Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 312047, 15 pages
http://dx.doi.org/10.1155/2015/312047

Research Article

Prediction of Cancer Proteins by Integrating Protein Interaction, Domain Frequency, and Domain Interaction Data Using Machine Learning Algorithms

Chien-Hung Huang,[1] Huai-Shun Peng,[1] and Ka-Lok Ng[2,3]

[1]*Department of Computer Science and Information Engineering, National Formosa University, 64 Wen-Hwa Road, Huwei, Yunlin 63205, Taiwan*
[2]*Department of Biomedical Informatics, Asia University, Wufeng Shiang, Taichung 41354, Taiwan*
[3]*Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 40402, Taiwan*

Correspondence should be addressed to Ka-Lok Ng; ppiddi@gmail.com

Received 2 December 2014; Revised 25 February 2015; Accepted 3 March 2015

Academic Editor: Xia Li

Many proteins are known to be associated with cancer diseases. It is quite often that their precise functional role in disease pathogenesis remains unclear. A strategy to gain a better understanding of the function of these proteins is to make use of a combination of different aspects of proteomics data types. In this study, we extended Aragues's method by employing the protein-protein interaction (PPI) data, domain-domain interaction (DDI) data, weighted domain frequency score (DFS), and cancer linker degree (CLD) data to predict cancer proteins. Performances were benchmarked based on three kinds of experiments as follows: (I) using individual algorithm, (II) combining algorithms, and (III) combining the same classification types of algorithms. When compared with Aragues's method, our proposed methods, that is, machine learning algorithm and voting with the majority, are significantly superior in all seven performance measures. We demonstrated the accuracy of the proposed method on two independent datasets. The best algorithm can achieve a hit ratio of 89.4% and 72.8% for lung cancer dataset and lung cancer microarray study, respectively. It is anticipated that the current research could help understand disease mechanisms and diagnosis.

1. Introduction

It has been known for a long time that cancer is a result of loss of cell cycle control. The loss of control is a result of series of genetic mutations involving activation of proto-oncogenes to oncogenes and inactivation of tumor-suppressing genes. Oncogenes and tumor suppressors may cause cancer by alternating the transcription factors, such as the p53 and ras oncoproteins, which in turn control expression of other genes. Therefore, understanding how oncoprotein-oncoprotein interacts and how oncoproteins drive the cell division cycle is indispensable for the study of molecular oncology. Predicting novel cancer-related proteins is an important topic in biomedical research; experimental techniques such as microarrays are being used to characterize cancer. However, the process could be time consuming and labor-intensive. Nagaraj and Reverter [1] proposed a Boolean logic based approach to predict colorectal cancer genes. Li et al. [2] took GO enrichment scores and KEGG enrichment scores as features to predict retinoblastoma related genes. The above two studies are confined to predict specific cancers. For general types of cancers, Hosur et al. [3] combined linear programming formulation for interface alignment to predict cancer related PPIs. Aragues et al. [4] used PPI data to predict cancer-related proteins. In this study, we extended Aragues's study by employing PPI data and domain information to attain improved performance.

Protein-protein interactions are inherent in almost every cellular process. In fact, PPI is the core of the entire interactomics system in living cells. PPI appears when two or more proteins bind together and perform a biological function [5]. Almost all major research topics in molecular biology involve PPI such as cellular function [6], genetic diseases [7], conserved patterns [8], and homologous relationships

[9]. The recent availability of PPI data has made it possible to study human disease at a system level. It is reported that since disease genes exhibit an increased tendency for their protein products to interact with one another, they tend to be coexpressed in specific tissues and display coherent functions [10]. Ideker and Sharan reported in a review article [11] on the applications of PPI networks to study disease in four major areas: (i) identifying new disease genes, (ii) studying their network properties; (iii) identifying disease-related subnetworks, and (iv) performing network-based disease classification. Another study [12] investigated the human cancer PPI network from a structural perspective, that is, protein interactions through their interfaces. Their findings indicated that cancer-related proteins have smaller, more planar, more charged, and less hydrophobic binding sites than noncancer proteins.

It is known that proteins are composed of multiple functional domains. A domain is a unit of function associated with different catalytic functions or binding sites, as found in enzymes or regulatory proteins. It is hypothesized that cancer proteins, also known as tumor associated genes, may share common functional domains [13], and thus a weighted domain score for each tumor associated gene's domain is determined. Novel cancer proteins are determined by translating full cDNA sequences to the corresponding protein sequences and calculating the weighted domain scores. Another work [14] used established methods to identify the network topology of a cancer protein network. They showed that cancer proteins contain a high ratio of structural domains, which have a high propensity for mediating protein interactions. Recently, Clancy et al. designed a statistical method to infer the physical interactions between two complexes for the human and yeast species [15]. Domains such as the immunoglobulin domain, Zinc-finger, and the protein kinase domains are the top three most frequently observed cancer protein domains. Many other works also employed PPI and DDI to characterize disease networks [7, 16–20]. In our previous works [21, 22], a one-to-one DDI model was proposed to obtain specific sets of DDI for oncoproteins and tumor suppressor proteins, respectively. Three specific sets of DDI, that is, oncoprotein and oncoprotein, tumor suppressor protein and tumor suppressor protein, and oncoprotein and tumor suppressor protein, are derived from their PPIs.

Weka (http://www.cs.waikato.ac.nz/~ml/weka/) [23] is a well known software tool which provides environments related to machine learning, data mining, text mining, predictive analysis, and business analysis. Machine learning and data mining algorithms have been widely used in bioinformatics and computational biology [24–28]. The present authors adopted amino acid composition profile information with the SVM classifier to improve protein complexes classification [29]. Additionally, we also proposed identifying microRNAs target of *Arabidopsis thaliana* by integrating prediction scores from PITA, miRanda, and RNAHybrid algorithms [30]. Recently, Li et al. [31] used random forest machine learning algorithm and topology features to identify the functions of protein complexes.

In this research, we began by collecting cancerous protein interaction data. That is, only interactions involved with cancer proteins are considered. These types of interactions are known as cancerous PPIs. Noncancerous protein interaction means either one or both of the proteins are not yet identified in relation to cancer. Given the cancerous PPIs, a set of DDI rules for cancer proteins are derived. In addition to this set of DDI, we also considered other features: the weighted domain frequency scores (DFS): DFS_C for cancer proteins and DFS_X for noncancer proteins and the cancer linker degree (CLD) score. A total of four features (DDI, DFS_C, DFS_X, and CLD) are adopted to make novel cancer protein predictions by using 39 machine learning algorithms from the Weka tool. In addition, we also verified the accuracy of the predictive model on two independent datasets. Finally, using differentially expressed genes found in lung cancer microarray data as a case study, we discovered some potential cancer genes for further experimental investigation.

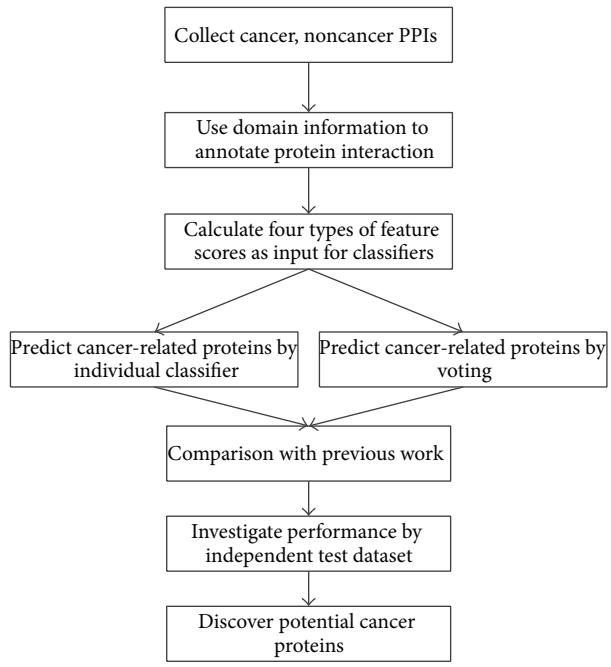

Figure 1: System flowchart for this study.

2. Methods

2.1. System Flowchart. In this study, a system was set up to predict cancer proteins by integrating four types of features. Firstly, cancer and noncancer PPIs were collected from biological databases, and then those interactions were annotated by using the domain information. Next, we determined four feature scores; data normalization is needed to ensure consistency in their distribution. The system flowchart of this study is illustrated in Figure 1.

2.2. Data Sources and Datasets Generation. Cancer proteins (tumor suppressor protein (TSP) or oncoprotein (OCP)) of the learning dataset were integrated from Institute of Biopharmaceutical Sciences of Taiwan National Yang Ming University, Tumor Associated Gene (TAG, http://www.binfo.ncku.edu.tw/TAG/GeneDoc.php) database [13], and Memorial

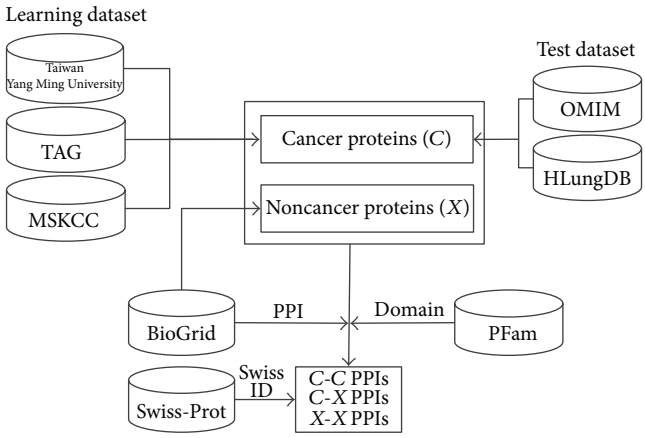

Figure 2: Data sources and datasets generation.

Sloan-Kettering Cancer Center (MSKCC, http://cbio.mskcc.org/research/cancergenomics/index.html). Noncancer proteins were from BioGrid (http://thebiogrid.org/) [32]. On the other hand, cancer proteins of the independent test dataset for Case Study 1 were obtained from two resources, that is, Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim) and the human lung cancer database (HLungDB) [33].

Protein domain information was downloaded from PFam database (http://pfam.xfam.org/) [34]. PPIs for both cancer and noncancer proteins were retrieved from BioGrid database. The Swiss ID of proteins was obtained from Swiss-Prot Database (http://www.expasy.org/sprot/).

There are three classified combinations of protein interaction adopted as input data in the current study: "*C*-*C*" indicates cancer-cancer protein interaction, "*C*-*X*" indicates cancer-noncancer protein interaction, and "*X*-*X*" indicates noncancer-noncancer protein interaction. Figure 2 depicts input data sources and dataset generation of this research.

2.3. Feature Scores Generation. We derived four feature scores based on three different approaches, which are domain-domain interaction, weighted domain frequency, and cancer linker degree. For the domain-domain interactions, it may happen that some relationships are derived from homologous sequences, which may produce a bias in the 10-fold validation; therefore, some redundancies (by a certain homology degree that includes domains) should be removed. By extracting the homologous PPIs by using the "UniRef50" dataset obtained from the UniProt Reference Clusters (UniRef, http://www.uniprot.org/uniref/), we found that, among the 123751 edges derived from BioGrid database, only 431 edges (approximately 0.348%) are homologous PPIs and 86 edges are cancer PPIs; it means that most of the PPIs are not homologous PPIs. We removed the 431 PPIs and the remaining 123320 PPIs are used in our experiment. In addition, the 431 homologous PPIs comprise 180 proteins, in which 179 proteins still appear in other nonhomologous PPIs; therefore, the total number of domains, that is, 3970, remains unchanged.

2.4. One-to-One Domain Interaction Model. Assuming that proteins P_i and P_j contain M and N domains, respectively, and then given an interacting protein pair (P_i, P_j), one considers that there are MN possible domain pairs. The set of domain pairs of two proteins P_i and P_j, $S_{i,j}$, is defined by

$$S_{i,j} = \left\{S\left(P_i\right) \times S\left(P_j\right)\right\}, \tag{1}$$

where $S(P_i)$ and $S(P_j)$ denote sets of protein domains in proteins P_i and P_j, respectively, and $\times$ denotes the Cartesian product of two sets $S(P_i)$ and $S(P_j)$.

To measure the likelihood of a DDI combination, a DDI pair interaction matrix I is introduced. The element $I_{\alpha,\beta}$ denotes the weighted combination probability of a domain pair (α, β) for a given protein pair (P_i, P_j), and it is given by

$$I_{\alpha,\beta} = \sum_{(P_i,P_j)} \frac{1}{\left|S\left(P_i\right)\right| * \left|S\left(P_j\right)\right|} \quad \text{if } \alpha \text{ in } S\left(P_i\right) \text{ and } \beta \text{ in } S\left(P_j\right), \tag{2}$$

where $|S(P_i)|$ and $|S(P_j)|$ denote the set sizes of $S(P_i)$ and $S(P_j)$, respectively, and $*$ is the multiplication operation; the summation is over all possible protein pairs of (P_i, P_j) such that α and β are an element of $S(P_i)$ and $S(P_j)$, respectively. Subsequently, protein domains are randomized while maintaining the number of domain assignments for each protein the same as the original set. The randomized counterpart of $I_{\alpha,\beta}$, $\langle I_{\alpha,\beta}^{\text{rand}} \rangle$, is performed in order to justify the protein domain pair calculation. Then, the domain pair score of the domain pair (α, β), $R_{\alpha,\beta}$, is defined by

$$R_{\alpha,\beta} = \frac{I_{\alpha,\beta}}{\left\langle I_{\alpha,\beta}^{\text{rand}} \right\rangle}, \tag{3}$$

where $\langle I_{\alpha,\beta}^{\text{rand}} \rangle$ denotes the ensemble average (we randomized the data 20 and 40 times, and it was found that the results converge after 40 times) of the randomized counterpart of $I_{\alpha,\beta}$. This result provides a criterion to rank the domain pairs. If the ratio $R_{\alpha,\beta}$ is larger than one, then the correlation is stronger than the randomized counterpart, so the domain pair (α, β) is a preferred DDI relation.

Given the set of domain annotation for any two proteins, one can turn around and compute a score that signifies PPI based on the set of $R_{\alpha,\beta}$ values for DDI. This derived PPI score can answer the question whether any two proteins interact or not given their domain components. The DDI score for the protein pair (P_i, P_j), $\text{DDI}_{i,j}$, is defined as follows:

$$\text{DDI}_{i,j} = \sum_{\alpha \in S(P_i), \beta \in S(P_j)} R_{\alpha,\beta} \quad \text{if } R_{\alpha,\beta} > 1, \tag{4}$$

where $S(P_i)$ and $S(P_j)$ denote the set of domains in proteins P_i and P_j, respectively.

2.5. Weighted Domain Frequency Score (DFS). The two feature scores (DFS_C and DFS_X) are defined in this section as the variations from the study by Chan [35]. Among the total of 3970 collected human domain types, the numbers

of 381 and 2750 of them appear only in cancer proteins and noncancer proteins, respectively, and 839 of them appear in both cancer proteins and noncancer proteins. This result supports the propensity that certain domain types reside in cancer and noncancer proteins.

Let $C = (C_1, C_2, \ldots, C_s)$ represent the set of cancer proteins, and let $D^C = (d_1^C, d_2^C, \ldots, d_m^C)$ be the set of domain types that appear in the cancer proteins; similarly, let $X = (X_1, X_2, \ldots, X_t)$ denote the set of noncancer proteins, and let $D^X = (d_1^X, d_2^X, \ldots, d_n^X)$ be the set of domain types that appear in noncancer proteins. For each domain α, let $C(\alpha)$ and $X(\alpha)$ denote the numbers of occurrence of domain α in cancer proteins and noncancer proteins, respectively. A higher score value suggested that the domain has a high propensity which resides in cancer or noncancer proteins. Then, two weighted DFS values for the protein pair (P_i, P_j), DFS_$C_{i,j}$ and DFS_$X_{i,j}$, are defined by the following, respectively:

$$\text{DFS\_}C_{i,j} = \frac{\sum_{\alpha \in S(P_i)} C(\alpha) + \sum_{\beta \in S(P_j)} C(\beta)}{m+n}, \quad (5)$$

$$\text{DFS\_}X_{i,j} = \frac{\sum_{\alpha \in S(P_i)} X(\alpha) + \sum_{\beta \in S(P_j)} X(\beta)}{m+n}, \quad (6)$$

where m and n are the total number of domain types that appear in cancer and noncancer proteins, respectively, and $S(P_i)$ and $S(P_j)$ denote sets of protein domains in proteins P_i and P_j, respectively.

The weighted DFS is adopted to measure the propensity of domain occurrence in cancer and noncancer proteins.

2.6. Cancer Linker Degree (CLD). The last feature is named the cancer linker degree (CLD) score which was adopted from the model proposed by Aragues et al. [4]. In organisms, proteins interact with each other to form a protein complex in order to perform special functions. We can conjecture the category of function and the level of activity by observing their interaction partners. For a given protein pair (P_i, P_j), let $n^C(P_i, P_j)$ and $n^X(P_i, P_j)$ denote the number of adjacent cancer proteins and noncancer proteins in PPI, respectively. Then, the cancer linker degree score for the protein pair (P_i, P_j), $\text{CLD}_{i,j}$, is defined by

$$\text{CLD}_{i,j} = \frac{n^C\left(P_i, P_j\right)}{n^C\left(P_i, P_j\right) + n^X\left(P_i, P_j\right)}. \quad (7)$$

As an illustration, an example is presented in Figure 3, where C_a and C_b are cancer proteins, and X_a, X_b, X_c, and X_d are noncancer proteins.

The CLD score represents the interaction ratio for a specified PPI interacting with a cancer partner. If the CLD score is close to one, it implies that the interaction edge is connecting many cancer nodes and could be located in the core of the cancer-related protein clusters.

2.7. Data Normalization. The four features scores consist of DDI score ($\text{DDI}_{i,j}$), weighted domain frequency scores (DFS_$C_{i,j}$ and DFS_$X_{i,j}$), and cancer linker degree score ($\text{CLD}_{i,j}$). Due to the fact that the distributions of the numerical values for the above four features are not consistent between each other, data normalization is needed. For the value after normalization, Y is defined by

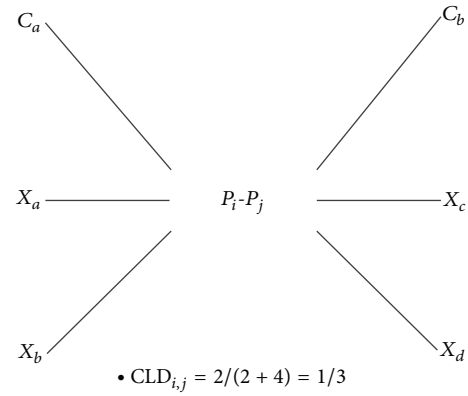

Figure 3: The calculation of score CLD for the protein pair (P_i, P_j).

$$Y = \frac{X^{D,W,C} - \min}{\max - \min}, \quad (8)$$

where $X^{D,W,C}$ denotes the unnormalized feature value and max and min are the maximum and minimum values for $X^{D,W,C}$, respectively.

2.8. Machine Learning Algorithms and Performance Statistical Measures. Since different choices of machine learning algorithms resulted in different predictions of performance, we conducted several comprehensive experiments to determine the optimal combinations of the algorithms. Thirty-nine machine learning algorithms in Weka are discussed in this study. Readers may refer to [23] for detailed descriptions about these algorithms. According to Weka, machine learning algorithms are divided into six classification types, that is, "Bayes" (6 algorithms), "functions" (6 algorithms), "Misc" (2 algorithms), "lazy" (4 algorithms), "rules" (9 algorithms), and "trees" (12 algorithms). A rigorous 10-fold cross validation test is performed to test the classification performance.

Six statistical measures are introduced to quantify the prediction performance, that is, accuracy (ACC), specificity (SPE), sensitivity (SEN), F-score ($F1$), Matthew's correlation coefficient (MCC), and positive predictive value (PPV), which are defined in terms of TP, TN, FP, and FN, where they denote true positive, true negative, false positive, and false negative events, respectively. Their definitions are listed in (9). SPE and SEN measure how well a true cancer protein or a true noncancer protein is identified. $F1$ conveys the balance between SPE and SEN. ACC and MCC provide an integrative measure of correct identification. PPV is positive predictive fraction. In addition, the AUC (area under the curve) score, which provides a global performance evaluation, is also included:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$F_1 = \frac{2 \times \text{SPE} \times \text{SEN}}{\text{SPE} + \text{SEN}} \times 100\%,$$

MCC

$$= \frac{(\text{TP} * \text{TN}) - (\text{FN} * \text{FP})}{\sqrt{(\text{TP} + \text{FN}) * (\text{TN} + \text{TP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})}},$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

3. Results

A total of 123320 PPIs, which are composed of 15214 cancer PPIs and 108106 noncancer PPIs, 2863 cancer proteins, and 3970 domains were used in our experiment. A 10-fold cross validation test was conducted to determine the optimal threshold settings for each classifier. We assumed the "*C-C*" type PPI as a positive set and the rest as a negative set. According to our previous work [30], the balanced trained dataset usually has better performance than the unbalanced one; hence, the algorithms are trained with an equal size ratio of 1 : 1 for the positive and negative dataset. Since the sizes of the original positive and negative sets differ by a factor of about 6 (unbalanced learning set), to generate a balanced learning set, the 15214 positive target interactions (cancer PPIs) were kept, and a total of 15214 noncancer PPIs were randomly selected from the negative set. Later on, the above-mentioned seven statistical measures are determined. For comparison, the corresponding results of unbalanced dataset are listed in Supplementary File 1, Appendix Tables S1 to S5, in Supplementary Material available online at http://dx.doi.org/10.1155/2015/312047, where the performance of MCC and PPV is much worse due to the very large TN and very small TP. Therefore, the use of balanced datasets is more preferable.

3.1. Performance Comparison by Individual Machine Learning Algorithm and Voting with the Majority. The performance comparison for the individual algorithm is listed in Table 1. The LMT algorithm of the "trees" type achieved the highest ACC (0.772), *F*1 (0.774), and MCC (0.548) among the 39 algorithms. Interestingly, according to either ACC, *F*1, or MCC, the top six algorithms (LMT, SimpleCart, J48, J48graft, REPTree, and FT) are all of the "trees" type. On the other hand, the LWL algorithm of the "lazy" type, the VFI algorithm of the "misc" type, the ConjunctiveRule algorithm of the "rules" type, and the DecisionStump algorithm of the "trees" type achieved the highest SPE and PPV (1.000), and the Nnge algorithm of the "rules" type achieved the highest SEN (0.858), while the Ridor algorithm of the "rules" type achieved the highest AUC (0.780). We also tried to combine the subsets of the four features for predicting the cancer proteins, but the predictive performance is not markedly improved.

The individual classifier has its own strengths and weaknesses; therefore, it is inspired to integrate multiple classifiers, that is, voting with the majority system, to improve the classification performance. Thirty-nine machine learning algorithms in Weka were selected and integrated using various types of voting with the majority system. In this study, the voting with the majority system involved an odd number of algorithms. For example, top 5 in Table 2 represents combining the top 5 algorithms extracted from Table 1, which are LMT, SimpleCart, J48, J48graft, and REPTree. Performance comparison by voting with the majority is listed in Table 2. The best performance in voting is attained when the top 23 algorithms are selected, which have the highest *F*1 (0.786) and PPV (0.890). When comparing Table 1 with Table 2, except SEN, the top voting with the majority system (top 23) is better than top individual algorithm (LMT) in the other six performance measures, and it also outperformed all thirty-nine individual algorithm in ACC, *F*1, MCC, and AUC. We also noted that the best performance of voting with the majority system (top 23) delivered lower FP events, that is, 186, than those of the top three individual algorithms, which are 303, 297, and 307, respectively. In other words, the voting approach does not introduce spurious events.

3.2. Performance Comparison with Group Voting with the Majority. Performance comparison by voting with the majority for each group is listed in Table 3 (Misc type is omitted here, because it contains only two algorithms). For instance, under the "functions" type, the "Top: 3" classifier in Table 3 represents the combination of "functions" type algorithms, that is, MultilayerPerceptron, Logistic, and SimpleLogistic algorithms (see Table 1). The results indicated that "trees" type achieved the highest ACC (0.781), SEN (0.749), *F*1 (0.786), MCC (0.567), MCC (0.513), and AUC (0.788), while the "rules" type had the highest SPE (0.838), and "lazy" type achieved the highest PPV (0.883).

For clarity, Figure 4 illustrates the majority voting results for various types of algorithms; the results suggest that "trees" type algorithms perform better in most of the performance measures.

3.3. Performance Comparison with the Competing Study. Since there are four features that are considered in this study, it is necessary to study their significance in classification performance. To study the prediction performance of the four features individually, we evaluated the feature importance by the area under the curve (AUC) value [36, 37]. Features with a higher AUC score are ranked as more important than features with a low score. The results of AUC values for the four features are given in Table 4. The DFS_C feature ranks at the top in AUC value, while the DFS_X feature has the lowest AUC value. These results suggest that DFS_C feature has the greatest discrimination information between positive datasets and negative datasets.

To demonstrate the effectiveness of the present study, we compared our results with the work by Aragues et al. [4], which uses the CLD feature only. As shown in Table 5 and Figure 5, for a single classifier, our method achieved better performance than Aragues's in all seven performance

Table 1: Performance comparison for the individual algorithm sorted by $F1$ value.

Type	Algorithm	ACC	SPE	SEN	$F1$	MCC	PPV	AUC
Trees	LMT	0.772	0.802	0.748	0.774	0.548	0.821	0.774
Trees	SimpleCart	0.770	0.804	0.742	0.773	0.546	0.825	0.775
Trees	J48	0.767	0.799	0.741	0.770	0.538	0.818	0.771
Trees	J48graft	0.767	0.799	0.742	0.769	0.538	0.818	0.771
Trees	REPTree	0.766	0.796	0.741	0.767	0.536	0.818	0.767
Trees	FT	0.763	0.798	0.735	0.766	0.528	0.821	0.766
Rules	DTNB	0.760	0.804	0.728	0.763	0.527	0.833	0.765
Trees	NBTree	0.760	0.796	0.733	0.763	0.527	0.819	0.764
Trees	RandomForest	0.760	0.777	0.744	0.761	0.524	0.791	0.761
Rules	Ridor	0.718	0.910	0.650	0.758	0.492	0.954	0.780
Rules	Jrip	0.754	0.790	0.726	0.757	0.512	0.817	0.757
Rules	DecisionTable	0.752	0.790	0.722	0.756	0.509	0.817	0.757
Rules	PART	0.744	0.758	0.745	0.748	0.494	0.754	0.751
Lazy	Kstar	0.744	0.766	0.725	0.744	0.491	0.784	0.744
Functions	MultilayerPerceptron	0.724	0.792	0.683	0.734	0.462	0.835	0.739
Trees	LADTree	0.720	0.762	0.696	0.727	0.447	0.788	0.729
Trees	RandomTree	0.721	0.721	0.720	0.721	0.440	0.723	0.721
Lazy	LWL	0.610	1.000	0.560	0.720	0.350	1.000	0.780
Misc	VFI	0.610	1.000	0.560	0.720	0.350	1.000	0.780
Rules	ConjunctiveRule	0.610	1.000	0.560	0.720	0.350	1.000	0.780
Trees	DecisionStump	0.610	1.000	0.560	0.720	0.350	1.000	0.780
Trees	ADTree	0.709	0.762	0.685	0.719	0.433	0.787	0.724
Bayes	BayesNet	0.714	0.755	0.683	0.717	0.431	0.796	0.719
Lazy	IB1	0.713	0.716	0.711	0.713	0.427	0.718	0.713
Lazy	Ibk	0.713	0.716	0.711	0.713	0.427	0.718	0.713
Functions	Logistic	0.669	0.652	0.692	0.672	0.342	0.606	0.673
Bayes	BayesianLogisticRegression	0.670	0.655	0.687	0.671	0.342	0.622	0.673
Functions	SimpleLogistic	0.669	0.652	0.691	0.670	0.342	0.605	0.672
Functions	SMO	0.665	0.641	0.699	0.667	0.334	0.580	0.670
Functions	VotedPerceptron	0.642	0.606	0.721	0.657	0.305	0.467	0.664
Rules	Nnge	0.522	0.511	0.858	0.641	0.128	0.052	0.686
Rules	OneR	0.635	0.629	0.641	0.635	0.269	0.610	0.634
Functions	RBFNetwork	0.624	0.603	0.655	0.627	0.254	0.529	0.629
Bayes	NaiveBayes	0.598	0.571	0.660	0.614	0.214	0.410	0.615
Bayes	NaiveBayesUpdateable	0.598	0.571	0.660	0.614	0.214	0.410	0.615
Bayes	NaiveBayesSimple	0.598	0.571	0.660	0.613	0.214	0.411	0.614
Bayes	NaiveBayesMultinomial	0.576	0.554	0.634	0.590	0.168	0.366	0.594
Misc	HyperPipes	0.570	0.568	0.579	0.571	0.143	0.534	0.572
Rules	ZeroR	0.510	0.510	0.510	0.510	0.010	0.510	0.510

measures. From Tables 1 and 5, we can see that the best single classifier of the current method (LMT) outperforms the best single classifier of Aragues's method (SimpleCart) in all seven performance measures; the difference value of ACC is 11.9%, SPE is 15.1%, SEN is 9.6%, $F1$ is 12.1%, MCC is 24.5%, PPV is 17.4%, and AUC is 12.1%.

From Tables 5 and 6, we can see that, using the CLD feature only, the best performance in voting is attained when

Table 2: Performance comparison by voting with the majority sorted by $F1$ value.

Classifiers	ACC	SPE	SEN	$F1$	MCC	PPV	AUC
TOP: 23	0.774	0.855	0.721	0.786	0.562	0.890	0.787
TOP: 11	0.781	0.829	0.744	0.785	0.569	0.855	0.787
TOP: 13	0.781	0.827	0.746	0.785	0.567	0.851	0.786
TOP: 9	0.783	0.820	0.751	0.784	0.568	0.844	0.786
TOP: 19	0.776	0.841	0.734	0.784	0.566	0.872	0.787
TOP: 21	0.775	0.856	0.723	0.784	0.564	0.889	0.789
TOP: 25	0.775	0.855	0.723	0.784	0.562	0.888	0.789
TOP: 27	0.774	0.847	0.727	0.784	0.562	0.879	0.787
TOP: 17	0.779	0.826	0.742	0.783	0.565	0.852	0.784
TOP: 7	0.780	0.819	0.748	0.782	0.563	0.840	0.784
TOP: 15	0.779	0.824	0.742	0.782	0.564	0.852	0.785
TOP: 29	0.773	0.836	0.730	0.780	0.556	0.867	0.783
TOP: 3	0.777	0.811	0.749	0.779	0.558	0.831	0.780
TOP: 5	0.776	0.811	0.748	0.779	0.556	0.830	0.780
TOP: 31	0.775	0.827	0.738	0.777	0.556	0.853	0.783
TOP: 33	0.774	0.819	0.738	0.776	0.552	0.847	0.778
TOP: 35	0.771	0.811	0.738	0.773	0.545	0.837	0.775
TOP: 37	0.768	0.802	0.739	0.770	0.540	0.826	0.772
TOP: 39	0.768	0.802	0.739	0.770	0.541	0.826	0.771

Table 3: Performance comparison by voting with the majority for five classification types.

Type	Classifiers	ACC	SPE	SEN	$F1$	MCC	PPV	AUC
Bayes	TOP: 3	0.672	0.653	0.695	0.672	0.346	0.609	0.673
	TOP: 5	0.599	0.571	0.660	0.614	0.215	0.410	0.614
Functions	TOP: 5	0.673	0.653	0.700	0.676	0.349	0.600	0.676
	TOP: 3	0.669	0.652	0.692	0.672	0.342	0.606	0.673
Lazy	TOP: 3	0.743	0.837	0.688	0.755	0.504	0.883	0.763
Rules	TOP: 5	0.764	0.825	0.721	0.769	0.537	0.856	0.774
	TOP: 7	0.764	0.826	0.721	0.769	0.537	0.857	0.774
	TOP: 9	0.765	0.817	0.726	0.769	0.534	0.848	0.771
	TOP: 3	0.755	0.838	0.705	0.766	0.528	0.877	0.771
Trees	TOP: 11	0.781	0.831	0.744	0.786	0.567	0.859	0.788
	TOP: 9	0.780	0.820	0.748	0.783	0.563	0.842	0.785
	TOP: 7	0.779	0.816	0.748	0.781	0.561	0.838	0.782
	TOP: 3	0.777	0.811	0.749	0.779	0.558	0.831	0.780
	TOP: 5	0.776	0.811	0.748	0.779	0.556	0.830	0.780

Table 4: The AUC value of the four features.

Feature	AUC	Rank
DFS_C	0.677	1
CLD	0.651	2
DDI	0.546	3
DFS_X	0.526	4

the top 3 algorithms (SimpleCart, REPTree, and FT) are selected using the CLD feature only, but the performance of voting with the majority is approximately equal to that of the individual algorithm. As shown in Table 6 and Figure 6, the proposed method significantly outperformed Aragues's in all seven performance measures. From Tables 2 and 6, we can see that the best voting with the majority of the current method (top 23) outperforms the best voting with the majority of Aragues's method (top 3) in all seven performance measures; the difference value of ACC is 12.2%, SPE is 20.2%, SEN is 6.9%, $F1$ is 13.4%, MCC is 25.8%, PPV is 24.0%, and AUC is 13.4%. These results suggest that the proposed method is superior to the competing study.

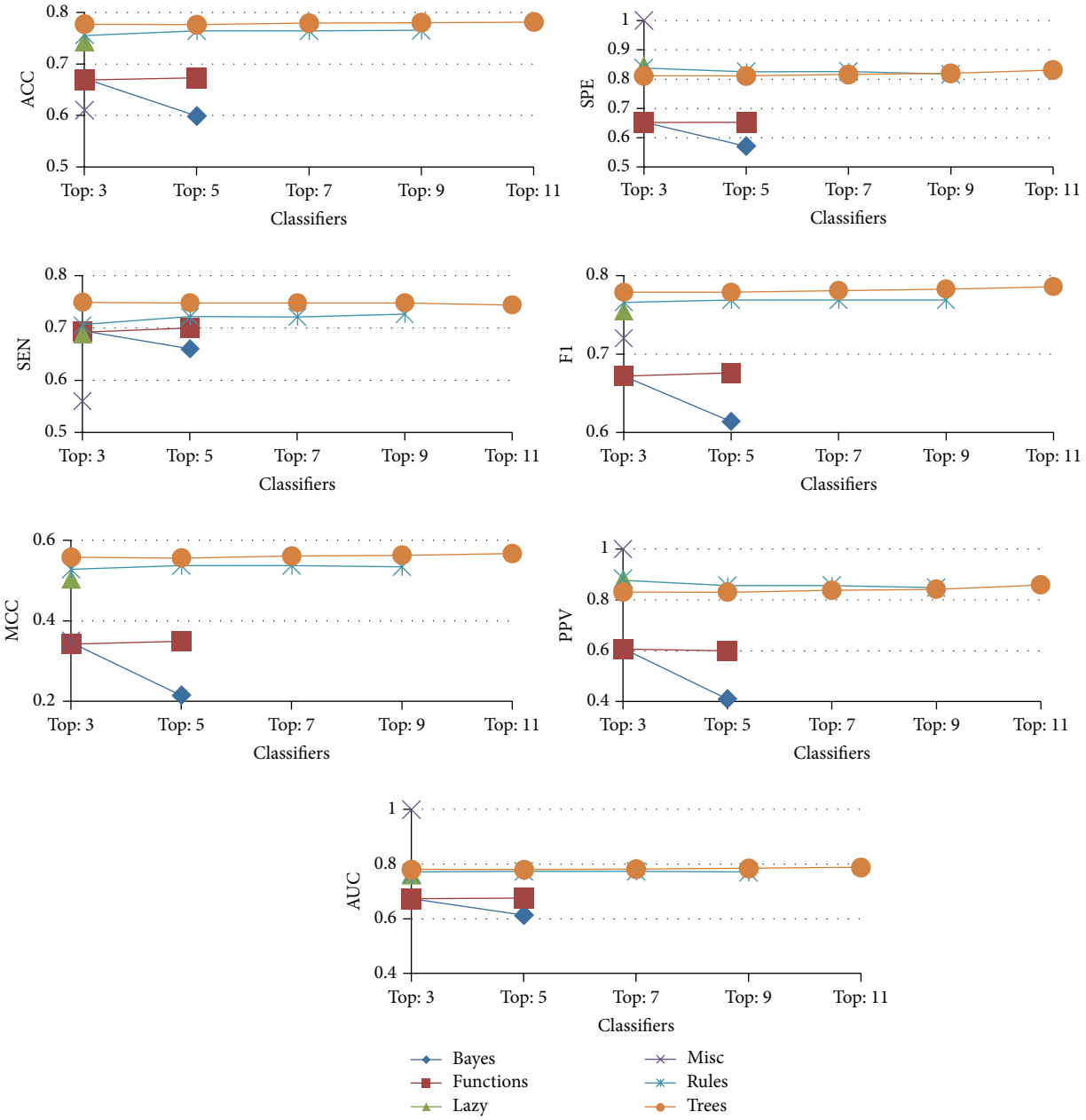

Figure 4: The seven performance measures for group voting with the majority.

4. Discussion

In this section, we designed two case studies to demonstrate the performance of the proposed method and showed how to discover potential cancer proteins, respectively.

4.1. Case Study 1. To investigate the performance of the proposed method, we retrieved lung cancer protein data from OMIM and HLungDB databases. Among the 2599 experimentally confirmed lung cancer proteins, there are a total of 1302 cancer proteins not appearing in our original training dataset, which could be used as the independent test dataset. List of the 1302 cancer proteins can be found in Supplementary File 2. The LWL, VFI, ConjunctiveRule, and DecisionStump were excluded from Case Study 1 because of their intrinsically high PPV which may bias the performance estimation. Consequently, as shown in Table 7, the hit number and hit ratio denote how many cancer proteins are true positive events and true positive ratios, respectively. Most of the algorithms had a hit ratio remarkably consistent over 75%, especially, the Ridor algorithm which achieved 89.4%. Compared with classifiers with higher ranks in $F1$ (Table 1), the results appear to suggest that classifiers with high PPV achieve better hit ratios.

Table 5: Performance comparison for the individual algorithm using the CLD feature sorted by *F*1.

Type	Algorithm	ACC	SPE	SEN	*F*1	MCC	PPV	AUC
Trees	SimpleCart	0.653	0.651	0.652	0.653	0.303	0.647	0.653
Trees	REPTree	0.650	0.650	0.649	0.650	0.300	0.649	0.650
Trees	FT	0.645	0.649	0.641	0.646	0.288	0.658	0.646
Rules	DecisionTable	0.642	0.650	0.639	0.644	0.288	0.663	0.644
Rules	DTNB	0.642	0.650	0.639	0.644	0.288	0.663	0.644
Trees	NBTree	0.643	0.648	0.639	0.644	0.287	0.661	0.644
Bayes	BayesNet	0.642	0.647	0.640	0.643	0.286	0.657	0.643
Trees	ADTree	0.642	0.644	0.641	0.643	0.283	0.646	0.643
Rules	Ridor	0.614	0.659	0.639	0.642	0.257	0.635	0.647
Trees	LADTree	0.642	0.648	0.641	0.642	0.287	0.653	0.644
Trees	LMT	0.639	0.656	0.625	0.640	0.280	0.693	0.640
Rules	PART	0.639	0.655	0.625	0.639	0.278	0.695	0.639
Trees	J48	0.639	0.655	0.627	0.639	0.280	0.689	0.641
Trees	J48graft	0.639	0.655	0.627	0.639	0.280	0.689	0.641
Rules	Jrip	0.637	0.639	0.638	0.638	0.277	0.634	0.638
Rules	OneR	0.635	0.629	0.641	0.635	0.269	0.610	0.634
Functions	VotedPerceptron	0.611	0.579	0.697	0.632	0.248	0.392	0.638
Lazy	LWL	0.612	0.582	0.683	0.629	0.246	0.431	0.632
Trees	DecisionStump	0.612	0.582	0.684	0.629	0.246	0.428	0.632
Rules	ConjunctiveRule	0.613	0.583	0.681	0.628	0.245	0.437	0.631
Bayes	NaiveBayes	0.619	0.595	0.656	0.623	0.242	0.496	0.627
Bayes	NaiveBayesSimple	0.619	0.595	0.656	0.623	0.242	0.496	0.627
Bayes	NaiveBayesUpdateable	0.619	0.595	0.656	0.623	0.242	0.496	0.627
Trees	RandomForest	0.623	0.616	0.631	0.623	0.247	0.591	0.624
Lazy	Ibk	0.622	0.613	0.632	0.622	0.242	0.586	0.622
Trees	RandomTree	0.622	0.613	0.632	0.622	0.242	0.586	0.622
Bayes	BayesianLogisticRegression	0.621	0.612	0.630	0.621	0.240	0.583	0.621
Functions	Logistic	0.621	0.612	0.631	0.621	0.241	0.582	0.621
Functions	SimpleLogistic	0.621	0.612	0.633	0.621	0.241	0.578	0.621
Functions	SMO	0.619	0.607	0.640	0.621	0.245	0.549	0.622
Lazy	Kstar	0.619	0.606	0.640	0.620	0.242	0.547	0.623
Functions	MultilayerPerceptron	0.618	0.598	0.649	0.620	0.242	0.518	0.621
Functions	RBFNetwork	0.618	0.598	0.647	0.620	0.240	0.517	0.620
Lazy	IB1	0.594	0.564	0.683	0.619	0.214	0.351	0.624
Rules	Nnge	0.544	0.601	0.529	0.562	0.106	0.832	0.564
Misc	HyperPipes	0.510	0.510	0.510	0.510	0.010	0.510	0.510
Rules	ZeroR	0.510	0.510	0.510	0.510	0.010	0.510	0.510
Bayes	NaiveBayesMultinomial	0.500	0.561	0.447	0.480	−0.008	0.489	0.569
Misc	VFI	0.500	NA	0.500	NA	NA	1.000	NA

4.2. Case Study 2. It is known that interacting proteins are often coexpressed; one can identify differentially expressed genes (DEGs) among a large number of gene expressions and understand the mechanism of lung cancer formation induced by these DEGs [38]. We further explored the potential cancer genes from DEGs in microarray data. Four sets of lung cancer microarray data were downloaded from the GEO database [39] and summarized in Table 8. Experiments GSE7670 [40] and GSE10072 [41] use the HG-U133A array, where GSE19804 [42] and GSE27262 [43] use HG-U133 plus 2.0 chip.

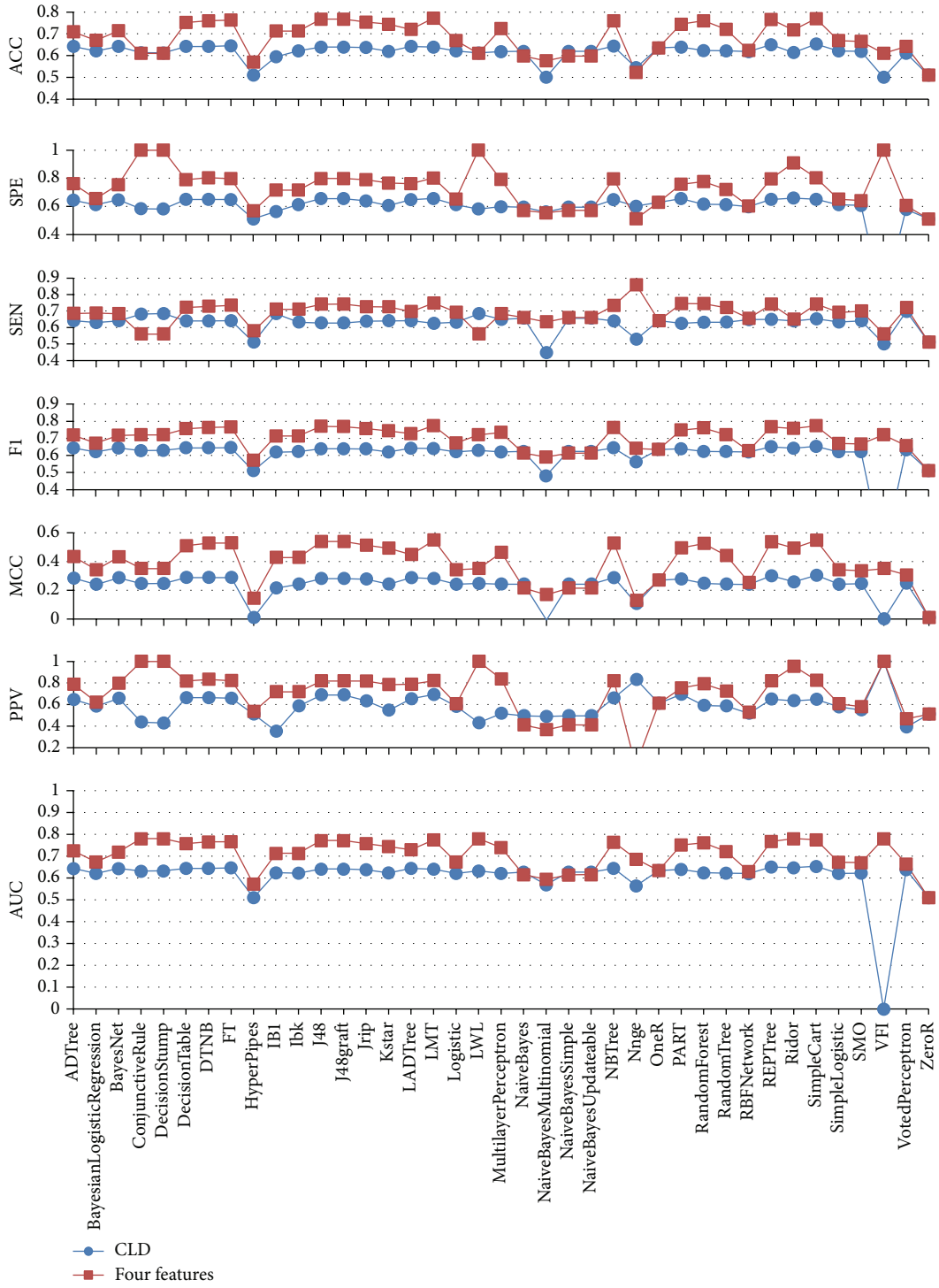

Figure 5: The performance comparison of the individual algorithm for Aragues (blue) and the proposed method (red).

The tested DEGs are collected from the intersection set of the above four microarray datasets. Among the 1345 common DEGs in the four microarray datasets, 360 DEGs were excluded because of their appearance in the original training set; another 209 DEGs are also removed due to their lacking of domain data or PPI data. The remaining 776 DEGs serve as input data.

Five classifiers, including the top three classifiers according to the $F1$ measure, LMT, SimpleCart, and J48 algorithms, as well as the top one classifier according to PPV measure, LWL algorithm, along with the top one classifier according to Case Study 1, Ridor algorithm, were selected for evaluating potential cancer genes under strictly uniformed voting; that is, only the one with five votes which all five classifiers

Table 6: Performance comparison by voting with the majority using the CLD feature sorted by *F*1.

Classifiers	ACC	SPE	SEN	*F*1	MCC	PPV	AUC
TOP: 3	0.652	0.653	0.652	0.652	0.304	0.650	0.653
TOP: 27	0.648	0.646	0.650	0.648	0.296	0.643	0.648
TOP: 17	0.646	0.653	0.639	0.647	0.291	0.671	0.647
TOP: 25	0.646	0.647	0.648	0.647	0.296	0.643	0.648
TOP: 5	0.646	0.651	0.639	0.646	0.293	0.666	0.646
TOP: 15	0.645	0.654	0.638	0.646	0.291	0.674	0.646
TOP: 19	0.643	0.651	0.640	0.646	0.289	0.663	0.645
TOP: 23	0.644	0.648	0.645	0.646	0.294	0.648	0.647
TOP: 29	0.646	0.645	0.648	0.646	0.292	0.641	0.647
TOP: 13	0.644	0.653	0.639	0.645	0.291	0.669	0.646
TOP: 31	0.644	0.642	0.648	0.645	0.292	0.634	0.644
TOP: 7	0.643	0.650	0.639	0.644	0.288	0.663	0.644
TOP: 9	0.643	0.650	0.639	0.644	0.290	0.662	0.644
TOP: 11	0.644	0.651	0.639	0.644	0.290	0.665	0.644
TOP: 21	0.642	0.649	0.642	0.644	0.289	0.659	0.645
TOP: 39	0.644	0.638	0.648	0.644	0.288	0.627	0.644
TOP: 37	0.642	0.635	0.649	0.643	0.285	0.620	0.644
TOP: 35	0.641	0.635	0.648	0.642	0.286	0.619	0.642
TOP: 33	0.641	0.635	0.648	0.641	0.285	0.620	0.642

predict as a cancer protein was considered. Among the 776 DEGs, a total of 565 DEGs (72.8%, a higher value can be obtained if we relax the strictly uniformed requirement) were evaluated as potential cancer genes. The complete 565 potential cancer genes derived from the second case study are listed in Supplementary File 3.

To validate our findings, we conducted a study of the literature by randomly selecting five genes (DBF4, MCM2, ID3, EXOSC4, and CDKN3) from the 565 potential cancer genes. The results indicated that while EXOSC4 remains unclear, the others are in agreement with our predictions. Barkley proposed that miR-29a targets the 3-UTR of DBF4 mRNA in lung cancer cells [44] and Bonte stated that most cell lines with increased Cdc7 protein levels also had increased DBF4 abundance, and some tumor cell lines had extra copies of the DBF4 gene [45]. Alexandrow noted that Stat3-P and the proliferative markers MCM2 were expressed in mice lung tissues *in vivo* [46]. Yang et al. observed that patients with higher levels of MCM2 and gelsolin experienced shorter survival time than patients with low levels of MCM2 and gelsolin [47]. Langenfeld et al. [48] indicated that Oct4 cells give rise to lung cancer cells expressing nestin and/or NeuN, and BMP signaling is an important regulator of ID1 and ID3 in both Oct4 and nestin cell populations. Tang claimed that CDKN3 has significant biological implications in tumor pathogenesis [49]. In [50], a metasignature was identified in eight separate microarray analyses spanning seven types of cancer including lung adenocarcinoma, and these included many genes associated with cell proliferation, and CDKN3 is among them.

Given a single protein as testing data, we can first treat this protein as X. If both PPIs and domains information are available, one can then apply the present method to classify the interaction type "C-X".

Any "C-X" type is classified as "C-C", and then there are two possible explanations for this: (i) the classifier is not completely specific; therefore, one has FP prediction, and (ii) prediction is a TP event. If one can exclude the first explanation option, then the present calculation provides a potential way to assign X as C; in other words, it provides a feasible solution for predicting cancer proteins.

If the PPI information is missing, given the FASTA sequence information, one can make use of the STRING database [51]. STRING is a database that provides known and predicted PPI derived from four sources: genomic context, high-throughput experiments, conserved coexpression, and published literature. On domain prediction, one can carry out the analysis by using the online tool "SEQUENCE SEARCH" under PFam [34] to find matching domains. Then, given the PPIs and domains information, one can conduct the same analysis as described in the last paragraph; otherwise, one is facing a difficult task, which requires further discussion or work.

5. Conclusion

Identifying cancer protein is a critical issue in treating cancer; however, identifying cancer protein experimentally is extremely time consuming and labor-intensive. Alternative

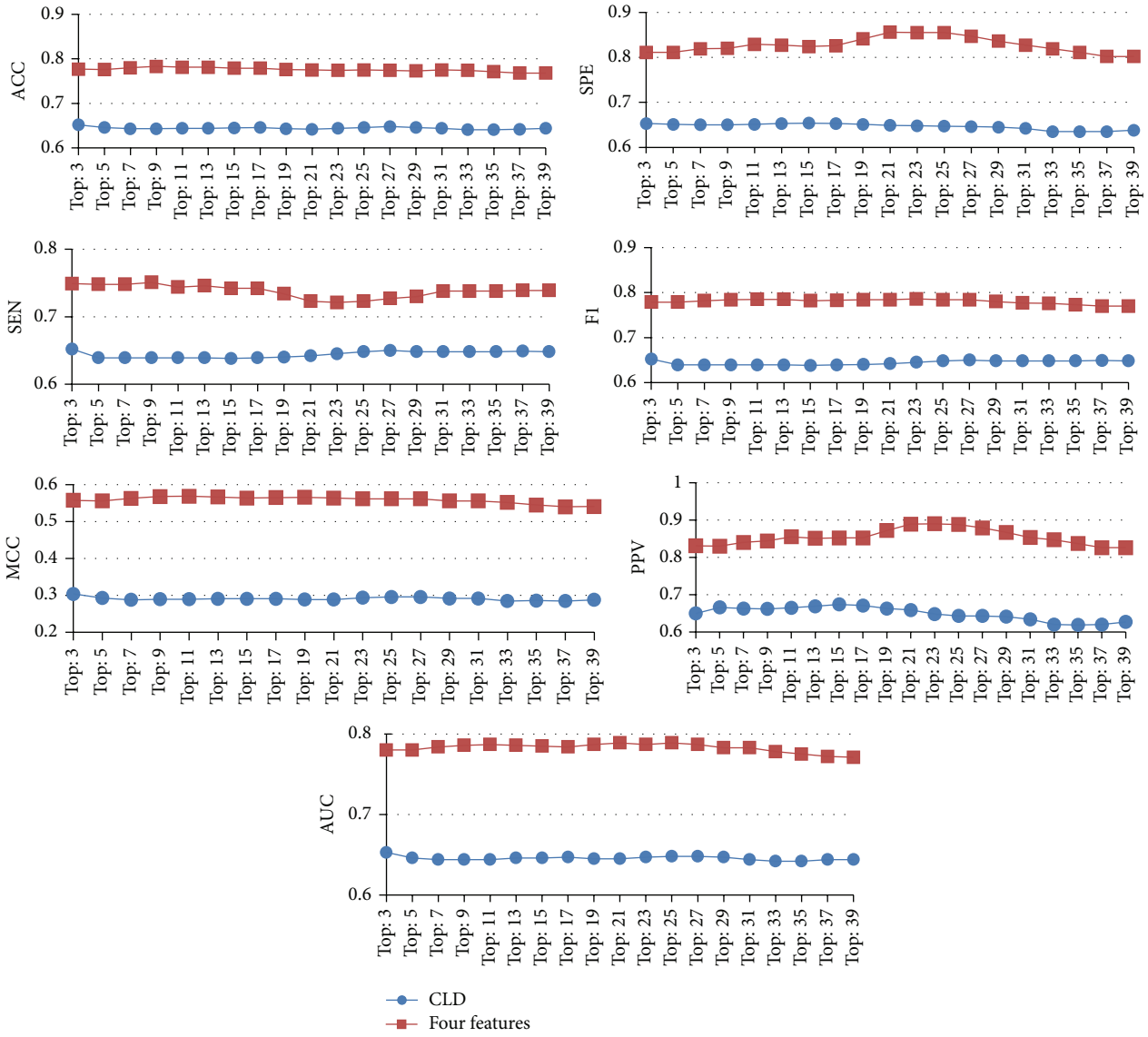

Figure 6: The performance comparison of the voting with the majority for Aragues (blue) and the proposed method (red).

methods must be developed to discover cancer proteins. We have integrated several proteomic data sources to develop a model for predicting cancer protein-cancer protein interactions on a global scale based on domain-domain interactions, weighted domain frequency score, and cancer linker degree. A one-to-one interaction model was introduced to quantify the likelihood of cancer-specific DDI. The weighted DFS is adopted to measure the propensity of domain occurrence in cancer and noncancer proteins. Finally, the CLD is defined to gauge cancer and noncancer proteins' interaction partners. As a result, voting with a majority system achieved ACC (0.774), SPE (0.855), SEN (0.721), $F1$ (0.786), MCC (0.562), PPV (0.890), and AUC (0.787) when the top 23 algorithms were selected, which is better than the best single classifier (LMT) in six performance measures except SEN.

We compared our performance with the previous work [4]. It was shown that the present approach outperformed Aragues's in all seven performance measures in both individual algorithm and combining algorithms. Effectiveness of the current research is further evaluated by two independent datasets; experimental results demonstrated that the proposed method can identify cancer proteins with high hit ratios. The current research not only significantly improves the prediction performance of cancer proteins, but also discovered some potential cancer proteins for future experimental investigation. It is anticipated that the current research could provide some insight into disease mechanisms and diagnosis.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Table 7: Performance evaluation for OMIM and HLungDB datasets.

Type	Algorithm	Hit number	Hit ratio
Rules	Ridor	1164	0.894
Rules	ZeroR	1159	0.890
Trees	ADTree	1119	0.859
Misc	HyperPipes	1115	0.856
Functions	MultilayerPerceptron	1076	0.826
Trees	LADTree	1061	0.815
Rules	OneR	1047	0.804
Lazy	IB1	1023	0.786
Lazy	IBk	1023	0.786
Bayes	BayesNet	1020	0.783
Rules	PART	1019	0.783
Trees	J48graft	1018	0.782
Trees	J48	1017	0.781
Rules	DecisionTable	1011	0.776
Trees	FT	1004	0.771
Lazy	KStar	999	0.767
Bayes	BayesianLogisticRegression	998	0.767
Trees	NBTree	995	0.764
Rules	DTNB	991	0.761
Trees	RandomTree	988	0.759
Trees	LMT	983	0.755
Trees	REPTree	975	0.749
Trees	SimpleCart	973	0.747
Trees	RandomForest	973	0.747
Rules	JRip	966	0.742
Functions	Logistic	964	0.740
Functions	SimpleLogistic	964	0.740
Functions	SMO	944	0.725
Functions	RBFNetwork	925	0.710
Functions	VotedPerceptron	844	0.648
Bayes	NaiveBayesSimple	827	0.635
Bayes	NaiveBayes	826	0.634
Bayes	NaiveBayesUpdateable	826	0.634
Bayes	NaiveBayesMultinomial	796	0.611
Rules	NNge	94	0.072

Table 8: Summary of microarray datasets.

GEO ID	Organization name	Number of DEGs
GSE7670	Taipei Veterans General Hospital	1874
GSE10072	National Cancer Institute, NIH	3138
GSE19804	National Taiwan University	5398
GSE27262	National Taiwan Yang Ming University	8476
		1345 (intersection)

Acknowledgments

The work of Chien-Hung Huang is supported by the Ministry of Science and Technology of Taiwan (MOST) and National Formosa University under Grants nos. NSC 101-2221-E-150-088-MY2 and EN103B-1005, respectively. The work of Ka-Lok Ng is supported by MOST and Asia University under Grants nos. NSC 102-2632-E-468-001-MY3 and 103-asia-06, respectively. The authors thank Yi-Wen Lin and Nilubon Kurubanjerdjit for their contributions to implementation assistance and providing the Case Study 1 dataset.

References

[1] S. H. Nagaraj and A. Reverter, "A Boolean-based systems biology approach to predict novel genes associated with cancer: application to colorectal cancer," *BMC Systems Biology*, vol. 5, article 35, 2011.

[2] Z. Li, B.-Q. Li, M. Jiang et al., "Prediction and analysis of retinoblastoma related genes through gene ontology and KEGG," *BioMed Research International*, vol. 2013, Article ID 304029, 8 pages, 2013.

[3] R. Hosur, J. Xu, J. Bienkowska, and B. Berger, "IWRAP: an interface threading approach with application to prediction of cancer-related protein-protein interactions," *Journal of Molecular Biology*, vol. 405, no. 5, pp. 1295–1310, 2011.

[4] R. Aragues, C. Sander, and B. Oliva, "Predicting cancer involvement of genes from heterogeneous data," *BMC Bioinformatics*, vol. 9, article 172, 2008.

[5] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, "Protein-protein interaction networks and biology—what's the connection?" *Nature Biotechnology*, vol. 26, no. 1, pp. 69–72, 2008.

[6] C. von Mering, R. Krause, B. Snel et al., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[7] B. Schuster-Böckler and A. Bateman, "Protein interactions in human genetic diseases," *Genome Biology*, vol. 9, no. 1, article R9, 2008.

[8] R. Sharan, S. Suthram, R. M. Kelley et al., "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974–1979, 2005.

[9] C.-C. Chen, C.-Y. Lin, Y.-S. Lo, and J.-M. Yang, "PPISearch: a web server for searching homologous protein-protein interactions across multiple species," *Nucleic Acids Research*, vol. 37, no. 2, pp. W369–W375, 2009.

[10] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[11] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.

[12] G. Kar, A. Gursoy, and O. Keskin, "Human cancer protein-protein interaction network: a structural perspective," *PLoS Computational Biology*, vol. 5, no. 12, Article ID e1000601, 2009.

[13] J.-S. Chen, W.-S. Hung, H.-H. Chan, S.-J. Tsai, and H. Sunny Sun, "In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma," *Bioinformatics*, vol. 29, no. 4, pp. 420–427, 2013.

[14] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, 2006.

[15] T. Clancy, E. A. Rødland, S. Nygard, and E. Hovig, "Predicting physical interactions between protein complexes," *Molecular and Cellular Proteomics*, vol. 12, no. 6, pp. 1723–1734, 2013.

[16] D. P. Ryan and J. M. Matthews, "Protein-protein interactions in human disease," *Current Opinion in Structural Biology*, vol. 15, no. 4, pp. 441–446, 2005.

[17] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

[18] A. Platzer, P. Perco, A. Lukas, and B. Mayer, "Characterization of protein-interaction networks in tumors," *BMC Bioinformatics*, vol. 8, article 224, 2007.

[19] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2007.

[20] J. Goñi, F. J. Esteban, N. V. de Mendizábal et al., "A computational analysis of protein-protein interaction networks in neurodegenerative diseases," *BMC Systems Biology*, vol. 2, article 52, 2008.

[21] Y.-L. Lee, J.-W. Weng, W.-C. Chiang et al., "Investigating cancer-related proteins specific domain interactions and differential protein interactions caused by alternative splicing," in *Proceedings of the 11th IEEE International Conference on Bioinformatics and Bioengineering (BIBE '11)*, pp. 33–38, Taichung, Taiwan, October 2011.

[22] K. L. Ng, J. S. Ciou, and C. H. Huang, "Prediction of protein functions based on function-function correlation relations," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 300–305, 2010.

[23] H. Ian and E. F. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.

[24] Y.-D. Cai, L. Lu, L. Chen, and J.-F. He, "Predicting subcellular location of proteins using integrated-algorithm method," *Molecular Diversity*, vol. 14, no. 3, pp. 551–558, 2010.

[25] C. R. Peng, L. Liu, B. Niu et al., "Prediction of RNA-Binding proteins by voting systems," *Journal of Biomedicine and Biotechnology*, vol. 2011, Article ID 506205, 8 pages, 2011.

[26] C.-Y. Hor, C.-B. Yang, Z.-J. Yang, and C.-T. Tseng, "Prediction of protein essentiality by the support vector machine with statistical tests," *Evolutionary Bioinformatics Online*, vol. 9, pp. 387–416, 2013.

[27] S. K. Dhanda, D. Singla, A. K. Mondal, and G. P. S. Raghava, "DrugMint: a webserver for predicting and designing of drug-like molecules," *Biology Direct*, vol. 8, no. 1, article 28, 2013.

[28] T. Wilhelm, "Phenotype prediction based on genome-wide DNA methylation data," *BMC Bioinformatics*, vol. 15, no. 1, article 193, 2014.

[29] C.-H. Huang, S.-Y. Chou, and K.-L. Ng, "Improving protein complex classification accuracy using amino acid composition profile," *Computers in Biology and Medicine*, vol. 43, no. 9, pp. 1196–1204, 2013.

[30] N. Kurubanjerdjit, C.-H. Huang, Y.-L. Lee, J. J. P. Tsai, and K.-L. Ng, "Prediction of microRNA-regulated protein interaction pathways in Arabidopsis using machine learning algorithms," *Computers in Biology and Medicine*, vol. 43, no. 11, pp. 1645–1652, 2013.

[31] Z.-C. Li, Y.-H. Lai, L.-L. Chen, Y. Xie, Z. Dai, and X.-Y. Zou, "Identifying functions of protein complexes based on topology similarity with random forest," *Molecular BioSystems*, vol. 10, no. 3, pp. 514–525, 2014.

[32] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke et al., "The BioGRID interaction database: 2013 update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D816–D823, 2013.

[33] L. Wang, Y. Xiong, Y. Sun et al., "HlungDB: an integrated database of human lung cancer research," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp945, pp. D665–D669, 2009.

[34] R. D. Finn, A. Bateman, J. Clements et al., "Pfam: the protein families database," *Nucleic Acids Research*, vol. 42, no. 1, pp. D222–D230, 2014.

[35] H. H. Chan, *Identification of novel tumor-associated gene (TAG) by bioinformatics analysis [M.S. thesis]*, National Cheng Kung University, Tainan City, Taiwan, 2006.

[36] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Cambridge, Mass, USA, 2nd edition, 2011.

[37] I. R. Galatzer-Levy, K.-I. Karstoft, A. Statnikov, and A. Y. Shalev, "Quantitative forecasting of PTSD from early trauma responses: a machine Learning application," *Journal of Psychiatric Research*, vol. 59, pp. 68–76, 2014.

[38] C.-H. Huang, M.-H. Wu, P. M.-H. Chang, C.-Y. Huang, and K.-L. Ng, "In silico identification of potential targets and drugs for non-small cell lung cancer," *IET Systems Biology*, vol. 8, no. 2, pp. 56–66, 2014.

[39] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D991–D995, 2013.

[40] L. J. Su, C. W. Chang, Y. C. Wu et al., "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme," *BMC Genomics*, vol. 8, article 140, 2007.

[41] M. T. Landi, T. Dracheva, M. Rotunno et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, Article ID e1651, 2008.

[42] T.-P. Lu, M.-H. Tsai, J.-M. Lee et al., "Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women," *Cancer Epidemiology Biomarkers and Prevention*, vol. 19, no. 10, pp. 2590–2597, 2010.

[43] T.-Y. W. Wei, C.-C. Juan, J.-Y. Hisa et al., "Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade," *Cancer Science*, vol. 103, no. 9, pp. 1640–1650, 2012.

[44] L. R. Barkley and C. Santocanale, "MicroRNA-29a regulates the benzo[a]pyrene dihydrodiol epoxide-induced DNA damage response through Cdc7 kinase in lung cancer cells," *Oncogenesis*, vol. 2, article e57, 2013.

[45] D. Bonte, C. Lindvall, H. Liu, K. Dykema, K. Furge, and M. Weinreich, "Cdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivation," *Neoplasia*, vol. 10, no. 9, pp. 920–931, 2008.

[46] M. G. Alexandrow, L. J. Song, S. Altiok, J. Gray, E. B. Haura, and N. B. Kumar, "Curcumin: a novel Stat3 pathway inhibitor for chemoprevention of lung cancer," *European Journal of Cancer Prevention*, vol. 21, no. 5, pp. 407–412, 2012.

[47] J. Yang, N. Ramnath, K. B. Moysich et al., "Prognostic significance of MCM2, Ki-67 and gelsolin in non-small cell lung cancer," *BMC Cancer*, vol. 6, article 203, 2006.

[48] E. Langenfeld, M. Deen, E. Zachariah, and J. Langenfeld, "Small molecule antagonist of the bone morphogenetic protein type I receptors suppresses growth and expression of Id1 and Id3 in lung cancer cells expressing Oct4 or nestin," *Molecular Cancer*, vol. 12, no. 1, article 129, 2013.

[49] H. Tang, G. Xiao, C. Behrens et al., "A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients," *Clinical Cancer Research*, vol. 19, no. 6, pp. 1577–1586, 2013.

[50] D. M. MacDermed, N. N. Khodarev, S. P. Pitroda et al., "MUC1-associated proliferation signature predicts outcomes in lung adenocarcinoma patients," *BMC Medical Genomics*, vol. 3, article 16, 2010.

[51] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.