# Predicting Cancer-Related Proteins in Protein-Protein Interaction Networks using Network Approach and SMO-SVM Algorithm

Richard Enyinnaya
Leiden University Institute of
Advanced Computer Science

## ABSTRACT

An early diagnosis of cancer is crucial to improving the survival rate and to prolong the lives of patients. With the large amounts of medical data available in the medical field, applying data mining tools and an efficient prediction methodology to diagnose diseases can lead to useful knowledge to support medical professionals in saving lives. This paper explores genomic interactions networks, investigating protein-protein interaction networks to predict cancer related proteins using sequential minimal Optimization (SMO) for training Support Vector Machine (SVM). The WEKA software was utilized as the data mining tool, which is an open source collection of machine learning algorithms. The provided data set was studied and analyzed in order to build a useful and reliable model to predict cancer and non-cancer related proteins.

## Keywords

Data mining, Support Vector Machine (SVM), Protein-Protein Interaction (PPI), Sequential Minimal Optimization (SMO)

## 1. INTRODUCTION

Protein-protein interaction networks (PPIN) are crucial to understanding biological processes and protein functions [6, 1]. PPIN have been used to explore disease behaviors by taking advantage of the network properties of interactions. PPI become visible when two or more proteins bind together and perform a biological function [20]. Networks being an efficient abstraction of biological systems [8,10,1], when used in combination with cell networks and gene diseases, provide useful models that shows the complex relationships of human diseases and its various interactions [6,15].

PPIN have been applied in several areas such as identifying new disease genes, proteins, lung cancer [6, 1, 18]. Artificial Neural networks [14, 16] and support vector machine [12, 17, 13] have been utilized as a classification task in medical diagnosis. Lo *et al* [14] utilized artificial neural network (ANN) approach to develop computer-aided diagnosis of mammography using an optimally minimized number of input features. The result showed that he ANN with the four optimized features was significantly better than expert radiologists. Polat *et al* [17] conducted breast cancer diagnosis using least square support vector machine (LS-SVM) classifier algorithm. The obtained classification accuracy was 98.53%, utilizing LS-SVM, the obtained results show that the machine learning method can be effective in diagnosing breast cancer and point on a direction of designing a new intelligent assistance diagnosis systems using SVM. In addition, Chuang *et al* [15] applied a protein-network-based approach that identified markers not as individual genes but as subnetworks extracted from protein interaction databases. The research result subnetworks provides hypotheses for pathways involved in tumor progression.

This paper deals with discovering cancer-related proteins by applying network properties and SMO-SVM machine learning algorithm. In addition developing a model that can accurately predict and discover potential cancerous proteins to enhance early diagnosis of cancer; however large numbers of cancer-related proteins are yet to be discovered.

Properties of PPIN are employed in investigating and predicting disease genes i.e. in-degree (a node with a high in-degree is more likely to be associated with cancer protein), shortest paths connecting two pair of nodes and proximity of candidate gene pairs with a known disease gene. These properties can be utilized to predict diseases as a result of genes with similar characteristics, that are in close proximity often carry similar functions [6,10] and cluster around common neighborhoods. Furthermore, have a higher tendency to interact with each other. This approach of exploring properties of networks and its interactions have proved successful in predicting new disease proteins [6, 1].

Figure 1, NetLogo simulation model by Asymptote [7]; the normal proteins (noncancerous) blue and the cancerous red. Normal protein initially spreads out, presence of cancerous protein in contact with a noncancerous protein begins to disrupt the surrounding environment due PPIN network. The cancer progresses, as seen from the graph, and the healthy proteins declines sharply downward.

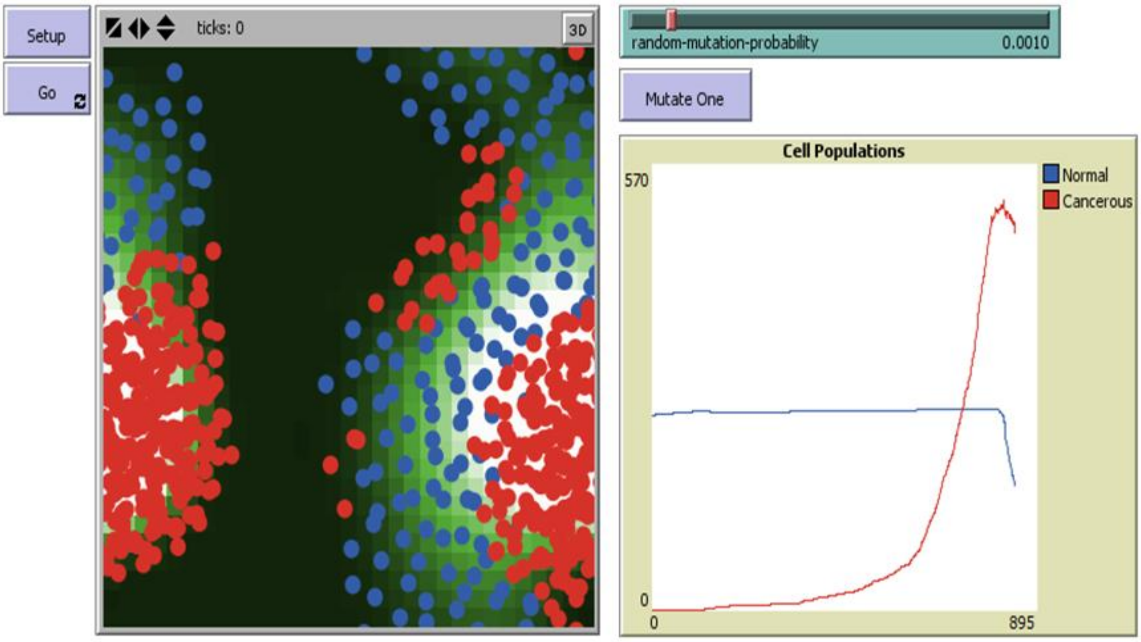

**Figure 1: Simulation of spread of cancerous and non-cancerous protein**

## 2. RESEARCH METHOD

### 2.1 Data set

The data set used in this study was anonymized provided and not available online. The data network property consists of an undirected protein-protein interaction (PPI) between two or more proteins. The humanPPI dataset served as the basis for classification distance (proximity) calculation. The human PPI dataset, functions and labelled properties were mapped and integrated to a training data set. The final instances consist of 918 labeled non-cancer proteins and 175 labeled cancer. This data set was utilized in training and various learning curve experiments in other to achieve a high performing, reliable SVM model.

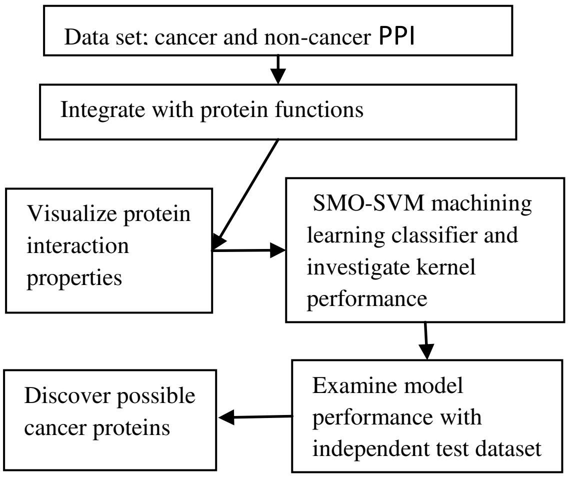
Data set: cancer and non-cancer PPI
Integrate with protein functions
Visualize protein interaction properties
SMO-SVM machining learning classifier and investigate kernel performance
Examine model performance with independent test dataset
Discover possible cancer proteins

**Figure 2: Flowchart of this study**

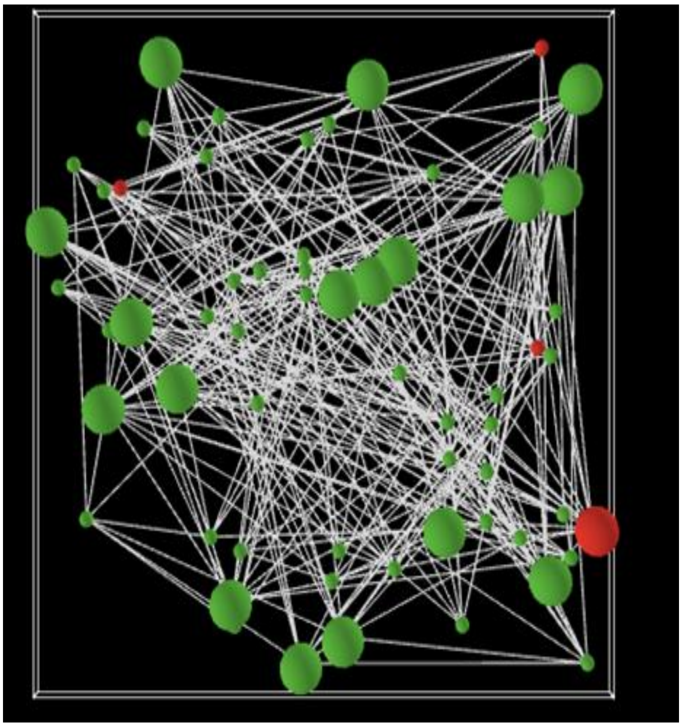

**Figure 3: Network simulation of interaction of cancer protein spread.**

### 2.2 Technique

The technique employed falls into the category of “guilt by proximity”, that is, proteins in a network that are closer to each other most likely behave the same and lead to same disease [1]. Semi - supervised learning data mining was employed; the humanPPI dataset and its respective functions was transformed into feature sets. Protein interaction network was integrated with its function, protein sequence and property (i.e cancer or not). This integrated feature dataset was trained with support vector machine (SVM) algorithm with 10 folds cross-validation which produced a model that was able to predict cancer and non-cancer proteins with high accuracy. The humanPPI served as the basis for calculating the distance between candidate proteins and disease causing proteins [1, 6]. Furthermore, the model was evaluated with an independent validation dataset that had an accuracy of 99.9% of correctly classified instances.

In Figure 3, utilizing Gephi and Yifan Hu Proportional algorithm, an algorithm that draws undirected graph with force-directed method [5]. The figure blue indicates non-cancer and red indicates cancer. Which indicates "guilt by proximity", that is proteins with the same characteristics cluster in common neighborhood [2].

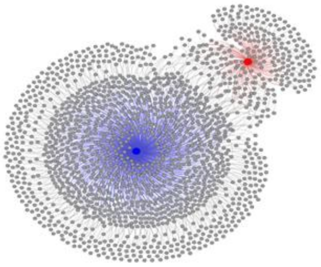

**Figure 4: Training data set visualization with Gephi using Yifan Hu Proportional algorithm**

## 2.3 Tools

WEKA, a collection of machine learning algorithms, was used for data preprocessing, training and prediction of cancerous and non-cancerous protein. Gephi an open-source software for visualizing and analyzing large networks graphs. NetLogo, also an open source software, was employed to model and learn the interaction of protein-protein interaction network and the proximity effect ("guilty of proximity") of non-cancer and cancer proteins.

## 3. TRAINING SVM USING SEQUENTIAL MINIMAL OPTIMIZATION (SMO)

Support vector machines (SVM) is a binomial classification algorithm that builds computational classification models that assign samples into two or more classes, which can be applied to prediction or diagnosis. Extensions of basic SVM algorithm such as the sequential Minimal optimization developed by John C. Platt of Microsoft research, which is applied in this research, implemented in WEKA can be used to train SVM faster, better group samples clusters based on similarity, build computational regression models to predict values of outcomes and accurate prediction of variable of interest [3,4]. SVM is fundamental because of theoretical reasoning; it is robust to a large number of variables and small samples, can learn both simple and high complex classification models, and avoids over fitting using complex mathematical principles and its reliable results [4].

The main idea of SVM is to discover a decision surface ("hyperplane") that can separate two classes with the largest distance ("gap" or "margin") within a border line (support vectors). If a decision surface is not found, data is mapped into a higher dimensional space, which is constructed via mathematical projection ("kernel trick") where separating decision surface is found.

Sequential Minimization algorithm (SMO) for training SVM is simple, faster and more scalable. SMO uses an analytical QP (quadratic programming) step, rather than numerical QP that previous methods of training SVM use. SMO spends most time in evaluating decision functions rather than executing QP, therefore it can exploit sparse data sets efficiently. SMO makes less use of matrix storage, hence very large SVM training problems can fit into the memory of a personal computer, as a result of SMO avoidance of large matrix manipulation. SMO scales between linear and quadratic in training set size of testing problems, which makes it efficient for large of amount of protein training data used in this research. SMO update two Lagrange multipliers as a SMO Step as shown below:

$$Given\ two\ examples\ E1\ and\ E2:$$

$$\alpha_2^{new} = \alpha_2 + \frac{y_2(E_2 - E_1)}{\eta} \quad \text{Where:}$$

$$\eta = K(\vec{x}_1, \vec{x}_1) + K(\vec{x}_2, \vec{x}_2) - 2K(\vec{x}_1, \vec{x}_2)$$

Clips the value at t the end of the segment:

$$\alpha_2^{\text{new,clipped}} = \begin{cases} H & \text{if} \quad \alpha_2^{\text{new}} \geq H; \\ \alpha_2^{\text{new}} & \text{if} \quad L < \alpha_2^{\text{new}} < H; \\ \text{L} & \text{if} \quad \alpha_2^{\text{new}} \leq L. \end{cases}$$

$$if\ y1 = y2\ then:$$

$$L = \max(0, \alpha_2 + \alpha_1 - C)$$

$$H = \min(C, \alpha_2 + \alpha_1)$$

Otherwise:

$$L = \max(0, \alpha_2 - \alpha_1)$$

$$H = \min(C, C + \alpha_2 - \alpha_1)$$

$$\alpha_1^{\text{new}} = \alpha_1 + s(\alpha_2 - \alpha_2^{\text{new,clipped}})$$

*Where* $s = y_1 y_2$

Major components of SMO [3] are an analytical method to solve for two Lagrange multipliers (Lagrange multiplier is a common calculus problem that is used to find maxima or minima of a function, a heuristic for choosing which multipliers to optimize and a method to compute*).*

**Table 1: Comparison of Result of SMO algorithm on different kernels**

| Kernels | | Training time (sec) | Correctly classified instances (%) | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|
| Puk | Training set | 16.57 | 83.9 | 1.0 | 0.83 | 0.90 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Validation | 15.0 | 99.9 | 1.0 | 0.99 | 0.99 |
| PolyKernel | Training set | 15.44 | 83.0 | 0.971 | 0.85 | 0.91 |
| | Validation | 10.0 | 99.9 | 1.0 | 0.99 | 0.99 |
| RBFKernel | Training set | 14.4 | 83.9 | 1.0 | 0.83 | .90 |
| | Validation | 11.0 | 90.7 | 1.0 | 0.91 | 0.95 |

## 4. RESULTS AND DISCUSSION

In this research, first the impact of kernel function on the performance of classification was examine, therefore, a comprehensive experiment to determine the optimal combinations of kernel function and SMO-SVM that gives the best prediction performance was performed. SMO training of SVM with 10-folds cross-validation with different kernel functions (Puk, Polykernel and RBFKernel). The results are presented in Table 1. Results were obtained using a training set of 2,095, and validation set of 1,005 and WEKA default parameters. This investigation was conducted to select the best kernel function based on performance (precision, recall and F-measure) as criteria. Precision, recall and F-measure are common evaluation measures utilized in evaluating machine learning performance experiments. Precision represents the proportion of predicted positive cases that are correctly real positives, recall is the proportion of real positive cases that are correctly predicted positive while F-measure references the true positives to the Arithmetic Mean of predicted positives and real positives, normalized to an idealized value [19].

$$Precision = \frac{tp}{tp+fp}$$

$$Recall = \frac{tp}{tp+fn}$$

$$F-measure = 2.\frac{precision \ \cdot recall}{Precision \ +recall}$$

The results indicates that the best kernel is polyKernel. Therefore polyKernel was selected as the choice for building a classification model, further calculations and experiments.

Secondly, the learning model accuracy was investigated as a function of training-set size. The plot Figure 4, indicates the percentage error versus the training set size by using percentage split; that is, how better does the model get at predicting the target, in consideration of number of instances used for training, decreasing the training dataset by 10%, 20%, 30%....90%. As shown in Figure 4, it can be deduced that the minimum percentage error can be attained by cutting the dataset by 90% and the highest error rate occurs if 20% of the training data set is used. However, a large training dataset will be more representative of actual characteristics of the data set.

The model that provided the most prediction accuracy based on precision, recall and F-measure and the model with the smallest difference between the training set and validation set was selected as the SMO-SVM prediction model. The best SMO-SVM prediction model and result is presented in the Table 2. The selected model indicates a precision of 100%, recall of 100% and F-measure of 99%. This result is above the baseline accuracy obtained by *Polat et al* [17] utilizing LS-SVM. Furthermore, this method can prove useful in designing effective intelligent assistance diagnosis systems to enhance cancer diagnosis.

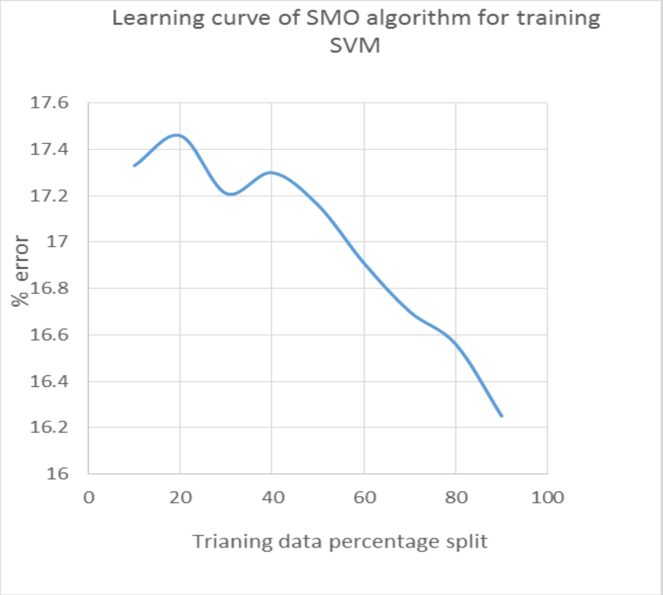

**Figure 4: Learning curve prediction accuracy as a function of percentage split of training data set**

**Table 2: Prediction power of the selected SMO-SVM model**

| | TP | FP | FN | TN | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|---|
| Training set | 892 | 26 | 161 | 14 | 0.971 | 0.85 | 0.91 |
| Validation set | 911 | 0 | 1 | 92 | 1.0 | 0.99 | 0.99 |

## 5. CONCLUSION

This work demonstrates the prediction of cancer related proteins and modeling of cancer proteins as a network and classification task. Furthermore, the implementation of using SMO for training Support vector machine. The result indicates that SMO-SVM are able to produce reliable models to effectively predict and classify cancer and non-cancer proteins using data for early diagnosis of cancer patients. It is fundamental to develop effective prediction models for cancer proteins that can be utilized as cost-effective tool for early diagnosis of cancer in the medical community to save lives. Protein networks are becoming a fundamental source of disease classification. Future research work and directions will be extend this method to other human disease networks. Network analysis can provide insightful knowledge about mechanism of cause-effect relationships of molecular interactions. In addition, an important area of application of network analysis approach is drug targeting and discovery; i.e. combination of molecular interaction networks and chemical-genetic interactions that can identify pathways of drugs reaction and effects. Exploiting networks and data mining in view of human biology can enhance disease diagnosis and treatment.

## 6. REFERENCES

[1] Wu, Xuebing, and Shao Li. "Cancer gene prediction using a network approach." Cancer Systems Biology (2010): 191-212.

[2] Lin, C., Cho, Y. R., Hwang, W. C., Pei, P., & Zhang, A. (2007). Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*, 1-35.

[3] Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).

[4] Hardin, Douglas, Isabelle Guyon, and Constantin F. Aliferis. A Gentle Introduction to Support Vector Machines in Biomedicine. World Scientific, 2011.

[5] Hu, Yifan. "Efficient, high-quality force-directed graph drawing." Mathematica Journal 10.1 (2005): 37-71.

[6] Yang, Lei, Xudong Zhao, and Xianglong Tang. "Predicting Disease-Related Proteins Based on Clique Backbone in Protein-Protein Interaction Network." International journal of biological sciences 10.7 (2014): 677.

[7] "Tumor - Nutrients." NetLogo User Community Models. NetLogo, 27 Dec. 2008. Web. 2 Jan. 2015.

[8] Schwikowski, Benno, Peter Uetz, and Stanley Fields. "A network of protein–protein interactions in yeast." Nature biotechnology 18.12 (2000): 1257-1261

[9] Cardelli, Luca. "Abstract machines of systems biology." Transactions on Computational Systems Biology III. Springer Berlin Heidelberg, 2005. 145-168.

[10] Barabási,Albert-László,Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." Nature Reviews Genetics 12.1 (2011): 56-68.

[11] Ergün, A., Lawrence, C. A., Kohanski, M. A., Brennan, T. A., & Collins, J. J. (2007). A network biology approach to prostate cancer. *Molecular systems biology*, *3*(1), 82.

[12] George, G., and V. Cyril Raj. "Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile." arXiv preprint arXiv:1109.1062 (2011).

[13] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389-422.

[14] Lo, J. Y., Baker, J. A., Kornguth, P. J., & Floyd, C. E. (1995). Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. *Academic radiology*, *2*(10), 841-850.

[15] Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, *3*(1).

[16] Janghel, R. R., Shukla, A., Tiwari, R., & Kala, R. (2010, June). Breast cancer diagnosis using artificial neural network models. In *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on* (pp. 89-94). IEEE.

[17] Polat, Kemal, and Salih Günes. "Breast cancer diagnosis using least square support vector machine." Digital Signal Processing 17.4 (2007): 694-701.

[18] Yu, W., He, L. R., Zhao, Y. C., Chan, M. H., Zhang, M., & He, M. (2013). Dynamic protein-protein interaction subnetworks of lung cancer in cases with smoking history. *Chinese journal of cancer*, *32*(2), 84.

[19] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[20] Huang, C. H., Peng, H. S., & Ng, K. L. (2015). Prediction of Cancer Proteins by Integrating Protein Interaction, Domain Frequency, and Domain Interaction Data Using Machine Learning Algorithms. *BioMed Research International*, *2015*.