Disease Gene Prioritization and the Novel Un-normalized Graph (p-) Laplacian Ranking Methods

Le Trung Hieu, Hoang Trang, Loc Hoang Tran, and Linh Hoang Tran

***Abstract*—The biological motivated problem that we want to solve in this paper is to predict the new members of a partially known set of genes involved in specific disease (i.e. disease gene prioritization). In this problem, we are given a core set of genes (i.e. the queries) involved in the specific disease. However, the biologist experts do not know whether this core set is complete or not. Our objective is to find more potential members of this core set by ranking genes in gene-gene interaction network. One of the solutions to this problem is the random walk on graphs method. However, the random walk on graphs method is not the current state of the art network-based method solving bioinformatics problem. In this paper, the novel un-normalized graph (p-) Laplacian based ranking method will be developed based on the un-normalized graph p-Laplacian operator definitions such as the curvature operator of graph (i.e. the un-normalized graph 1-Laplacian operator) and will be used to solve the disease gene prioritization problem. The results from experiments shows that the un-normalized graph p-Laplacian ranking methods are at least as good as the current state of the art network-based ranking method (*p*=2).**

***Index Terms*—Graph, p-Laplacian, ranking, disease gene prioritization.**

I. Introduction

A genetic disorder is a disease caused by changes and mutations that happen in a single gene or multiple genes in the genome. Some examples to famous single-gene disorders are sick cell anemia, Marfan syndrome, and Huntingtons disease. Oppositely, some diseases such as heart disease, high blood pressure, and cancer are complex diseases that need the interaction of multiple genes. Identification of the genes associated with the latter kind of diseases is a greater challenge since the impact of each gene involved can be insignificant and difficult to identify separately. Identification of disease genes is significant to better understand gene functions as well as to improve medical care [1]. Genome-wide linkage and association studies in healthy and affected populations provide chromosomal regions that most probably have disease-associated genes [2]. Experimental identification of most significant genes among hundreds of candidates by employing an extensive range of data sources is an expensive and time consuming task. Thus, computational methods based on gene expression data [3] and protein-protein interaction (PPI) network data [3], [4] are proposed up to now.

In this paper, we mainly focus on PPI networks that model physical interactions between proteins. These interactions are captured via a variety of experimental and computational methods [5], [6]. Reconstructing a reliable and comprehensive protein interaction network for different species is one of the most significant challenges in molecular biology. Thorough analyses on these networks are shown to be very significant in understanding cells and diseases on a system-wide level. Despite the rise of high-throughput technology, reconstruction of a full network is still far from a realistic goal. Present public PPI databases only cover a portion of these interactions. Thus, accumulations of data from several sources are used in many applications. PPI networks are frequently abstracted by graph models, in which the proteins are nodes of the graph and the physical interactions among them are the undirected edges. This concept supports the application of graph theoretical approaches to the investigation of cellular organization.

Recently, several algorithms have been proposed to include topological properties of PPI networks in understanding genetic diseases [7]-[9]. These algorithms mostly concentrate on prioritization of candidate genes and mainly utilize the idea that the products of genes associated with similar diseases have a higher possibility of being connected in the PPI networks. However, a significant challenge for these applications is the partial and noisy nature of the PPI networks [10]. Missing interactions and false positives affect the accurateness of “local methods” based on local information such as direct interactions and shortest distances. Few “global methods” based on simulation of information flow in the network (e.g., random walks [8], [9] or network propagation [11]) avoid this problem by considering numerous different paths and the whole topology of PPI networks. To solve the disease gene prioritization problem, [8], [9] simulates a random walker that starts from a set of nodes (i.e. the queries or the set of genes involved in the specific disease) instead of a single node. Thus, given a set of proteins that is involved in a specific disease as the start set, the random walk on graphs method ranks the remaining proteins in the protein-protein interaction network with respect to their proximity to the queries’ complex. This ranking method [8], [9] has also been employed by Google Company to exploit the global hyperlink structure of the Web and produce better rankings of search results [12]. Its idea [8], [9], [12] has also been employed in [13] to solve the protein function prediction problem (i.e. the classification problem). However, based on [13], the random walk on graphs method is not the best network-based method solving

Manuscript received October 10, 2015; revised January 18, 2016.

Hieu Le is with the IC Design Lab at Hochiminh City University of Technology, Vietnam (e-mail: lehieu.ee@gmail.com).

Hoang Trang is with Ho Chi Minh City University of Technology, Vietnam (e-mail: hoangtrang@hcmut.edu.vn).

Loc Tran is with University of Technology, Sydney, Australia (e-mail: tran0398@umn.edu).

Linh Tran is with Portland State University, Portland (e-mail: linht@pdx.edu).

doi: 10.18178/ijmlc.2016.6.1.575

the classification bioinformatics problems such as protein function prediction [13], [14] and cancer classification [15]. Unlike the random walk method utilizing the random walk graph Laplacian, the network propagation method [11] employs the symmetric normalized graph Laplacian. Moreover, to the best of our knowledge, the un-normalized graph (p-) Laplacian based semi-supervised learning method is considered the current state of the art network-based method solving protein function prediction problem [14], [16] and cancer classification problem [15]. However, the un-normalized graph (p-) Laplacian based ranking method has not yet been developed and obviously has not been applied to any practical applications. In this paper, the un-normalized graph (p-) Laplacian based ranking method will be developed based on the un-normalized graph p-Laplacian operator definitions such as the curvature operator of graph (i.e. the un-normalized graph 1-Laplacian operator). Finally, this proposed method will be used to solve the disease gene prioritization problem.

We will organize the paper as follows: Section II will introduce the preliminary notations and definitions used in this paper. Section III will introduce the definition of the gradient and divergence operators of graphs. Section IV will introduce the definition of Laplace operator of graphs and its properties. Section V will introduce the definition of the curvature operator of graphs and its properties. Section VI will introduce the definition of the p-Laplace operator of graphs and its properties. Section VII will show how to derive the algorithm of the un-normalized graph p-Laplacian based ranking method from regularization framework. In section VIII, we will compare the accuracy performance measures of the un-normalized graph Laplacian based ranking algorithm and the un-normalized graph p-Laplacian based ranking algorithms. Section IX will conclude this paper and the future direction of researches of other practical applications in bioinformatics utilizing discrete operator of graph will be discussed.

II. Preliminary Notations and Definitions

Given a graph *G=(V,E,W)* where *V* is a set of vertices with $|V| = n$, $E \subseteq V * V$ is a set of edges and W is a $n * n$ similarity matrix with elements $w_{ij} > 0$ $(1 \leq i, j \leq n)$.

Also, please note that $w_{ij} = w_{ji}$.

The degree function $d: V \rightarrow R^{+}$ is

$$d_i = \sum_{j \sim i} w_{ij}, \tag{1}$$

where $j \sim i$ is the set of vertices adjacent with *i*.

Define $D = diag(d_1, d_2, \dots, d_n)$.

The inner product on the function space R^V is

$$< f, g >_V = \sum_{i \in V} f_i g_i \tag{2}$$

Also define an inner product on the space of functions R^E on the edges

$$< F, G >_E = \sum_{(i,j) \in E} F_{ij} G_{ij} \tag{3}$$

Here let $H(V) = (R^V, <.,.>_V)$ and $H(E) = (R^E, <$ *.,.>E)* be the Hilbert space real-valued functions defined on the vertices of the graph *G* and the Hilbert space of real-valued functions defined in the edges of *G* respectively.

III. Gradient and Divergence Operators

We define the gradient operator $d: H(V) \rightarrow H(E)$ to be

$$(df)_{ij} = \sqrt{w_{ij}}(f_j - f_i), \tag{4}$$

where $f: V \rightarrow R$ be a function of $H(V)$.

We define the divergence operator $div: H(E) \rightarrow H(V)$ to be

$$< df, F >_{H(E)} = < f, -divF >_{H(V)}, \tag{5}$$

where $f \in H(V), F \in H(E)$

Next, we need to prove that

$$(divF)_j = \sum_{i \sim j} \sqrt{w_{ij}}\,(F_{ji} - F_{ij})$$

Proof:

$$\begin{aligned}
< df, F > &= \sum_{(i,j) \in E} df_{ij} F_{ij} \\
&= \sum_{(i,j) \in E} \sqrt{w_{ij}}\,(f_j - f_i) F_{ij} \\
&= \sum_{(i,j) \in E} \sqrt{w_{ij}}\, f_j F_{ij} - \sum_{(i,j) \in E} \sqrt{w_{ij}}\, f_i F_{ij} \\
&= \sum_{k \in V} \sum_{i \sim k} \sqrt{w_{ik}}\, f_k F_{ik} - \sum_{k \in V} \sum_{j \sim k} \sqrt{w_{kj}}\, f_k F_{kj} \\
&= \sum_{k \in V} f_k (\sum_{i \sim k} \sqrt{w_{ik}}\, F_{ik} - \sum_{i \sim k} \sqrt{w_{ki}}\, F_{ki}) \\
&= \sum_{k \in V} f_k \sum_{i \sim k} \sqrt{w_{ik}}\,(F_{ik} - F_{ki})
\end{aligned}$$

Thus, we have

$$(divF)_j = \sum_{i \sim j} \sqrt{w_{ij}}\,(F_{ji} - F_{ij}) \tag{6}$$

IV. Laplace Operator

We define the Laplace operator $\Delta: H(V) \rightarrow H(V)$ to be

$$\Delta f = -\tfrac{1}{2} div(df) \tag{7}$$

Next, we compute

$$\begin{aligned}
(\Delta f)_j &= \frac{1}{2} \sum_{i \sim j} \sqrt{w_{ij}}\,((df)_{ij} - (df)_{ji}) \\
&= \frac{1}{2} \sum_{i \sim j} \sqrt{w_{ij}}\,(\sqrt{w_{ij}}(f_j - f_i) - \sqrt{w_{ij}}(f_i - f_j)) \\
&= \sum_{i \sim j} w_{ij}\,(f_j - f_i) \\
&= \sum_{i \sim j} w_{ij} f_j - \sum_{i \sim j} w_{ij}\, f_i \\
&= d_j f_j - \sum_{i \sim j} w_{ij}\, f_i
\end{aligned}$$

Thus, we have

$$(\Delta f)_j = d_j f_j - \sum_{i \sim j} w_{ij}\, f_i \tag{8}$$

The graph Laplacian is a linear operator. Furthermore, the graph Laplacian is self-adjoint and positive semi-definite.

Let $S_2(f) = < \Delta f, f >$, we have the following **theorem 1**

$$D_f S_2 = 2\Delta f \tag{9}$$

The proof of the above theorem can be found from [15], [17].

V. Curvature Operator

We define the curvature operator $\kappa : H(V) \to H(V)$ to be

$$\kappa f = -\frac{1}{2} div\left(\frac{df}{||df||}\right) \tag{10}$$

Next, we compute

$$(\kappa f)_j = \frac{1}{2}\sum_{i\sim j} \sqrt{w_{ij}} \left(\left(\frac{df}{||df||}\right)_{ij} - \left(\frac{df}{||df||}\right)_{ji}\right)$$

$$= \frac{1}{2}\sum_{i\sim j} \sqrt{w_{ij}} \left(\frac{1}{||d_i f||}\sqrt{w_{ij}}(f_j - f_i) - \frac{1}{\|d_j f\|}\sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \frac{1}{2}\sum_{i\sim j} w_{ij} \left(\frac{1}{||d_i f||} + \frac{1}{\|d_j f\|}\right)(f_j - f_i)$$

Thus, we have

$$(\kappa f)_j = \frac{1}{2}\sum_{i\sim j} w_{ij} \left(\frac{1}{||d_i f||} + \frac{1}{\|d_j f\|}\right)(f_j - f_i) \tag{11}$$

From the above formula, we have

$$d_i f = ((df)_{ij} : j\sim i)^T \tag{12}$$

The local variation of f at i is defined to be

$$||d_i f|| = \sqrt{\sum_{j\sim i} (df)_{ij}^2} = \sqrt{\sum_{j\sim i} w_{ij} (f_j - f_i)^2} \tag{13}$$

To avoid the zero denominators in (11), the local variation of f at i is defined to be

$$||d_i f|| = \sqrt{\sum_{j\sim i} (df)_{ij}^2 + \epsilon}, \tag{14}$$

where $\epsilon = 10^{-10}$.

The graph curvature is a non-linear operator.

Let $S_1(f) = \sum_i ||d_i f||$, we have the following **theorem 2**

$$D_f S_1 = \kappa f \tag{15}$$

The proof of the above theorem can be found from [15], [17].

VI. p-Laplace Operator

We define the p-Laplace operator $\Delta_p : H(V) \to H(V)$ to be

$$\Delta_p f = -\frac{1}{2} div(||df||^{p-2} df) \tag{16}$$

Clearly, $\Delta_1 = \kappa$ and $\Delta_2 = \Delta$. Next, we compute

$$\left(\Delta_p f\right)_j = \frac{1}{2}\sum_{i\sim j} \sqrt{w_{ij}} \left(||df||^{p-2} df_{ij} - ||df||^{p-2} df_{ji}\right)$$

$$= \frac{1}{2}\sum_{i\sim j} \sqrt{w_{ij}} \left(||d_i f||^{p-2} \sqrt{w_{ij}}(f_j - f_i) - \|d_j f\|^{p-2} \sqrt{w_{ij}}(f_i - f_j)\right)$$

$$= \frac{1}{2}\sum_{i\sim j} w_{ij} \left(||d_i f||^{p-2} + \|d_j f\|^{p-2}\right)(f_j - f_i)$$

Thus, we have

$$\left(\Delta_p f\right)_j = \frac{1}{2}\sum_{i\sim j} w_{ij} \left(||d_i f||^{p-2} + \|d_j f\|^{p-2}\right)(f_j - f_i) \tag{17}$$

Let $S_p(f) = \frac{1}{p}\sum_i ||d_i f||^p$, we have the following **theorem 3**

$$D_f S_p = p\Delta_p f \tag{18}$$

The proof of the above theorem can be found from [15], [17].

VII. Discrete Regularization on Graphs and Protein Function Classification Problems

Given a protein network $G=(V, E)$. V is the set of all proteins in the network and E is the set of all possible interactions between these proteins. Let y denote the initial function in $H(V)$. y_i can be defined as follows

$$y_i = \begin{cases} 1 \text{ if protein } i \text{ is the query} \\ 0 \text{ if protein } i \text{ is not the query} \end{cases}$$

Our goal is to look for an estimated function f in $H(V)$ such that f is not only smooth on G but also close enough to an initial function y. Then each protein i is ranked as value of f_i. This concept can be formulated as the following optimization problem

$$\text{argmin}_{f\in H(V)}\{S_p(f) + \frac{\mu}{2}||f - y||^2\} \tag{19}$$

The first term in (19) is the smoothness term. The second term is the fitting term. A positive parameter μ captures the trade-off between these two competing terms.

A. 2-Smoothness

When p=2, the optimization problem (19) is

$$\text{argmin}_{f\in H(V)}\{\frac{1}{2}\sum_i ||d_i f||^2 + \frac{\mu}{2}||f - y||^2\} \tag{20}$$

By theorem 1, we have

Theorem 4: The solution of (20) satisfies

$$\Delta f + \mu(f - y) = 0 \tag{21}$$

Since Δ is a linear operator, the closed form solution of (21) is

$$f = \mu(\Delta + \mu I)^{-1} y, \tag{22}$$

Where I is the identity operator and $\Delta = D - W$. (22) is the algorithm proposed by [13], [16].

B. 1-Smoothness

When p=1, the optimization problem (19) is

$$\text{argmin}_{f\in H(V)}\{\sum_i ||d_i f|| + \frac{\mu}{2}||f - y||^2\}, \tag{23}$$

By theorem 2, we have

Theorem 5: The solution of (23) satisfies

$$\kappa f + \mu(f - y) = 0, \tag{24}$$

The curvature κ is a non-linear operator; hence we do not have the closed form solution of equation (24). Thus, we have to construct iterative algorithm to obtain the solution. From (24), we have

$$\frac{1}{2}\sum\nolimits_{i\sim j} w_{ij}\left(\frac{1}{\|d_i f\|}+\frac{1}{\|d_j f\|}\right)(f_j - f_i) + \mu(f_j - y_j) = 0 \quad (25)$$

Define the function $m: E \to R$ by

$$m_{ij} = \frac{1}{2} w_{ij}\left(\frac{1}{\|d_i f\|}+\frac{1}{\|d_j f\|}\right) \quad (26)$$

Then (25)

$$\sum_{i\sim j} m_{ij}\,(f_j - f_i) + \mu(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum\nolimits_{i\sim j} m_{ij} + \mu\right) f_j = \sum\nolimits_{i\sim j} m_{ij}\, f_i + \mu y_j \quad (27)$$

Define the function $p: E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i\sim j} m_{ij}+\mu} & if\ i \neq j \\ \frac{\mu}{\sum_{i\sim j} m_{ij}+\mu} & if\ i = j \end{cases} \quad (28)$$

Then

$$f_j = \sum\nolimits_{i\sim j} p_{ij} f_i + p_{jj} y_j \quad (29)$$

Thus we can consider the iteration

$$f_j^{(t+1)} = \sum\nolimits_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j \ \forall j \in V$$

to obtain the solution of (23).

C. p-Smoothness

For any number p, the optimization problem (19) is

$$argmin_{f\in H(V)}\left\{\frac{1}{p}\sum\nolimits_i \|d_i f\|^p + \frac{\mu}{2}\|f - y\|^2\right\}, \quad (30)$$

By theorem 3, we have

Theorem 6: The solution of (30) satisfies

$$\Delta_p f + \mu(f - y) = 0, \quad (31)$$

The *p-Laplace* operator is a non-linear operator; hence we do not have the closed form solution of equation (31). Thus, we have to construct iterative algorithm to obtain the solution. From (31), we have

$$\frac{1}{2}\sum\nolimits_{i\sim j} w_{ij}\left(\|d_i f\|^{p-2} + \left\|d_j f\right\|^{p-2}\right)(f_j - f_i) + \mu(f_j - y_j) = 0 \quad (32)$$

Define the function $m: E \to R$ by

$$m_{ij} = \frac{1}{2} w_{ij}(\|d_i f\|^{p-2} + \left\|d_j f\right\|^{p-2}) \quad (33)$$

Then equation (32) which is

$$\sum_{i\sim j} m_{ij}\,(f_j - f_i) + \mu(f_j - y_j) = 0$$

can be transformed into

$$\left(\sum\nolimits_{i\sim j} m_{ij} + \mu\right) f_j = \sum\nolimits_{i\sim j} m_{ij}\, f_i + \mu y_j \quad (34)$$

Define the function $p: E \to R$ by

$$p_{ij} = \begin{cases} \frac{m_{ij}}{\sum_{i\sim j} m_{ij}+\mu} & if\ i \neq j \\ \frac{\mu}{\sum_{i\sim j} m_{ij}+\mu} & if\ i = j \end{cases} \quad (35)$$

Then

$$f_j = \sum\nolimits_{i\sim j} p_{ij} f_i + p_{jj} y_j \quad (36)$$

Thus we can consider the iteration

$$f_j^{(t+1)} = \sum\nolimits_{i\sim j} p_{ij}^{(t)} f_i^{(t)} + p_{jj}^{(t)} y_j \ \forall j \in V$$

to obtain the solution of (30).

VIII. Experiments and Results

A. Datasets

In this paper, we use the dataset available from [18] and references therein. This dataset contains the gene-gene interaction network containing 8959 genes and 68360 undirected interactions (i.e. edges). In order to evaluate the performance of the un-normalized graph p-Laplacian based ranking algorithms, we used the default seed set of three genes available from [18] and is involved in one specific disease. The IDs of these three genes are 673, 2064, and 5071.

B. Experiments

In this section, we experiment with the above proposed un-normalized graph p-Laplacian ranking methods with *p*=1.7, 1.8, 1.9 and the current state of the art method (i.e. the un-normalized graph Laplacian based ranking method *p*=2) in terms of accuracy performance measure. All experiments were implemented in Matlab 6.5 on virtual machine. The leave-one-out testing strategy is used to compute the accuracy performance measures of all methods used in this paper. For the default seed set, one member gene is left out and the remaining genes are used as the core set in the membership queries. Effective ranking methods should report the left out gene in top *k* ranks. The parameter μ is set to 1. The accuracy performance measures of the above proposed methods and the current state of the art method is given in the following Table I.

The results from the above table shows that the un-normalized graph p-Laplacian ranking methods are at least as good as the current state of the art method (*p*=2).

TABLE I: The Comparison of Accuracies of Proposed Methods with Different *P*-Values

Top *k* ranks		*k*=1000	*k*=1500	*k*=2000
Accuracy performance measures (%)	*p*=1.7	0	0	33
	p=1.8	0	0	**100**
	p=1.9	0	**67**	**100**
	p=2 (current state of the art method)	0	**67**	**100**

IX. Conclusions

We have developed the detailed regularization frameworks for the un-normalized graph p-Laplacian ranking methods applying to disease gene prioritization problem. Experiments show that the un-normalized graph p-Laplacian ranking methods are at least as good as the current state of the art method (i.e. *p*=2).

Recently, to the best of my knowledge, the un-normalized directed graph p-Laplacian based ranking methods have not yet been developed and applied to any practical problems. This method is worth investigated in the future because of its difficult nature and its close connection to partial differential equation on directed graph field.

REFERENCES

[1] H. G. Brunner and M. A. van Driel, "From syndrome families to functional genomics," *Nat Rev Genet*, vol. 5, no. 7, pp. 545–551, Jul. 2004

[2] E. Evangelou, D. M. Maraganore, and J. P. Ioannidis, "Meta-analysis in genome-wide association datasets: Strategies and application in Parkinson disease," *PLoS ONE*, vol. 2, no. 2, p. e196, 2007

[3] K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Biotech*, vol. 25, no. 3, pp. 309–316, Mar. 2007.

[4] O. Vanunu and R. Sharan, "A propagation based algorithm for inferring gene-disease associations," in *Proc. German Conference on Bioinformatics*, 2008

[5] D. Auerbach, S. Thaminy, M. Hottiger, and I. Stagljar, "The post-genomic era of interactive proteomics: Facts and perspectives," *Proteomics*, vol. 2, pp. 611–623, 2002.

[6] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," in *Proc Natl Acad Sci USA*, April 2001, vol. 98, no. 8, pp. 4569–4574.

[7] K. Lage, E. Karlberg *et al*., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Bio*, vol. 25, no. 3, pp. 309-316, 2007.

[8] J. Chen, B. Aronow, and A. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics*, no. 10, p. 73, 2009.

[9] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Am J Hum Genet*, vol. 82, no. 4, pp. 949-958, 2008.

[10] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein, "Bridging structural biology and genomics: assessing protein interaction data with known complexes," *Trends in Genetics*, vol. 18, no. 10, pp. 529-536, 2002.

[11] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *Plos Computational Biology*, 2010.

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks And ISDN Systems*, no. 30, pp. 107-117, 1998.

[13] L. Tran, "Application of three graph Laplacian based semi-supervised learning methods to protein function prediction problem," CoRR abs/1211.4289, 2012.

[14] L. Tran, "The un-normalized graph p-Laplacian based semi-supervised learning method and protein function prediction problem," in *Proc. The Fifth International Conference on Knowledge Systems and Engineer*, 2013.

[15] L. Tran and L. Tran, "Un-normalized graph p-Laplacian semi-supervised learning method applied to cancer classification problem," in *Proc. The Second International Conference on Intelligent and Automation Systems*, 2014.

[16] K. Tsuda, H. H. Shin, and B. Schoelkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, pp. 59-65, 2005.

[17] D. Zhou and B. Schölkopf, "Discrete Regularization," *Semi-Supervised Learning*, MIT Press, Cambridge, MA, pp. 221-232, 2006.

[18] S. Erten, G. Bebek, R. Ewing, and M. Koyuturk, "DADA: Degree-aware algorithms for network-based disease gene prioritization," *BMC BioData Mining*, vol. 4, no. 19, 2011.

Hieu Le completed the bachelor of technology in electronics and communications at Ho Chi Minh University of Technology in 2011, and completed the master of science in communication engineering at National Chiao Tung University in 2013. Currently, he's a researcher in IC Design Lab at Hochiminh University of Technology.

Hoang Trang completed the bachelor of science and master of science at Ho Chi Minh City University of Technology in 2002 and 2004 respectively. He completed his PhD degree in 2008. Currently, he's a lecturer at Ho Chi Minh City University of Technology.

Loc Tran completed the bachelor of science and master of science in computer science at University of Minnesota in 2003 and 2012 respectively. Currently, he's a PhD student at University of Technology, Sydney.

Linh Tran completed his bachelor of science and master of science in electrical and computer engineer at Portland State University. Currently, he's a PhD student at Portland State University.