ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(5.9)

Computational Prediction of Aging-Related Proteins Using Machine Learning Algorithm

Siti Raihan Halilan[1*] and Sakhinah Abu Bakar[2]

[1,2]*School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor.*

Human aging is the main social and medicinal challenge as it is among the important factor lead to many diseases. Exploring the underneath of the aging process will give advantages for human. As the network of protein-protein interaction (PPI) in this study is dynamic, we are able to observe topological changes of network over time and make prediction of the key players of aging. The dynamic networks consist of thirteen elderly ages, they are within 80 – 99 years. Integration of static protein-protein interaction network data with microarray human brain gene expression is an approach to construct the dynamic network. The network topological features are then will be used to carry out prediction using Support Vector Machine (SVM), artificial neural network (ANN) and *K*-Nearest Neighbour (KNN) models. This paper aims to provide a computational approach within the aging related proteins prediction by network topological features. SVM outperform the two others method in predicting aging-related proteins by reaching Area Under ROC Curve (AUC) value of 74.6%.

Keywords: Aging, Dynamical Network, PPI, SVM, RBF, KNN

I. INTRODUCTION

Aging is a phenomenon in all living things and a fundamental biological process across a life span. Aging process is continuing from younger to older people as many factors contribute to accelerate or decelerate the process. However, this biological process is poorly understood and therefore, aging-related genes gain more attraction from academician and medicinal field. Studies on aging in recent years are highly increasing as the awareness of relation of disease and aging are connected and therefore will be advantageous in the drugs and medicinal fields. Aging is a process where organs and tissues in our body physiologically lose its integrity where our body starts to lose coordination between parts and the systems begin to malfunction over time (Soltow et al., 2010). Understanding of aging process is very challenging, therefore, with the extensively increasing of biological data nowadays, aging becomes a needed field to study using the existed data by computational approach (Fabris et al., 2017).

Study by Kenyon (2010) shows that aging has a strong genetic components (Kenyon, 2010). Therefore, the identification of genes that are related to aging process is very helpful to reveal the underneath of the process. Distinguished feature of aging-related proteins are called non-aging-related proteins will also help us in understanding the mechanisms of aging process (Kenyon, 2010). In protein function, there will be no protein functions or working alone as it is working together with others. It is shown that topological features in network representation of PPI helps a lot in identifying characteristics of aging-related proteins (Li et al., 2010). The aging-related proteins have *i)* higher degree centrality *ii)* higher number of aging-related proteins neighbours *iii)* centralized in PPI network *iv)* higher correlation coefficients

*Corresponding author's e-mail: sitiraihan_srh@yahoo.com

with other genes (Li et al., 2010). However, to our knowledge, most of study deals with static representation of PPI network, there are no study on computational prediction of dynamical network representation of PPI network in investigating aging-related proteins.

In this study, we apply the computational approach in predicting human aging-related proteins from dynamical network of aging. The static PPI network downloaded from DIP, we integrate the network with genes expression profiles to get the dynamical network of PPI. From the dynamical network, the network measurements are calculated and then used as features in prediction of aging-related protein using Support Vector Machine (SVM), Radial Basis Function Neural Network (RBFNN) and *K*-Nearest Neighbour (KNN).

II. MATERIALS AND METHODS

A. Data Sources

- Aging-related gene expression data.

We use a microarray human brain gene expression data set containing 173 samples from 55 individuals, with 13 different ages between 80 until 99 years (Berchtold et al., 2008).

- PPI data set.

Human PPI data set is obtained from Database of Interacting Protein (DIP) consist of 7794 PPI with self-interactions as downloaded on 31/7/2016 (Salwinski et al., 2004). However, after considering the unique protein with no self-interactions, the number of PPI reduced to 7285. The id used in this study is uniprot kb.

- Dynamic age-specific network (Fig. 1).

Static PPI data was integrated with gene expressions data to form dynamic age-specific networks.

B. Development of Dynamic Network

Microarray human brain gene expression data is obtained via Affymetrix Hg-U133plus 2.0 microarray experiments using 54675 probes in each individual sample. The probe is then converted into protein id using uniprot mapping ids for mapping the probes into PPI network.

Detection of *p*-value < 0.04 is used to determine whether the probe is expressed(Lu et al., 2004) via Presence-Absence calls with Negative Probesets (PANP) in R software package that outperform MAS-P/A in gene detection method (Waren et al., 2007). Based on majority vote rule, gene m will be considered as expressed at age n if more than 50% of $x \times y$probes are found to be expressed at age n (or at least $\frac{x \times y}{2} + 1$ probes)(Faisal & Milenković, 2014). Figure 1 illustrate the integration of static PPI with gene expression data. Dynamic network in this study consist of 13 subnetworks. Size of each network is within range of 913 – 1156 interactions with 883 – 1027 proteins.

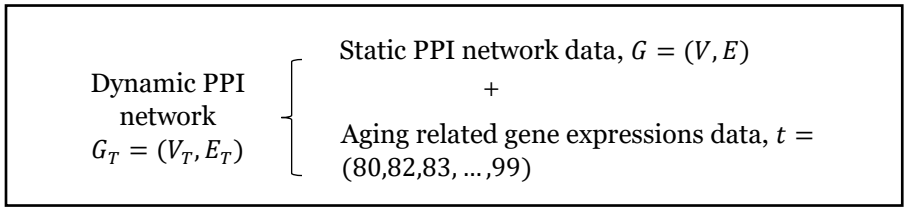

Figure 1. Integration of gene expression data with static PPI network data

C. Network Feature Computation

The computation of network topological features is Degree Centrality (DC), Closeness Centrality (CC), Betweenness Centrality (BC), Local Clustering Coefficient (LCC), Eigenvector centrality(Newman, 2010)and Local Assortativity (LA) (Piraveenan et al., 2012). Table 1 shows the details of the features.

Table 1. Network Measurements

Name	Function	Description
Degree	K_i	K is a number of neighbour. K_i is the number of neighbors of node i.
Closeness Centrality	$\frac{n}{\sum_j d(v, v_j)}$	n is the number of nodes while $d(v, v_j)$ are the shortest path from v to v_j.
Betweenness Centrality	$\sum_{j \neq v \neq k \in V} \frac{g_{jk}(v)}{g_{jk}}$	$g_{jk}(v)$ is the number of shortest paths between protein j and k while g_{jk} is the number of shortest paths between protein j and k that passes through protein v.
Local Clustering Coefficient	$\frac{E_v}{[k_v(k_v-1)]/2}$	E_v is the number of pairs of v's neighbors that are connected. k_v is the number of neighbours of v or degree of v.
Eigenvector	$x_i = \sum_j A_{i,j} x_j$	$A_{i,j}$ is an element of adjacency matrix and x is a non-zero vector (centrality of each vertex i)
Local Assortativity	$\frac{j(j+1)(\bar{k} - \mu_q)}{2n\sigma_q^2}$	$\bar{k}$ is the remaining degree of node's neighbor, n is the number of links in the network, μ_q and σ_q are the mean and standard deviation of remaining degree distribution of a network while *q(k)* is given by $q(k) = \frac{(k+1)p(k+1)}{\sum_j jp(j)}$ where j and is the excess degrees of protein v while k is the excess degree of protein v's neighbor

D. Support Vector Machine (SVM)

Binary classification of SVM with Radial Basis Function (Jiang & Ching, 2011)was chosen as underlying kernel function in this study. A radial basis function is a kernel that map data from single dimension vector to higher dimension vector space to classify aging related proteins or non-aging related proteins. SVM perform classification in prediction of aging-related proteins by constructing hyperplanes in multidimensional space that splits the aging-related or non-aging related protein. The software MATLAB is employed in this study for all methods.

E. Radial Basis Function Neural Network (RBFNN)

Neural network or artificial neural network is a non-linear computing system that inspired by biological neural network (Rajan & Kaur, 2016). In this study, radial basis function is used as activation function in neural network with combination of linear function. RBFNN consist of three layers, they are input, hidden layer and output. Each layer consists of nodes and each of the nodes are connected to previous layer. Input variables are directly passed to hidden layer without weight (Wang et al., 2010). The output can be obtained by

$$y(x) = \sum_{i=1}^{c} w_i \phi_i(x) \quad (1)$$

where c is the number of hidden layer neuron while w represents weights and ϕ is the radial basis function which considered as Gaussian function,

$$\phi_i x = \exp\left(-\frac{(x-\mu_i)^2}{2{\sigma_i}^2}\right). \quad (2)$$

F. K-Nearest Neighbour (KNN)

In KNN(Li et al., 2010), K=9 is chosen in this study. In this method, a protein is allocated to the class most common among its K nearest neighbour. K=3, 5, 7 and 9 is tested and K=9 gives a better prediction compare to others. In this case of binary classification, an odd value of K is chosen because the majority vote rule for the classes might be in tie if the even number of K is chosen. The Euclidean distance,

$$d_{st} = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3)$$

is used to determine a protein's *K* nearest neighbour.

G. Cross Validation

In this study, 10-fold cross validation(Feng et al., 2012) is adopted in each methods for assessing performance. The data is divided into 10 equal subsets and in each iteration, a single subset retained as test set while the remaining subsets are used as training set. AUC value in prediction by SVM for all involved ages are above70% while the AUC value of other two methods are lower than 62%.

Prediction accuracy is measured regarding to AUC (Area Under ROC Curve) value by 10-fold cross validation (Jiang & Ching, 2011). After the prediction is done, the number of true positive (*TP*), false positive (*FP*), true negative (*TN*) and false negative (*FN*) are counted to plot ROC curve. ROC curve is a plot of true positive rate (TPR) against false positive rate (FPR)(Sonego et al., 2008). TPR also known as sensitivity and calculated as

$$TPR(sensitivity) = \frac{TP}{TP + FN}, \quad (4)$$

while FPR is calculated as

$$FPR = 1 - specificity = \frac{FP}{FP + TN}\ . \quad (5)$$

ROC curve is plotted at various thresholds and the higher value of AUC indicate the better predictive ability of the methods.

III. RESULTS AND DISCUSSION

A. Prediction Methods of Aging-Related Proteins

Based on the six features of network measurements, the aging-related proteins are predicted using three methods through the dynamic network mentioned above. Figure 2 shows the performance of SVM, RBFN and KNN in each of the network. The prediction performance of AUC value in the three methods in the ages are slightly different within range 70.1% - 74.6% for SVM, 55.6% - 61.7% for RBFNN and 51.4% - 53.5% for KNN. The AUC value for all methods in each age close to each other indicate the methods are stable in predicting aging-related proteins using the group of six network measurements. A method of learning algorithm is said to be stable if the prediction is consistent when we change the training set (Ji et al., 2014; Nogueira et al., 2018).Hence, as the features used in this study gives a stable prediction methods, then the features are suitable in predicting aging-related proteins.

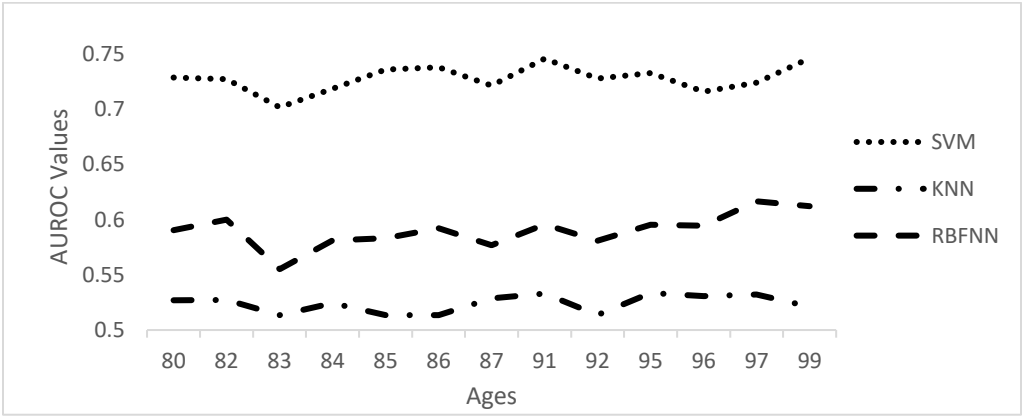

Figure 2. AUC values in respective ages

Figure 2 shows the values of AUC in each of the respective ages for SVM, KNN and RBFNN. In each age, SVM outperform the two other methods with above 70% of AUC value. For all of the methods, with the same standard of 10-fold cross validation utilized for training and testing, RBFNN and KNN are able to attain below 61.7% and 55% of AUC value. The highest accuracy for SVM reach up to 74.6%. Even the lowest AUC value for SVM is around 70.1% which much higher than the other two methods that can only reach highest AUC value on 55.6% and 51.4% for RBFNN and KNN, respectively. It is a significant improvement in SVM that reach AUC value much higher than both of the other methods. Higher AUC value shows the better predictive ability of the method (Jiang & Ching, 2011).

Instead of studying of aging in static network, this study considering aging process through thirteen ages of human as the process is continuing with age. Therefore, from the results of prediction performance methods in Figure 2, it shows that SVM are suitable methods for each age because the accuracy prediction in each age is highest among the three methods. The highest accuracy for SVM among the ages is 74.6% in age 99, for RBFNN the highest accuracy is 61.7% in age 97 while KNN can only reach highest at 53.5% in age 95. Therefore, SVM show a significant increase in prediction performance of AUC value. On the other hand, network features used in this study bring an important impact on aging field as it is showing a stable prediction performance not only in SVM but in RBFNN and KNN.

Previous study on aging prediction with the same method, SVM by using topological features are listed in Table 2. Based on the studies, by comparing methods used in prediction, they also found that SVM outperform any other methods used. For example, study by Li et al. (2010) on prediction of aging-related genes in *Caenorhabditis elegans* found out that SVM with RBF as kernel function outperform KNN, decision tree, and SVM as linear and polynomial as kernel function (Li et al., 2010). Study by Feng et al. (2012) was successfully predicts 110 new candidates of aging genes of *Mus musculus* using SVM(Feng et al., 2012) while with the same method, Song et al. (2012) predicted 51 new candidates of aging genes in *Drosophila melanogaster*.

Table 2. Previous study on aging prediction

	Topological Feature Used	Species
Li et al. (2010)	Degree, betweenness, neighbour ratio, shortest distance, clustering coefficient, *K*-core	*Caenorhabditis elegans*
Feng et al. (2012)	Degree, betweenness, neighbour ratio, cluster coefficient	*Mus musculus*

Table 2. (continued)

Song et al. (2012)	Degree, betweenness, *K*-core, shortest distance and clustering coefficient	*Drosophila melanogaster*
This study	Degree, closeness, betweenness, local clustering coefficient, local assortativity and Eigenvector.	*Human*

In this study, we added new feature such as LA and Eigenvector in our prediction. Other study listed in Table 2 mostly shares common network topological features. We also found out that LA helps in increase the performance of prediction as LA is one of the important features in analysing network topology. Eigenvector refers to the neighbours of important character in network (Newman, 2010), therefore,

added in as one of the features for prediction helps in predicting the aging-related proteins.

B. New Prediction of Aging-Related Proteins

Aging genes from GenAge databases is selected based on their presence in each age. For example, in age 83, total number of proteins are 866. From that number, 71 proteins are known to be aging-related proteins. Therefore, 783 proteins are not known to be aging-related proteins. From the prediction of binary in SVM as chosen method in this study, there are 144 new candidates for aging-related proteins. For these predictions, some of the proteins are supported by previous study. For example, Q92945 is one of the differentially expressed proteins in association with the aging follicle (McReynolds et al., 2012). Previous study showed that O14773 are involved in aging of human brain (Zucca et al., 2018).

IV. CONCLUSION

In this study, we successfully build a dynamic network of aging-related proteins. By using network topological features calculated from the network, we use them as features in prediction of aging-related proteins by employing SVM, RBFNN, and KNN. We successfully predict 144 new candidates of aging-related proteins and some of them have been validated in previous study.

V. ACKNOWLEDGEMENT

This research is funded by the National University of Malaysia grant, GUP-2017-118.

VI. REFERENCES

Berchtold, N. C. et al., "Gene expression changes in the course of normal brain aging are sexually dimorphic.," Proc. Natl. Acad. Sci. U. S. A., vol. 105, no. 40, pp. 15605–15610, 2008.

Fabris, F., Magalhães J. P. de and Freitas A. A. (2017). "A review of supervised machine learning applied to ageing research," Biogerontology, vol. 18, no. 2, pp. 171–188.

Faisal, F. E. and Milenković, T. (2014). "Dynamic networks reveal key players in aging," Bioinformatics, vol. 30, no. 12, pp. 1721–1729.

Feng, K., Song, X., Tan, F., Li, Y. H., Zhou, Y. C., and Li, J. H. (2012). "Topological analysis and prediction of aging genes in Mus musculus," 2012 Int. Conf. Syst. Informatics, ICSAI 2012, no. Icsai, pp. 2268–2271.

Ji, Z., Z. Lipton, C., and Elkan, C. (2014). "Differential Privacy and Machine Learning: a Survey and Review," CoRR, vol. abs/1412.7, pp. 1–30.

Jiang, H. and Ching, W.-K. (2011). "Classifying DNA repair genes by kernel-based support vector machines." Bioinformation, vol. 7, no. 5, pp. 257–263.

Kenyon, C. J. (2010). "The genetics of ageing," Nature, vol. 464, no. 7288, pp. 504–512.

Li, Y. H., Dong, M. Q., and Guo, Z. (2010). "Systematic analysis and prediction of longevity genes in Caenorhabditis elegans," Mech. Ageing Dev., vol. 131, no. 11–12, pp. 700–709.

Li, Y. H., Zhang, G. G., Guo, and Z. (2010). "Computational prediction of aging genes in human," 2010 Int. Conf. Biomed. Eng. Comput. Sci. ICBECS 2010, pp. 0–3, 2010.

Lu, T. et al. (2004). "Gene regulation and DNA damage in the ageing human brain," Nature, vol. 429, no. 6994, pp. 883–891.

McReynolds, S. et al. (2012). "Impact of maternal aging on the molecular signature of human cumulus cells," Fertil. Steril. vol. 98, no. 6, p. 1574–1580.e5.

Newman, M. (2010). Networks: An Introduction. New York: Oxford university press.

Nogueira, S., Sechidis, K., and Brown, G. (2018). "On the Stability of Feature Selection Algorithms," J. Mach. Learn. Res., vol. 18, pp. 1–54.

Piraveenan, M., Prokopenko, M., and Zomaya, A. (2012). "Assortative mixing in directed biological networks," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 9, no. 1, pp. 66–78.

Rajan, D. and Kaur, L. (2016). "Applications of Artificial Neural Networks: A Review," Indian J. Sci. Technol., vol. 9, no. 47.

Salwinski, L., Miller, C. S., Smith A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). "The Database of Interacting Proteins: 2004 update," Nucleic Acids Res., vol. 32, no. Database Issue, pp. D449–D451.

Soltow, Q. A., Jones, D. P., and Promislow D. E. L., (2010). "A network perspective on metabolism and aging," Integr. Comp. Biol., vol. 50, no. 5, pp. 844–854.

Sonego, P., Kocsor, A., and Pongor, S. (2008). "ROC analysis: Applications to the classification of biological sequences and 3D structures," Brief. Bioinform. vol. 9, no. 3, pp. 198–209.

Song, X., Zhou, Y. C., Feng, K., Li, Y. H., and Li, J. H. (2012). "Discovering aging-genes by topological features in Drosophila melanogaster protein-protein interaction network," Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW 2012, pp. 94–98.

Wang, B., Chen, P., Wang, P., Zhao, G., and Zhang, X. (2010). "Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes," Protein Pept Lett, vol. 17, no. 9, pp. 1111–1116.

Warren, P., Taylor, D., Martini, P. G. V., Jackson, J., and Bienkowska, J. (2007). "PANP - A new method of gene detection on oligonucleotide expression arrays," Proc. 7th IEEE Int. Conf. Bioinforma. Bioeng. BIBE, pp. 108–115.

Zucca, F. A. et al. (2018). "Neuromelanin organelles are specialized autolysosomes that accumulate undegraded

proteins and lipids in aging human brain and are likely involved in Parkinson's disease," npj Park. Dis., vol. 4, no. 1, p. 17.