An Enhanced Prediction Model for Essential Proteins Prediction for Human Diseases

D. Narmadha, A. Pravin

Abstract: ***Proteins play an important role in human biological system. Proteins interact with other molecules such as DNA, RNA and other proteins to perform biological activities. Essential proteins are indispensable for the survival of an organism. The identification of essential proteins is important for finding the disease treatment, develop novel drugs. Numerous topological and machine learning approaches have been introduced in recent past for essential protein prediction but they have not attained promising results. In order to improve the prediction accuracy of essential protein identification the proposed prediction model is constructed by incorporating graph coloring and machine learning approaches. Numerous performance measures namely accuracy, precision, recall and f-measure were employed to predict the performance of the proposed model. After analysis, it is identified that the proposed model produced promising results as compared to state-of art methods.***

Keywords: Classification algorithms, Decision tree, Graph Coloring, Protein-protein interaction, Random Forest, SVM.

I. INTRODUCTION

Essential proteinsplay a key important role in the cellular functions such as regulations of body issues, biochemical reactions and perform regulation of body processes to maintain fluid balance in our body[1]. In recent times, the researches on essential protein identification have attracted many researchers. Essential proteins are those proteins which are crucial for drug development, disease diagnosis and treatment. In previous years, both experimental and computational based approaches have been most widely employed for essential protein identification [2]. The experimental based approaches such as gene knock out[3], RNA interference [4]and conditional gene knock out[5] are extremely timeconsuming. In the meantime, with the rapid generation of vast amount of biological data, many computational based approaches have been proposed for the identification of essential proteins based on topology of protein network and machine learning based methodologies such as supervised and unsupervised approaches. Many insight have been provided by the researchers on finding the essential proteins based on network topology such as degree centrality (DC)[6], eigenvector centrality(EC)[7], edge clustering(EC)[8], closeness centrality(CC)[9], betweenness centrality(BC)[10]. However, presently available computational based approaches result in high false negative and low true negative.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)

D. Narmadha, Research Scholar, Sathyabama Institute of Science and Technology, Chennai, India.
D. Narmadha, Assistant Professor, Karunya Institute of Technology and Science, Coimbatore, India.
Dr. A.Pravin, Assistant Professor, Sathyabama Institute of Science and Technology, Chennai, India.

During the past, many machine learning based approaches have been applied effectively applied to PPI to find essential proteins. In the supervised approach, a training dataset includes input as well as output response values. Some of the most commonly used approaches are logistic regression, decision tree, random forest, NB classifier and KNN classifier. The input for building the classification model can be of either sequential or structural features based extraction.Yu Dong et al.[10] gave insight on using SVM classifier for predicting structural classes of protein such as . Long-Hui Wang and Juan Liu[11] examined the impact on predicting protein using physical chemical properties and structural properties of amino acids.HaijiangGeng et al.[12] have developed NB classifier based approach that used two distinct features such a PSSM(Position Specific Score Matrix) and Relative Solvent Accessibility for predicting the protein binding site.Jian Cheng et al. [13] developed a combination of machine learning model (FMW) based on logistic regression, NB classifier and genetic algorithm.FMW based model produces better accuracy that using single machine learning model.Santosh Philips et al.[14] tested the importance of machine learning algorithms in predicting the essential genes for diagnosing diseases and treating it. They employed algorithms such as J48, SMO and Naïve Bayes for predicting essentialgenes.

Marcio L Acencio andNey Lemke[15] have developed a combinatorial framework of machine learning and topological approach for essential protein prediction. They used 12 network topological features for prediction. A series of research have been proposed in [16-18] that portrayed the importance of combining topological and biological features in identifying the essential proteins.

In this review, we proposed a powerful and accurate prediction model using graph coloring and machine learning model. In this proposed architecture, Protein interaction data is extracted from stringDB and graph coloring approach has been employed to extract the essential proteins by extracting the primary and secondary colors. Secondly, physiochemical properties of the proteins are used to extract the most essential proteins. Performance measures such as precision, recall, f-measure are utilized to evaluate the performance of the proposed architecture.

II. MATERIALS AND METHODS

Datasets:

In this research, we have collected four different benchmark datasets for various diseases such as cancer, diabetes, allergy and NIPAH virus. There are various sources of protein interaction databases such as stringDB[19], bioGRID[20], IntACT[21]. The dataset was constructed by retrieving all the interactions of a specific protein from stringDB.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Further, to predict the essentiality of protein the physiochemical properties of individual proteins were constructed from UniProtKB using the identifier.

Feature Extraction Methods:

Physiochemical properties of the proteins were used to extract the prominent features of the protein which are used as training and testing the dataset.

Framework of proposed Architecture:

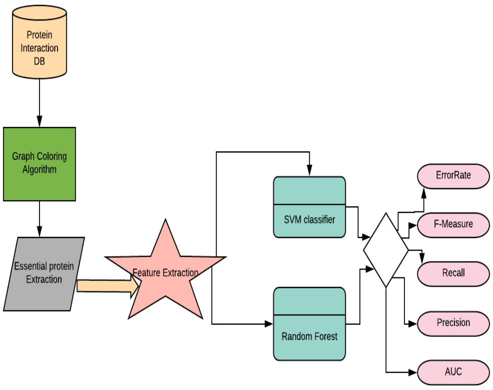

Fig. 1: Framework of proposed Architecture

Fig. 1 shows the proposed framework for extracting essential protein from protein interaction database. The dataset from our experimental study was collected from stringDB. StringDB is a biological database which holds information about protein-protein interaction, functions of protein and huge collection of experimental data. Nodes represent protein and edges represent the interaction between proteins. In the graph, graph coloring algorithm has been employed by checking the connectivity between neighbors. The proteins are separated based on colors and the algorithm extracts the primary and secondary layers of interaction from the main target protein. Furthermore, the physiochemical properties of the proteins are obtained using bioinformatics tool and most essential proteins are further analyzed using SVM and random forest classification algorithms.

2.3.1 Graph Coloring:

Fig. 2 shows the outcome of graph coloring algorithm. From the PPI, start at a chosen protein and check whether it is safe to color the adjacent proteins of the currently picked target protein. Perform three condition checks at the currently selected protein. Firstly, there may not be any proteins adjacent to the currently chosen protein. Secondly, there may be adjacent proteins which are already colored. Thirdly, there may be adjacent proteins which are not colored. The first and second condition need not be considered for coloring. The third condition can be checked, by using the formula’s (1) and (2). It is used to find the proteins that are adjacent to the target protein and to check whether it is already colored.

$$G[p][j] == 1 \qquad (1)$$

$$C1 == x[j] \qquad (2)$$

For all the set of the proteins in a PPI, compute its color. Top percentages of essential proteins for diseases are computed by extracting the primary and secondary colors from PPI.

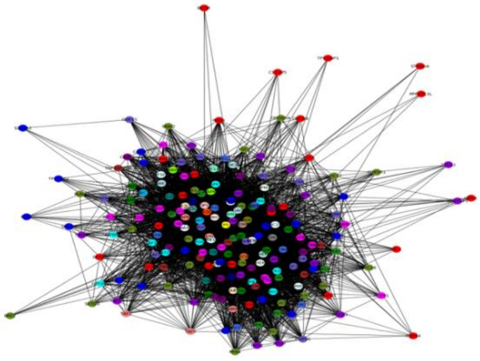

Fig. 2: Framework of Graph Coloring

2.3.2 SVM Classifier:

SVM classification algorithm is most popularly used algorithm in the field of bioinformatics for pattern recognition, classify diseases. It was developed by HavaSiegelmann and Vladimir Vapnik[22] for classification problems. The main objective of the algorithm is to finds a hyper plane that best divides the plane into two classes. The plane maximizes the distance between nearest data points of either side of classes.

2.3.3 Random Forest Classifier:

Random Forest is one of the most popularly used machine learning algorithm which makes uses a combination of decision trees to support strong decision making process. The forest chooses the classification having the most votes. It plays a wide important role in many applications such as bioinformatics, banking, marketing and Medicine. The first basic decision tree algorithm was developed by Tin Kam Ho[23], [24]. Later, an extension to the approach was given by Leo Breiman and Adele Cutler[25].

I. Evaluation Methodology:

Numerous performance parameters such as accuracy, precision, recall, F-measure and AUC have been utilized to measure the performance of the proposed predictor. In this test, protein interaction data have been collected from stringDB. A very large collection of proteins are given as input to the graph coloring algorithm. The most essential proteins for various diseases have been extracted through primary and secondary colors. In the next phase of the prediction, physiochemical properties of the proteins have been collected with the Uniprot ID. A 70% of the data have been assigned as training data and the remaining as test data. The evaluation metrics are calculated using the following formulas;

$$\text{Accuracy} = \frac{TN + TP}{TP + FN + FP + TN} \qquad (3)$$

$$\text{Precision} = \frac{TN}{TP + TN} \qquad (4)$$

$$\text{Recall} = \frac{TP}{FN + TP} \quad (5)$$

$$\text{F-measure} = 2\times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

where TP, TN, FP, FN represents true positive, true negative, false positive and false negative respectively. Accuracy, precision, recall and F-measure are computed using the formulas (3), (4),(5) and (6). Accuracy helps to measure the ratio of correctly classified essential proteins to the total proteins. Precision helps to measure the ratio of correctly classified essential proteins to the total number of essential proteins. Recall is the measure of the ratio of correctly classified essential proteins to all the observations in actual essential proteins.

III. RESULTS AND DISCUSSION

Table 1 describes the success rate of the proposed graph coloring and machine learning approaches for protein interaction data from stringDB and the physiochemical property based features for various diseases such as cancer, diabetes, allergy and NIPAH virus. After analyzing the proposed approach with varying window size such as 1001,1003,1005,1007 the proposed graph coloring and machine learning approach has attained a highest accuracy rate for the window size 1001. The accuracy, precision, recall and F-measure are 92.3 %, 93% and 92% and 91% respectively. The accuracy for SVM classifier is high for the window size 1007. The better accuracy for random forest has been attained for the window size 1007. The highest accuracy rate for decision tree is 86.7%. For NB classifier the accuracy, precision, recall and F-measure are 79.4%, 76%, 78% and 79% respectively. The success rate of the proposed approach is better than other classification algorithms such as SVM, Random forest, decision tree and NB classifier.

Table 1: Success rate of Graph coloring and classification algorithm for Cancer disease

Window size	Prediction Model	Acc(%)	Precision(%)	Recall(%)	F-measure(%)
1001	Graph Coloring + SVM + Random Forest	**92.3**	**93**	**92**	**91**
1003		91.4	92.2	91.2	91.4
1005		92.7	91.2	90	92.3
1007		91.8	92.3	90.6	94.4
1001	SVM Classifier	80.2	73.4	80	76.6
1003		81.3	74.4	81.2	77
1005		82.5	76.6	83	78
1007		83.3	78	84.5	79
1001	Random Forest	84.6	85	84.8	86
1003		86.6.	87.1	86.7	87.2
1005		88	89.2	87	88
1007		89.8	89	88.2	89
1001	Decision Tree	84.2	82.8	83.8	86
1003		83.3	83.3	84.7	87.2
1005		86.7	86.7	87.6	88
1007		85	88	88.5	86.4
1001	NB classifier	77	78.2	79	75.4
1003		76.2	78.8	78.2	76.6
1005		78	77.4	78.4	78
1007		79.4	76	78	79

Table 2 shows the success rate of the proposed graph coloring and machine learning based model for diabetes disease. The analysis of the proposed approach with varying window size has been recorded in the below mentioned table. The accuracy for the disease diabetes is high for the window size 1005. The accuracy, precision, recall and F-measure are 93.5%, 94.8%, 90.6% and 93.3% respectively. The prediction accuracy rate is high for the window size 1007 for the SVM classifier algorithm. For random forest the accuracy rate is high for the window size 1007. For decision tree classifier, the accurate rate is high for the window size 1007.The accuracy, precision, recall and F-measure of the proposed graph coloring and machine learning based model are 94.7%, 94.3%, 92% and 93.2% respectively.

Table 2:Success rate of Graph coloring and classification algorithm for diabetes disease.

Window size	Prediction Model	Acc(%)	Precision(%)	Recall(%)	F-measure(%)
1001	Graph Coloring + SVM + Random Forest	93	92	91.8	92
1003		92.7	93	93.1	92.4
1005		93.5	94.8	90.6	93.3
1007		94.7	94.3	92	93.2

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

1001	SVM Classifier	80.2	73.4	80	76.6
1003		81.3	74.4	81.2	77
1005		82.5	76.6	83	78
1007		83.3	78	84.5	79
1001	Random Forest	84.6	85	84.8	86
1003		86.6.	87.1	86.7	87.2
1005		88	89.2	87	88
1007		89.8	89	88.2	89
1001	Decision Tree	84.2	82.8	83.8	86
1003		83.3	83.3	84.7	87.2
1005		85	86.7	87.6	88
1007		86.7	88	88.5	86.4
1001	NB classifier	78.4	79	79.8	76.4
1003		77	79.8	76	76.2
1005		78.9	76.4	76.4	79
1007		79.2	79.2	77.8	78.3

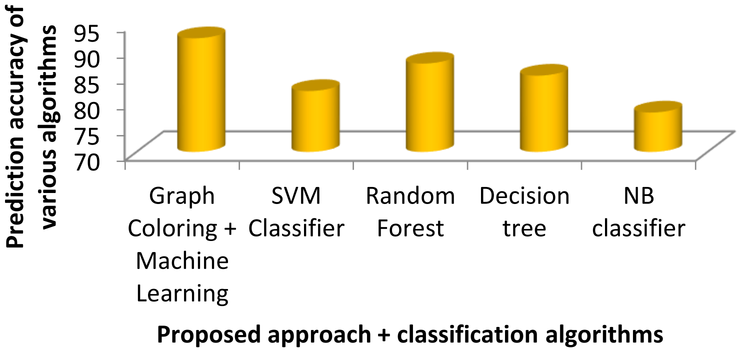

Figure 3: Comparison of prediction accuracy of proposed approach and classification algorithms

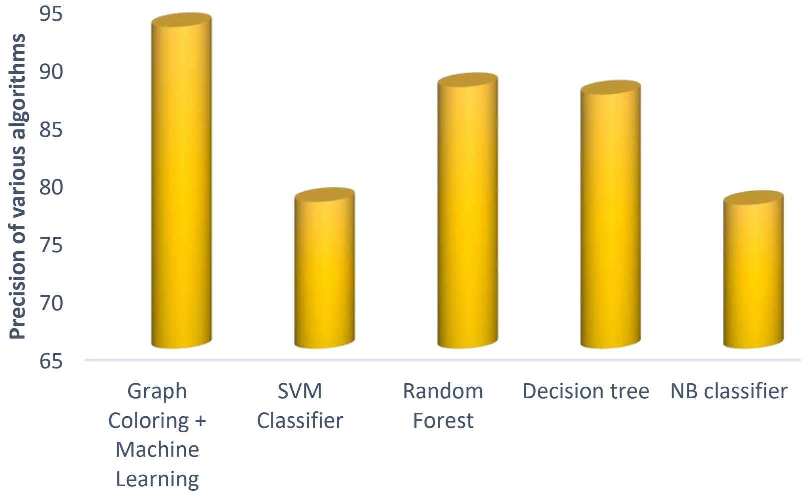

Figure 4: Comparison of precision % of proposed approach and classification algorithms

Figure 4 shows the comparison of the precision % of the proposed approach with existing state of art methods. The proposed method produces about 16% increase in performance than SVM classifier, 7% increase than random forest, 9% increase in performance than decision tree and 17% increase in precision performance than NB classifier.

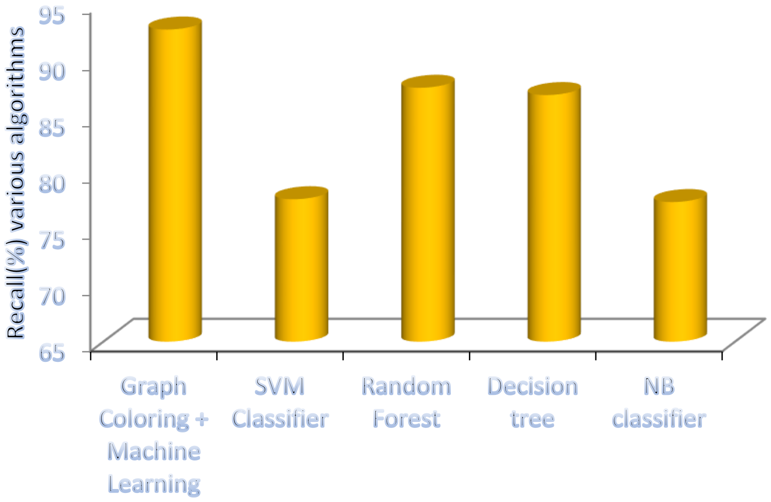

Figure 5: Comparison of recall(%) of proposed approach and classification algorithms

Figure 5 shows the comparison of recall in % than existing machine learning models. It is observed that the proposed method produces about 16% increase than SVM classifier, 7% increase than random tree, 10% increase than decision tree and 18% increase than NB classifier.

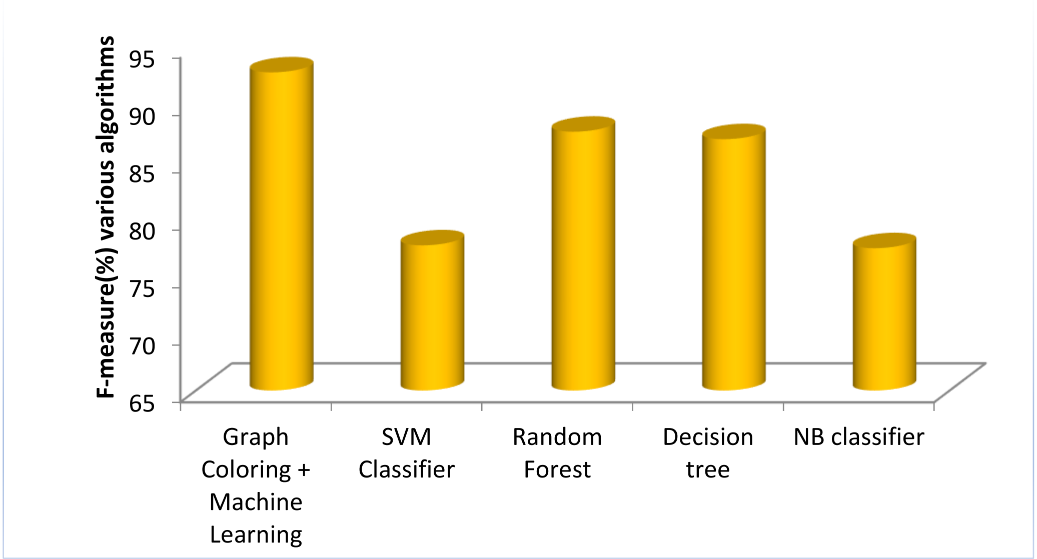

Figure 6: Comparison of F-measure(%) of proposed approach and classification algorithms

Figure 6 shows the performance comparison of the proposed method than existing machine learning model. It is clearly observable that the proposed method produces an increase in % of recall of about 17% than SVM classifier, 9% than Random Forest, 10.5% than decision tree and 18.5 % than NB classifier.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

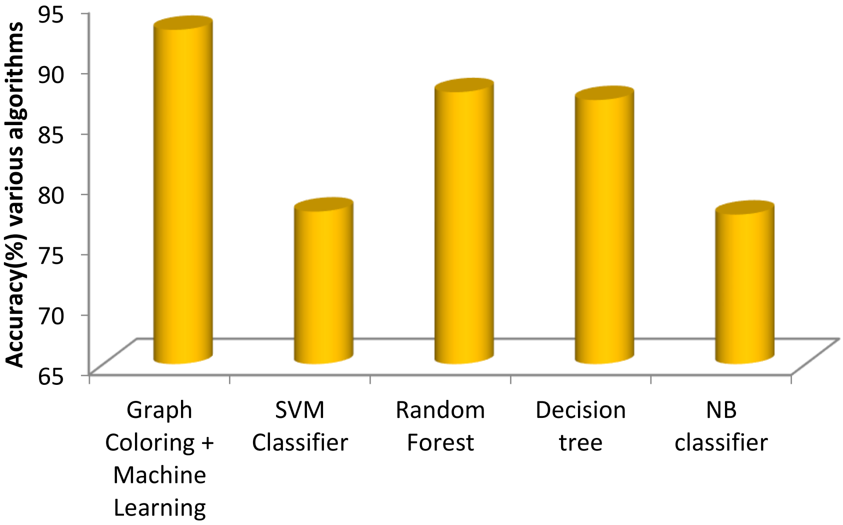

Figure 7: Comparison of prediction accuracy(%) of proposed approach and classification algorithms

Figure 7 shows the performance of the prediction accuracy of the proposed model with the existing machine learning model for diabetes disease. It is observable that the proposed approach produces about 18.5% increase than SVM classifier, 9% increase than random forest, 8.4% increase than decision tree and 19%increase than NB classifier.

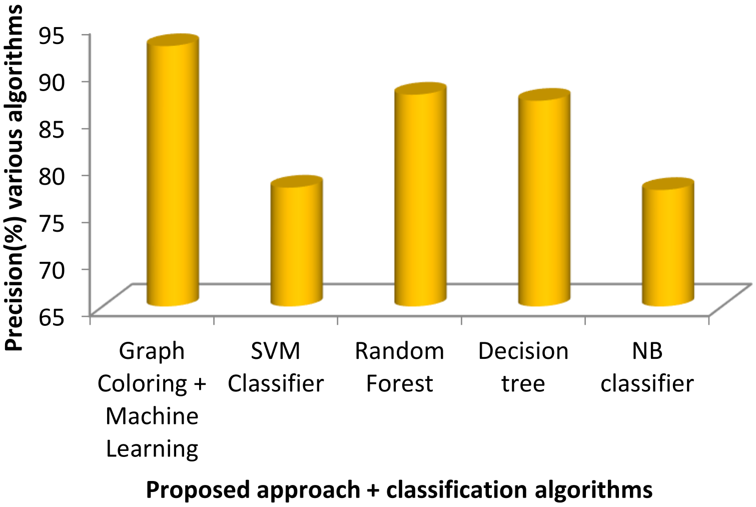

Figure8: Comparison of precision(%) of proposed approach and classification algorithm

Figure 8 shows the performance of the proposed approach with the existing machine learning model. It is observable that the proposed approach works better than SVM for about 19.2%, 8.3% increase in precision than Random forest, 7% increase than decision tree and 18% increase than NB classifier.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

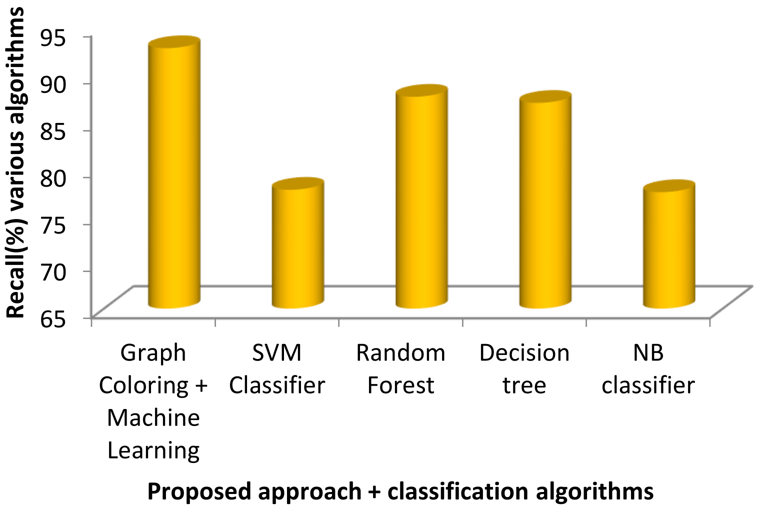

Figure 9: Comparison of Recall(%) of proposed approach and classification algorithms

Figure 9 records the comparison of the proposed method with the existing classification algorithms. It is observable that the combination of graph coloring and machine learning produces better result than the existing algorithms in terms of recall

IV. CONCLUSION

In this research an effort had been made for the accurate prediction of protein-protein interactions is developed. Two different approaches such as the combination of graph coloring and machine learning based model have been used for the accurate prediction. It is observed that the prediction outcome of the proposed approach is quite efficient for four different types of diseases than the traditional methods in the literature conducted hitherto. It is observed that the proposed prediction model will be a user friendly tool for predicting the essential proteins for different diseases. In order to provide an easy access for identifying the essential proteins, an effort shall be made in future to develop a web based interface. Thus, graph coloring and machine learning based model has been extensively applied for effective protein classification.

REFERENCE

1. Ideker, T. and Sharan, R., 2008. Protein networks in disease. Genome research, 18(4), pp.644-652.
2. Browne, F., Zheng, H., Wang, H. and Azuaje, F., 2010. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. Advances in Artificial Intelligence, 2010, p.7.
3. Galli-Taliadoros, L.A., Sedgwick, J.D., Wood, S.A. and Körner, H., 1995. Gene knock-out technology: a methodological overview for the interested novice. Journal of immunological methods, 181(1), pp.1-15.
4. Shukla, V., Coumoul, X. and Deng, C.X., 2007. RNAi-based conditional gene knockdown in mice using a U6 promoter driven vector. *International journal of biological sciences*, *3*(2), p.91..
5. Fritsch, L., Martinez, L.A., Sekhri, R., Naguibneva, I., Gerard, M., Vandromme, M., Schaeffer, L. and Harel-Bellan, A., 2004. Conditional gene knock-down by CRE-dependent short interfering RNAs. *EMBO reports*, *5*(2), pp.178-182.
6. Zotenko, E., Mestre, J., O'Leary, D.P. and Przytycka, T.M., 2008. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, *4*(8), p.e1000140.
7. Ruhnau, B., 2000. Eigenvector-centrality—a node-centrality?. *Social networks*, *22*(4), pp.357-365.
8. Wang, J., Li, M., Wang, H. and Pan, Y., 2012. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), pp.1070-1080.
9. Koschützki, D. and Schreiber, F., 2008. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, *2*, pp.GRSB-S702.
10. Cai, Y.D., Liu, X.J., Xu, X.B. and Zhou, G.P., 2001. Support vector machines for predicting protein structural class. *BMC bioinformatics*, *2*(1), p.3.
11. Wang, L.H., Liu, J., Li, Y.F. and Zhou, H.B., 2004. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Informatics*, *15*(2), pp.181-190.
12. Geng, H., Lu, T., Lin, X., Liu, Y. and Yan, F., 2015. Prediction of protein-protein interaction sites based on naive Bayes classifier. *Biochemistry research international*, *2015*.
13. Cheng, J., Wu, W., Zhang, Y., Li, X., Jiang, X., Wei, G. and Tao, S., 2013. A new computational strategy for predicting essential genes. *BMC genomics*, *14*(1), p.910.
14. Philips, S., Wu, H.Y. and Li, L., 2017. Using machine learning algorithms to identify genes essential for cell survival. *BMC bioinformatics*, *18*(11), p.397.
15. Acencio, M.L. and Lemke, N., 2009. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics*, *10*(1), p.290.
16. Chen, Y. and Xu, D., 2004. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics*, *21*(5), pp.575-581.
17. Hwang, Y.C., Lin, C.C., Chang, J.Y., Mori, H., Juan, H.F. and Huang, H.C., 2009. Predicting essential genes based on network and sequence analysis. *Molecular BioSystems*, *5*(12), pp.1672-1678.
18. Plaimas, K., Eils, R., and König, R. (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.* 4:56. doi: 10.1186/1752-0509-4-56
19. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P. and Jensen, L.J., 2010. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, *39*(suppl_1), pp.D561-D568.
20. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M., 2006. BioGRID: a general repository for interaction datasets. Nucleic acids research, 34(suppl_1), pp.D535-D539.

21. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A. and Margalit, H., 2004. IntAct: an open source molecular interaction database. Nucleic acids research, 32(suppl_1), pp.D452-D455.
22. Ben-Hur, A., Horn, D., Siegelmann, H.T. and Vapnik, V., 2001. Support vector clustering. *Journal of machine learning research*, *2*(Dec), pp.125-137.
23. Ho, T.K., 1995, August. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
24. Barandiaran, I., 1998. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*, *20*(8).
25. Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.