the largest size partition. The score H of a partition C is computed as,

$$H(I) = \frac{N_P(C)}{N_M} \quad (1)$$

In Equation (1), $N_P(C)$ is the amount of proteins in the partition C and N_M is the amount of proteins in the largest size partition. $H(I)$ is the Score of location Compartment (SC).

Interaction Level Sub-cellular Localization Score

A protein may appear in may appear in multiple sub-cellular locations. For a protein i, its sub-cellular localization score $Score(i)$ is described as the sum of $H(I)$ of all sub-cellular locations it appears. Here, for each protein i the NSL is calculated as,

$$NSL(i) = \frac{Score(i) + Max\_Score(i)}{Max\big(Score(i) + Max\_Score(i)\big)} \quad (2)$$

In Equation (2), $Max\_Score(i)$ is the maximum value of $Score(i)$ for all proteins in gene and $Max()$ takes for all proteins in gene.

A PMLP is calculated based to predict a list of locations for a protein. PMLP score computes the location of candidate disease gene from the known location of proteins. Consider $P = \{p_1, p_2, \dots p_n\}$ is a group of proteins with known locations, $L = \{l_1, l_2, \dots l_m\}$ is the group of all locations and p is candidate gene protein which has no data about its locations. Initially, a bipartite graph $G = (P \cup L, E)$ is constructed where $p_i \in P$ and $l_j \in L$, the edge $e = (p_i, l_j)$ belongs to E. A recommender matrix $Mat = [mat_{ij}]$ with n rows and m columns is calculated [14]. Assume $Ad = [ad_{ij}]_{n\times m}$ is the adjacency matrix of G, where ad_{ij} is assigned as 1 when p_i and l_j are neighbors otherwise ad_{ij} is assigned as 0. The kth row of R is calculated by $f(p_k) \times W^T$, where W^T is the transpose of W and $W = [w_{ij}]_{m\times m}$ which is calculated as,

$$w_{ij} = \frac{1}{d(l_j)} \sum_{t=1}^{n} \frac{ad_{ti} ad_{tj}}{d(p_t)} \quad (3)$$

In Equation (3), $d(l_j)$ and $d(p_t)$ are the grade of vertices l_j and p_t in G respectively. Then the interaction score between p and p_i is denoted as Int_{ppi} which is obtained from string database. Then, define $Int(p) = [Int_{pp_1}, Int_{pp_2}, \dots Int_{pp_n}]$ and predicts the location of p as $Pr(p) = Int(p) \times R$. $Pr(p)$ is the i-th component of the predicted score of location l_i for protein p. The location of a protein i is predicted from the location of other proteins in G. The PMLP score is calculated as,

$$PMLP_i = \frac{score}{high\_Pred(p)} \quad (4)$$

In Eq. (4), $score$ is the score of $Pred(p)$ and $high_{Pred(p)}$ is the highest score of $Pred(p)$. By using $NSL(i)$ and $PMLP_i$, an interaction level sub-cellular localization score of a protein p is calculated as,

$$ILSLS_p = NSL(i) + PMLP_i \quad (5)$$

$$(ILSLS_p)_{ij} = \begin{cases} \frac{\Lambda ILSLS_{ij}}{\sum_j ILSLS_{ij}}, if\ \sum_j ILSLS_{ij} \neq 0 \\ 0, otherwise \end{cases} \quad (6)$$

In Equation (6), Λ denotes jumping likelihood, $ILSLS_{ij} = 1$ represents that the j-th interaction level sub-cellular localization score is related to the i-th interaction level sub-cellular localization score and $ILSLS_{ij} = 0$ otherwise.

Based on topological similarity, protein sequence similarity, gene expression similarity and interaction level sub-cellular localization score of protein, the PPI network is reconstructed as,

$$RE\_PPI = W_1(T_p)_{ij} + W_2(S_p)_{ij} + W_3(E_p)_{ij} + W_4\,(ILSLS_p)_{ij} \quad (7)$$

In Equation (7), $T_p(i_l, j_l)$ is the topological similarity matrix, $(S_p)_{ij}$ is the protein sequence similarity, $(E_p)_{ij}$ is the gene expression similarity, $(ILSLS_p)_{ij}$ is the interaction level sub-cellular localization score and $\sum W_1 + W_2 + W_3 + W_4 = 1$.

B. Gene Phenotype Interaction by Phenotype Structure

Phenotype includes the appearance, growth, and behavior of the organism that is perceived as physical properties. A phenotype structure represents a group of blocks in which each block is a sample subset and a subset of genes to make a separation of p samples from all blocks and samples in a block connected to a phenotype. In order to differentiate this collection of samples from others, the gene expression pattern can furthermore be used in a block as a signature. The signature genes may suggest the possible biomarkers associated with the disease. Therefore, for candidate disease gene ranking, phenotype structure is considered.

A h^*-sequence model is presented to differentiate the phenotype structure. The model uses a definition of a significant chain to guarantee that the any pair of gene values is distinct. This improves the strength of the h^*-sequence model and allows highly discriminatory signatures to be identified using only a small amount of genes. The the candidate structure's value is also calculated based on the projection divergence. The variations between a pair of blocks are enumerated based on the differentiation of the signatures within the blocks.

h*-sequence

A gene expression dataset D is an $p \times l$ matrix, with p samples $Sam = \{Sam_1, Sam_2, \dots, Sam_p\}$ and l genes $Gen = \{Gen_1, Gen_2, \dots Gen_l\}$. The gene expression value of gene Gen_j on sample Sam_i is represented as Exp_{ij}. An idea of Equivalent Dimension Group (EDG) is introduced which denotes a group of genes with same gene expression values.

For a sample $Sam_i \in Sam$, a subset of genes $Gen' \subseteq Gen$ is an EDG, when Gen' assures the below conditions:

$$\max_{Gen_j, Gen_{j'} \in Gen'} \left| Exp_{ij} - Exp_{ij'} \right| < \beta \times \min_{Gen_j \in Gen'} Exp_{ij} \quad (8)$$

$$\forall Gen_j \in Gen', \min_{Gen_{j''} \in (Gen - Gen')} \left| Exp_{ij} - Exp_{ij''} \right| \quad (9)$$

Equation (8) restricts the total difference between any gene expression value pair in an EDG. Equation (9) guarantees the genes are always paired together with their closest neighbor. If a gene assures Equation (8), but is ruled out from an EDG by Equation (9), then that gene is called as breakpoint. According to its very noisy data, the analysis of gene expression data does not take close values into account as ordered. The EDG contains gene expression values together in a set of genes. The gene sequences where any gene pair does not possess the same EDG are durable for noise according to the group threshold β. In fact, the maximum size of the genes is much smaller if such genes are only taken into account than the original.

For a sample Sam_i, a sliding window technique is used to determine all EDGs. Initially, all genes are arranged in the increasing order of their expression values. Then a window is slide from left to right. The window size is determined by Equation (8) and then it is refined by Equation (9). If a breakpoint occurs, from the first breakpoint the next window begins. Else, right to the present left end of the window immediately starts the next window.

In the conventional sequence pattern definitions, there are non overlapping events. Whereas in the h^*-sequence model, the EDGs may overlay with each other. In this sense, a sequence of EDGs is denoted as h^*-sequence where h^* denotes any gaps. The h^*-sequence of sample Sam_i is denoted as δ_i. For a given δ_i, $\mathcal{R}(i,j)$ is a binary relation for a pair of genes i and j. If there subsist an EDG in δ_i containing both i and j genes, then $\mathcal{R}(i,j)$ is true. Or else $\mathcal{R}(i,j)$ is false. Given two h^*-sequence δ_i and δ_j, if $\forall i,j \in \delta_i, \mathcal{R}(i,j)$ always maintains the same value for both δ_i and δ_j. $\delta_i \sqsubseteq \delta_j$ represents δ_i is a subsequence of δ_j. In specific, if $\forall i,j \in \delta_i, \mathcal{R}(i,j)$ always false in δ_i and δ_j, then it is considered that δ_i is the significant chain of δ_j. In addition, δ_i is closed if there is no δ_i' subject to $\forall \delta_j, \delta_i \sqsubseteq \delta_i' \sqsubseteq \delta_j$.

Based on h^*-sequence model, the quality of phenotype structure is quantified. A sample Sam_i is signified by its h^*-sequence δ_i. If p h^*-sequence $\delta_i (i \in [1,p])$ are partitioned into o disjoint subsets $\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_o$. A subsequence S is a signature of subset $\mathcal{S}_l (l \in [1,o])$, iff $\forall S_i \in \mathcal{S}_l, S \sqsubseteq S_i$ and $\forall S_j \notin \mathcal{S}_l, S \not\sqsubseteq S_j$. In particular, if $\forall S_i \in \mathcal{S}_l, S$ is a significant chain of $S_i$and S is called as p-signature of $\mathcal{S}_l$. For a given p-signature sig and a sample Sam, the projection of sig on sample Sam is denoted as $sig|_{Sam}$, denotes the sequence of all genes in sig permuted based on their relative order in S.

A pair of gene is called as reverse pair when the pair of genes in sig has a reverse relative order in $sig|_{Sam}$.

Given sig and $sig|_{Sam'}$ for a gene i, if it is at the oth locus in sig and at the jth locus in $sig|_{Sam'}$, then call $|o-j|$ the distortion of i between sig and $sig|_{Sam'}$ is represented as $dist_i(sig, Sam)$. For a p-signature sig and sample Sam, the projection divergence of sig and $sig|_{Sam'}$ is calculated as,

$$Proj(sig, sig|_{Sam}) = \sum_{i,j \in sig, i \neq j} \varphi(i,j) \left[dist_i(sig, Sam) + dist_j(sig, Sam) \right] \quad (10)$$

In Equation (10),

$$\varphi(i,j) = \begin{cases} 1, & if\,(i,j)\ is\ a\ reverse\ pair \\ 0, & Othersiwe \end{cases} \quad (11)$$

The interconnection among genes is considered in projection divergence measure during computing the difference in individual gene. For gene expression data D, $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots \mathcal{S}_o\}$ be a partition of the p samples and $\mathbb{G} = \{sig_1, sig_2, \dots sig_o\}$ be a set of p-signatures, where sig_i is a p-signature of $\mathcal{S}_i (1 \leq i \leq o)$. A phenotype structure of all sub-matrices $P_{ij} = \{(\mathcal{S}_i, sig_i)\}$. The quality of phenotype structure is calculated as,

$$Q(\mathcal{S}, \mathbb{G}) = \frac{1}{C} \sum_{i=1}^{o} \sum_{j=i+1}^{o} \mathfrak{B}(i,j) \quad (12)$$

In Equation (12),

$$\mathfrak{B}(i,j) = \frac{\sum_{\forall Sam \in \mathcal{S}_j} Proj(sig_i, sig_i|_{Sam}) + \sum_{\forall Sam \in \mathcal{S}_i} Proj(sig_j, sig_j|_{Sam})}{|\mathcal{S}_i| + |\mathcal{S}_j|} \quad (13)$$

In Equation (13), $|\mathcal{S}_i|$ is the amount of samples in $\mathcal{S}_i$ and $|\mathcal{S}_j|$ is the amount of samples in $\mathcal{S}_j$. According to the quality of phenotype structure, the best phenotype structure is selected.

C. Ranking of Candidate disease genes

The phenotype structure and the reconstructed PPI network are utilized to rank the candidate disease genes. The RWR is the most general approach to rank the candidate genes. Generally, the RWR is given as follows:

$$p_{s+1} = (1+\gamma) M^T p_s + \gamma p_0 \quad (14)$$

In Equation (14), γ is the restart likelihood, p_0 is the primary likelihood vector and p_s is a vector in which the i-th element holds the likelihood of discovering the random walker at gene i at step s. The transition matrix of the network is denoted as M and the transition likelihood from gene i to gene j is denoted as M_{ij}. In RW-OTHSPN, uses a phenotype structure with a reconstructed PPI network to rank the candidate genes. The transition matrix of RHN is denoted as

$$M = \begin{bmatrix} M_R & M_{RP} \\ M_{PR} & M_P \end{bmatrix} \quad (15)$$

M_R is the likelihood of random walker to travel from r_i to r_j, where r_i and r_j are the genes in the reconstructed PPI network RE_PPI. M_P is the likelihood of random walker to transit from p_i to p_j which is given as follows:

$$M_P = \begin{cases} \frac{P_{ij}}{\sum_j P_{ij}}, & if \sum_j P_{ij} = 0 \\ (1-\lambda) \frac{P_{ij}}{\sum_j P_{ij}}, & otherwise \end{cases} \quad (16)$$

M_{RP} is the transition likelihood from r_i to p_j and M_{PR} is the transition likelihood from p_i to r_j which are given in Equation (17) and (18).

$$(M_{RP})_{ij} = \begin{cases} \frac{\acute{\Lambda} g_{ij}}{\Sigma_j g_{ij}}, & if \sum_j g_{ij} \neq 0 \\ 0, & otherwise \end{cases} \quad (17)$$

$$(M_{RP})_{ji} = \begin{cases} \frac{\acute{\Lambda} g_{ji}}{\Sigma_j g_{ji}}, & if \sum_j g_{ji} \neq 0 \\ 0, & otherwise \end{cases} \quad (18)$$

The reconstructed PPI network is denoted as $RE_{PPI} = (R_{ij})_{(n\times n)}$, where n is the amount of genes. The phenotype structure is denoted as $P = (P_{ij})$. The initial likelihood vector of the RHN is represented as $p_0 = \begin{bmatrix} (1-\vartheta)R_0 \\ \vartheta P_0 \end{bmatrix}$, where P_0 and R_0 denotes the initial likelihood of phenotype structure and reconstructed PPI network respectively and $\vartheta \in (0,1)$. The transition matrix M represented in Equation (15) and initial likelihood p_0 is applied in Equation (14). The likelihood will reach a steady state after some steps. Finally, the genes are prioritized based on the steady likelihood.

IV. RESULT AND DISCUSSION

Here, the performance of RW-OTHN, RW-OTHSN and RW-OTHSPN are tested in terms of accuracy, Receiver Operating Characteristics (ROC) curve, precision-recall curve, number of successful prediction, relative frequency and average degree. For the experimental purpose, PPI network, phenotype structure and gene expression data are obtained from Human Protein Reference Database (HPRD) which is available in http://www.hprd.org/index_html. The gene-disease association is obtained from Online Mendelian Inheritance in Man (OMIM) which is available in https://omim.org.

The top most candidate genes predicted by IRW-OTHN for six diseases is shown in Table I.

Table- I: Prioritization of Candidate disease gene by RW-OTHSPN

Phenotype	R1	R2	R3	R4	R5	R6	R7
Breast cancer	brca1	pik3ca	nbn	rad51	rb1	brip1	msh2
Lung cancer	egfr	braf	oip5	mras	kras	ralgds	raf1
Prostate cancer	rnasel	elac2	mvp	abce1	ring1	rnase2	cbx4
Gastric cancer	rnase2	ring1	abce1	rnasel	mvp	cbx4	elac2
Colon Cancer	zcchc10	axin2	csnk1e	ankrd6	apc	msh2	brca1
Diabetes mellitus	irs1	mapk8ip1	enpp1	gck	slc2a2	gpd2	pparg

A. Accuracy

Accuracy refers to the number of candidate genes accurately determined by the maximum amount of genes present in the data set. The analysis takes place as,

$$Accuracy = \frac{True\ Positive\ (TP) + False\ Negative\ (FN)}{TP + True\ Negative\ (TN) + False\ Positive(FP) + FN}$$

where, TP is actual candidate disease gene which are exactly predicted as candidate disease gene

TN is actual non-candidate disease gene which are exactly predicted as non-candidate disease gene

FP is known non candidate disease gene which are wrongly predicted as candidate disease gene

FN is known candidate disease gene which are wrongly predicted as non candidate disease gene.

Table II shows the accuracy of RW-OTHN, RW-OTHSN and RW-OTHSPN based candidate disease gene prediction and prioritization.

Table- II: Comparison of Accuracy

	RW-OTHN	RW-OTHSN	RW-OTHSPN
Accuracy	83	85	89

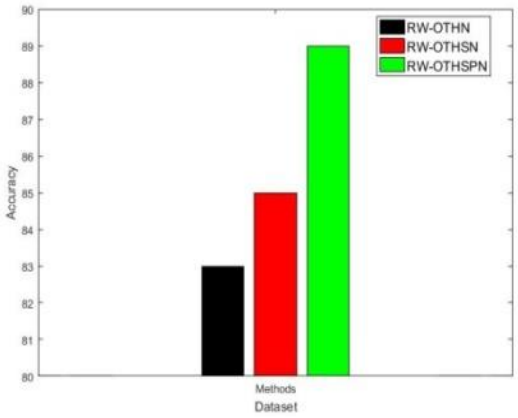

Fig. 2.Comparison of Accuracy

The accuracy for candidate disease gene prediction and prioritization based on RW-OTHN, RW-OTHSN and RW-OTHSPN methods is shown in Fig. 2. The candidate disease gene prediction and prioritization methods are taken in x-axis and the accuracy is taken in y-axis. The accuracy of RW-OTHSPN is 7.23% greater than RW-OTHN and 4.71% greater than RW-OTHSN. From this analysis, it came to know that the RW-OTHSPN high accuracy than other methods.

B. ROC Curve

ROC is a graphical presentation of the relationship between both True Positive Rate (TPR) and False Positive Rate (FPR). TPR is the proportion of TP that are correctly predicted and ranked by candidate disease gene prediction and prioritization methods. FPR is the proportion of TN that are correctly predicted and ranked by candidate disease gene prediction and prioritization methods.

Table III tabulates the ROC value of RW-OTHN, RW-OTHSN and RW-OTHSPN.

Table- III: Comparison of ROC Curve

FPR	TPR		
	RW-OTHN	RW-OTHSN	RW-OTHSPN
0.2	0.7	0.74	0.78
0.4	0.85	0.88	0.91
0.6	0.9	0.95	0.97
0.8	0.96	0.968	0.98
1	1	1	1

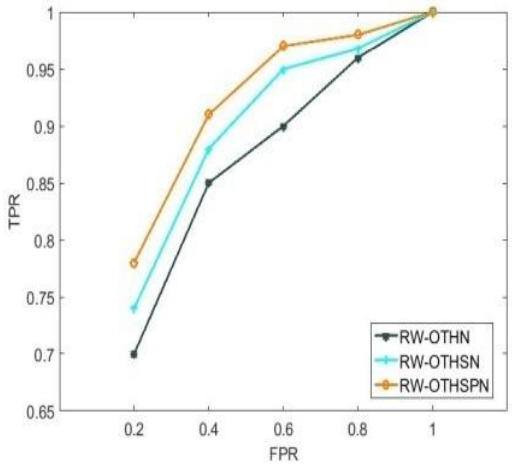

Fig. 3.TPR vs. FPR

Fig. 3 shows the TPR of RW-OTHN, RW-OTHSN and RW-OTHSPN for different range of FPR. When the FPR is 0.6, the TPR of RW-OTHSPN is 7.78% greater than RW-OTHN and 2.11% greater than RW-OTHSN. From this analysis, it came to know that the RW-OTHSPN has better ROC curve than RW-OTHN and RW-OTHSN methods for candidate disease gene prediction and prioritization.

C. Precision-Recall Curve

Precision is calculated based on the candidate disease gene prediction and prioritization at true positive and false positive predictions. It is calculated as,

$$Precision = \frac{TP}{(TP + FP)}$$

Recall is calculated based on the candidate disease gene prediction and prioritization at true positive and false negative predictions.

$$Recall = \frac{TP}{(TP + False\ Negative\ (FN))}$$

Table IV tabulates the precision-recall curve of RW-OTHN, RW-OTHSN and RW-OTHSPN.

Table- IV: Comparison of Precision-Recall Curve

Precision	Recall		
	RW-OTHN	RW-OTHSN	RW-OTHSPN
0.2	0.45	0.5	0.54
0.4	0.4	0.46	0.51
0.6	0.29	0.35	0.39
0.8	0.2	0.27	0.32
1	0.17	0.21	0.26

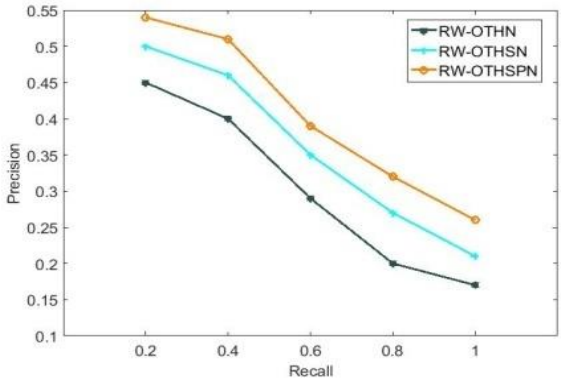

Fig. 4. Precision vs. Recall

Fig. 4 shows the precision of RW-OTHN, RW-OTHSN and RW-OTHSPN for different range of recall. When the precision is 0.6, the recall of RW-OTHSPN is 34.48% greater than RW-OTHN and 11.43% greater than RW-OTHSN. From this analysis, it came to know that the RW-OTHSPN has better precision-recall curve than RW-OTHN and RW-OTHSN methods for candidate disease gene prediction and prioritization.

D. Number of Successful Predictions

The number of successful candidate disease gene predictions for different jumping likelihood is given in Table V.

Table- V: Number of Successful Predictions vs. Jumping Likelihood

λ	Number of Successful Predictions		
	RW-OTHN	RW-OTHSN	RW-OTHSPN
0.2	240	246	250
0.4	245	252	258
0.6	244	250	256
0.8	250	255	262
1	280	286	290

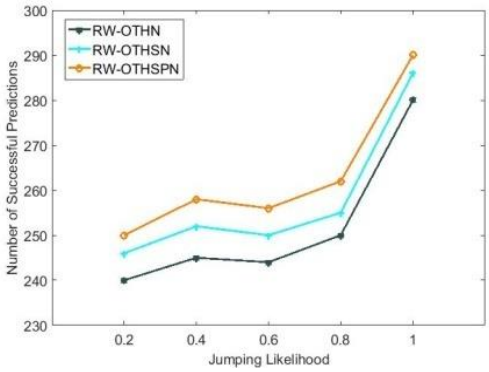

Fig. 5. Number of Successful Predictions vs. Jumping Likelihood

Fig. 5 shows the number of successful prediction by RW-OTHN, RW-OTHSN and RW-OTHSPN for different jumping likelihood. When the jumping likelihood is 0.6, the number of successful predictions by RW-OTHSPN is 4.92% greater than RW-OTHN and 2.4% greater than RW-OTHSN. From this analysis, it came to know that the RW-OTHSPN has high number of successful predictions than other methods.

E. Relative Frequency

It induces similar gene ranking across original PPI network and reconstructed PPI network obtained at different confidence score. Table VI tabulates the relative frequency of RW-OTHN, RW-OTHSN and RW-OTHSPN for different confidence score.

Table- VI: Relative Frequency vs. Confidence Score

Confidence Score	Relative Frequency		
	RW-OTHN	RW-OTHSN	RW-OTHSPN
0.15	6.5	7.1	7.4
0.4	6.1	6.7	7
0.7	5.7	6	6.5
0.9	5.3	5.8	6.3
1	5	5.5	6

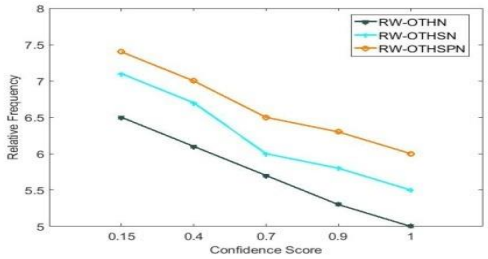

Fig. 6. Relative Frequency vs. Confidence Score

Fig. 6 shows the relative frequency of RW-OTHN, RW-OTHSN and RW-OTHSPN methods under different confidence score. When the confidence score is 0.7, the relative frequency of RW-OTHSPN is 14.04% greater than RW-OTHN and 8.33% greater than RW-OTHSN method. From this analysis, it came to know that the RW-OTHSPN has better relative frequency than other methods.

F. Average Degree

Average degree is a measure of how many proteins are in set compared to number of PPIs in set. Table VII tabulates the average degree of RW-OTHN, RW-OTHSN and RW-OTHSPN for different threshold.

Table- VII: Average Degree vs. Threshold

Threshold	Average Degree		
	RW-OTHN	RW-OTHSN	RW-OTHSPN
0.15	90	200	300
0.4	80	140	250
0.7	30	100	190
0.9	10	50	150
1	2	20	100

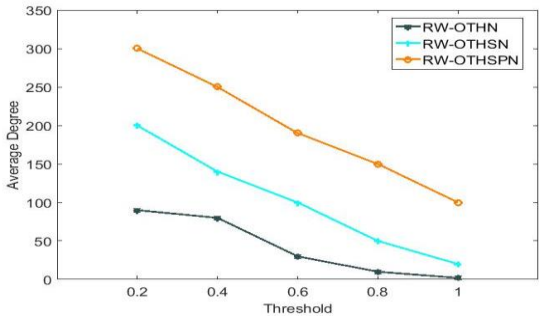

Fig. 7. Average Degree vs. Threshold

Fig. 7 shows the average degree of RW-OTHN, RW-OTHSN and RW-OTHSPN methods under different threshold. When the threshold is 0.7, the average degree of RW-OTHN is 30, RW-OTHSN is 100 and RW-OTHSPN is 190. From this analysis, it came to know that the RW-OTHSPN has better relative frequency than other methods.

V. CONCLUSION

In this article, RW-OTHSN and RW-OTHSPN are proposed to further refine the candidate disease gene prediction and prioritization process. The RW-OTHSN reconstructs the PPI network by computing the sub-cellular localization of proteins with the help of ILSLS. The sub-cellular localization of protein is computed based on an interaction between proteins. The sub-cellular localization is incorporated with the topological similarity, protein sequence similarity and gene similarity for PPI network reconstruction. In, RW-OTHSPN the phenotype structure is used for candidate disease prioritization. The phenotype structure and its quality is quantified by using h*-sequence model and projection divergence. The reconstructed PPI and phenotype structure is used to rank the candidate disease genes based on steady probabilities of random walker. The experimental result shows that the proposed IRW-OTHN method has better accuracy, ROC, precision-recall curve, number of successful predictions, relative frequency and average degree than other methods.

REFERENCES

1. J. Zhu, Y. Qin, T. Liu, J. Wang and X. Zheng, "Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles," in *BMC Bioinform.*, vol. 14, no. 5, p. S5, 2013.
2. Y. Xiao, C. Xu, Y. Ping, J. Guan, H. Fan, Y. Li and X. Li, "Differential expression pattern-based prioritization of candidate genes through integrating disease-specific expression data," *Genom.*, vol. 98, no. 1, pp. 64-71, 2011.
3. L. Zhang, X. Li, J. Tai, W. Li, W and L. Chen, "Predicting candidate genes based on combined network topological features: a case study in coronary artery disease," *PloS one*, vol. 7, no. 6, p. e39542, 2012.
4. J. Luo and S. Liang, "Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data," *J. Biomed. Inf.*, vol. 53, pp. 229-236, 2015.
5. M. R. Devi and J. M. Shyla, "Prioritization of candidate gene associated with diseases improved by random walker on optimized trustworthy heterogeneous network," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 4, pp. 2510-2516, 2019.
6. B. Liu, M. Jin and P. Zeng, "Prioritization of candidate disease genes by combining topological similarity and semantic similarity," *J. Biomed. Inf.*, vol. 57, pp. 1-5, 2015.
7. Z. Q. Zhao, G. S. Han, Z. G. Yu and J. Li, "Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization," *Comput. Biol. Chem.*, vol. 57, pp. 21-28, 2015.
8. M. Fang, X. Hu, Y. Wang, J. Zhao, X. Shen and T. He, "NDRC: a disease-causing genes prioritized method based on network diffusion and rank concordance," *IEEE Trans. Nanobiosci.*, vol. 14, no. 5, pp. 521-527, 2015.
9. Z. Razaghi-Moghadam, R. Abdollahi, S. Goliaei and M. Ebrahimi, "HybridRanker: Integrating network topology and biomedical knowledge to prioritize cancer candidate genes," *J. Biomed. Inf.*, vol. 64, pp. 139-146, 2016.
10. A. Vasighizaker and S. Jalili, "C-PUGP: A cluster-based positive unlabeled learning method for disease gene prediction and prioritization," *Comput. Biol. Chem.*, vol. 76, pp. 23-31, 2018.

11. D. T. Van, A. Sperduti and F. Costa, "The conjunctive disjunctive graph node kernel for disease gene prioritization," *Neurocomputing*, vol. 298, pp. 90-99, 2018.
12. S. Guo, B. Wei, B. Dong, W. Li, S. Wu, Y. He, and W. He, "Prioritizing complex disease risk genes by integrating multiple data," *Genom.*, vol. 111, no. 4, pp. 590-597, 2019.
13. A. Jalilvand, B. Akbari, F. Z. Mirakabad and F. Ghaderi, "Disease gene prioritization using network topological analysis from a sequence based human functional linkage network," *arXiv preprint arXiv:1904.06973*, 2019.
14. T. Zhou, J. Ren, M. Medo and Y. C. Zhang, "Bipartite network projection and personal recommendation," *Phy. Rev. E*, vol. 76, no. 4, p. 046115, 2007.

AUTHORS PROFILE

Dr. M. Renuka Devi is presently heading the Department of BCA in Sri Krishna Arts and Science College. She has more than 19 years experience in teaching professional. She has completed Ph.D in Computer Science on 2012. She has published articles in 47 International Journals, 18 National and 12 International conference proceedings. She has produced Ph.D scholar and M.Phil Scholars. Her areas of interests are Digital Image Processing, Data Mining & Data Warehousing, Wireless Sensor Networks etc.

J. Maria Shyla, pursuing Part-time Ph.D in Computer Science in Bharathiar University. She is working as an Assistant Professor, Department of IT, PSGR Krishnammal College for Women and has 12 years experience in teaching. She has published articles in 3 International Journals, 8 International and 5 National Conference proceedings. Her areas of interest are Data Mining & Data Warehousing, Bioinformatics etc.