

Prediction of Protein Essentiality by the Support Vector Machine with Statistical Tests*

Chiou-Yi Hor, Chang-Biau Yang[†], Zih-Jie Yang and Chiou-Ting Tseng
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung 80424, Taiwan

Abstract—Essential proteins affect the cellular life deeply, but it is extreme time-consuming and labor-intensive to discriminate them experimentally. The goal of this paper is to identify the features which are crucial for discriminating protein essentiality and build learning machines for prediction. We first collect features from a variety of sources. Then we adopt a backward feature selection method and use the selected features to build SVM predictors. The cross validations are conducted on the originally imbalanced data set as well as the down-sampling balanced data set. The performance of these feature subsets are then subject to the statistical test to confirm their significance. For the imbalanced data set, our best values of F-measure and MCC are 0.549 and 0.495, respectively. For balanced data set, our best values of F-measure and MCC of our models are 0.770 and 0.545, respectively. The results are superior to all previous results under various performance measures.

Index Terms—bioinformatics; essential protein; protein-protein interaction; support vector machine; statistical test

I. INTRODUCTION

The identification of essential proteins is important for understanding the cellular life. It is usually believed that once an essential protein is removed, it will cause the cell to lose its functionality or life because the function of the essential protein cannot be taken over by other proteins. As a result, lots of works are devoted to identify essential proteins. According to literatures, essential proteins can be distinguished by the experimental approaches, like gene deletion [3], RNA interference [6] and combination of gene replacement and conditional gene expression. However, the required tasks are extremely time-consuming and labor-intensive. Hence, it motivates many researchers to devise some economical ways for substitution. Among these, machine-learning based methods seems promising ways to address this problem.

The *protein-protein interaction* (PPI) is one of the significant characteristics between proteins. In the past, finding PPI is a time-consuming work. Recently, with the *yeast two-hybrid* [11] high-throughput technique, which can identify a lot of PPIs in one experiment, it therefore becomes easier to get the PPI information. Since a PPI network is similar to a social network in many ways, some researchers apply social network techniques to the PPI network studies. Consequently, many topological properties were extensively explored and

studied in recent years. It is shown that lots of biological researches, such as prediction and classification of protein functionality [16] or analysis of protein phenotype [4], can be benefited by involving these network properties. Besides, previous study [2] further suggests that essential proteins and nonessential proteins can be discriminated by means of topological properties derived from the PPI network.

The essential protein classification problem is to determine the essentiality of a protein, for given its various properties. Several researchers have devoted to the problem by means of machine learning techniques recently. Chin *et al.* [5] proposed a double screening scheme and built a framework, called *hub analyzer* (<http://hub.iis.sinica.edu.tw/Hubba/index.php>), to rank the proteins. Hwang *et al.* [10] applied the *support vector machine* (SVM) to classify the proteins. Acencio *et al.* [1] used *Waikato Environment for Knowledge Analysis* (WEKA) [21] to predict the essential proteins. These studies more or less adopt parts of topology and protein properties and do not consider some fundamental properties, like sequence or protein physiochemical ones. Although these preliminary properties may be somewhat less relevant to the essentiality, they can be extracted with merely sequence information and thus are relatively accessible. Hence, we take these properties together in our study.

In this paper, we get the PPI data set Scere20070107, which contains 4873 proteins and 17166 interactions, from the DIP database. Our feature set consists of the features obtained or extracted from the methods proposed by Chin *et al.* [5], Hwang *et al.* [10], Lin *et al.* [14] and Acencio *et al.* [1]. The LIBSVM tool [9] is used to predict the essential proteins. We adopt a *modified sequential backward feature selection method* and build SVM models with various subsets of features. Each model is applied on both imbalanced and balanced data sets to examine the significance against the previous works. The best values of F-measure and MCC of our experiments on the imbalanced data set are 0.549 and 0.495, respectively. For the balanced data set, we get 0.770 and 0.545, respectively. The best of our models outperforms other previous methods, which achieved 0.354 and 0.36 on the imbalanced data set as well as 0.737 and 0.492 on the balanced data set [10], with various performance measures. We finally adopt *bootstrap cross-validation* [7] to verify that the improvement is statistically significant.

The rest of this paper is organized as follows. In Section

*This research work was partially supported by the National Science Council of Taiwan under contract NSC100-2221-E-110-050.

[†]Corresponding author: cbyang@cse.nsysu.edu.tw

II, we will introduce some preliminary knowledge about this study. In Section III, we will present our method to the problem. Section IV shows the experimental results and compares them with some previous results. Finally, Section V gives a conclusion and some possible future works.

II. Preliminaries

In this section, we will introduce some fundamental knowledge used in our paper, which include the data set, significance tests, and the performance measures.

A. The Data set

In this paper, the experiments are performed with the data set Scere20070107, which is a PPI data set and downloaded from the DIP (<http://dip.doe-mbi.ucla.edu/>) database [18]. This data set contains 4873 proteins and 17166 interactions. We consider the largest connected component in the PPI, which has 4815 proteins, including 975 essential proteins and 3840 nonessential proteins. The information of essential proteins is extracted from the SGD (<http://www.yeastgenome.org/>). Since this data set has been adopted in several studies, we thus can obtain and incorporate various related features for training.

In the data set, the ratio of essential proteins to nonessential proteins is low, nearly 1:4, which will lead to biased fitting to nonessential proteins during the training and testing processes. Hence, we build another balanced data set, in which the number of essential data against that of non-essential ones are equal. That is, we randomly select 975 non-essential proteins and mix them with the essential ones to form a balanced data set.

B. Performance Evaluation

Receiver operating characteristic (ROC) curve [20] and area under curve (AUC) value are common ways to evaluate the classifier performance. In addition, we also adopt other kinds of performance measures, including *precision*, *recall*, *F-measure* (F1), *Matthews correlation coefficient* (MCC) and *top-percentage* of essential proteins. The formulas are given as follows.

1. Precision: $\frac{TP}{TP+FP}$
2. Recall: $\frac{TP}{TP+FN}$
3. F-measure: $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
4. MCC: $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$

In this paper, the positive samples denote essential proteins. TP (*true positive*), FP (*false positive*), FN (*false negative*) and TN (*true negative*) denote the number of true positive, false positive, true negative and false negative numbers, respectively. The value n is the total number of predictions.

III. OUR METHOD

In this section, we will present our feature sets and our prediction method with SVM.

A. Feature Extraction

The feature set includes *topological properties* (T), such as bit string of double screening scheme and betweenness centrality related to physical interactions; *protein properties* (P), such as cell cycle and metabolic process; *sequence properties* (S), such as amino acid occurrence and average amino acid PSSM; and *other properties* (O), such as phyletic retention and essential index. There are totally 45 properties and 90 features, whose names and sources are shown in Tables I.

B. Our Method with SVM

This paper uses the SVM program developed by Chang and Lin, called *LIBSVM* [9]. In this paper, we adopt *radial basis function* (RBF) as the kernel function because it yields the best results. We apply a *modified sequential backward feature selection* method on the data set. Instead of using only accuracy to guide the feature selection, we propose a composite score C in terms of *precision*(P), *recall*(R), *F-measure*(F) and *MCC*(M) as the objective function, where $C = w_P * P + w_R * R + w_F * F + w_M * M$. The weights, w_P , w_R , w_F and w_M , are introduced for compromise between these performance measures. To allow scores associated with fewer features can compete with those with more features, an additional punishment is imposed to C . That is,

$$C(S) = w_P * P(S) + w_R * R(S) + w_F * F(S) + w_M * M(S) - (|S| - t) * u(|S| - t) * e,$$

where S denotes the selected feature subset. The unit step function $u(|S| - t) = 1$ as $|S| - t > 0$, otherwise $u(|S| - t) = 0$. $|S|$ and t denote the size of selected feature set S and the goal number of features specified by a user, respectively. Finally, to make sure the improvement over feature changes is not merely a random process, a threshold v is adopted. The value v is estimated approximately by comparing average score difference corresponding to features of sizes p and $p+1$ in the preliminary run. The value e is the penalty score when an additional feature is selected, which is slightly larger than v to encourage feature subsets of smaller sizes. The feature selection procedure is given in Procedure Feature-Selection.

IV. EXPERIMENTAL RESULTS

In this section, we will show our experimental results and compare them with the methods proposed by Hwang *et al.* [10] and Acencio *et al.* [1].

A. Feature Selection Results

During the feature selection procedure, k -fold cross-validations are carried out to compute performance scores and 50% of data elements, which are $(3840 + 975) * 50\%$ observations, are involved for the backward feature selection procedure. In the first run, parameters k , w_P , w_R , w_F , w_M , v , e , t and r are set to 2, 1, 1, 1, 1, 0.005, 0, 90 and 5, respectively, in an attempt to find out a feature subset that is best compromise among all performance measures. Because no goal number of selected features is imposed, the procedure will try to exploit all available feature combinations to achieve

TABLE I
THE NAMES OF PROTEIN FEATURES.

| ID | Property name | Type | Size | Sub-names |
|----|--|------|------|-----------|
| 1 | Phyletic retention [10] | O | 1 | |
| 2 | Bit string of double screening scheme [23] | T | 1 | |
| 3 | Amino acid occurrence [14] | S | 20 | A...Y |
| 4 | Nucleus [1] | P | 1 | |
| 5 | Betweenness centrality related to physical interactions [1] | T | 1 | |
| 6 | Neighbors' intra-degree [10] | T | 1 | |
| 7 | Essential index [10] | O | 1 | |
| 8 | Clique level [10] | T | 1 | |
| 9 | Degree related to all interactions [12] | T | 1 | |
| 10 | Common function degree [10] | O | 1 | |
| 11 | Clustering coefficient [10] | T | 1 | |
| 12 | Betweenness centrality related to all interactions [24] | T | 1 | |
| 13 | Other process [1] | P | 1 | |
| 14 | Density of maximum neighborhood component [5] | T | 1 | |
| 15 | Maximum neighborhood component [5] | T | 1 | |
| 16 | Closeness centrality [22] | T | 1 | |
| 17 | Degree related to physical interactions [1] | T | 1 | |
| 18 | Edge percolated component [4] | T | 1 | |
| 19 | Other localization [1] | P | 1 | |
| 20 | Open reading frames length [10] | O | 1 | |
| 21 | Cytoplasm [1] | P | 1 | |
| 22 | Average amino acid PSSM [14] | S | 20 | A...Y |
| 23 | Cell cycle [1] | P | 1 | |
| 24 | Transcription [1] | P | 1 | |
| 25 | Mitochondrion [1] | P | 1 | |
| 26 | Metabolic process [1] | P | 1 | |
| 27 | Bottleneck [16] [24] | T | 1 | |
| 28 | Cysteine count [14] | S | 1 | |
| 29 | Endoplasmic reticulum [1] | P | 1 | |
| 30 | Cysteine odd-even index [14] | S | 1 | |
| 31 | Average hydrophobic [14] | S | 1 | |
| 32 | Signal transduction [1] | P | 1 | |
| 33 | Average cysteine position [14] | S | 1 | |
| 34 | Outdegree related to metabolic interaction [1] | T | 1 | |
| 35 | Outdegree related to transcriptional regulation interaction [1] | T | 1 | |
| 36 | Indegree related to metabolic interaction [1] | T | 1 | |
| 37 | Average distance of every two cysteines [14] | S | 1 | |
| 38 | Transport [1] | P | 1 | |
| 39 | Betweenness centrality related to metabolic interactions [1] | T | 1 | |
| 40 | Protein length [14] | S | 1 | |
| 41 | Betweenness centrality transcriptional regulation interactions [1] | T | 1 | |
| 42 | Identicalness [1] | O | 1 | |
| 43 | Indegree related to transcriptional regulation [1] | T | 1 | |
| 44 | Cysteine location [14] | S | 5 | 1...5 |
| 45 | Average hydrophobicity around cysteine [14] | S | 4 | 1...4 |

the best performance. In this run, the number of resultant selected features is 18. For the subsequent runs, t is set starting from 17 till that performances are significantly worse than Hwang's *et al.* [10] results. Here, we use the *bootstrap cross-validation* [7] for significance examination. Parameters k , w_P , w_R , w_F , w_M , v , e and r are set to 2, 1.03, 1, 1, 1, 0.005, 0.01, and 5 in order to obtain reduced feature subsets. The reason to set $w_P = 1.03$ is to encourage features that prevent the true positive rate from dropping too much. In addition, e is greater than v to allow the procedure to be able to proceed to fewer features.

For feature subsets of the same size, we run 10 times and obtain 10 feature subsets. Since resultant feature subsets with identical feature number may be slightly different in different runs, we compare them with a 5-fold cross-validation. The

one with the highest score are finally preserved for further examination. The selected feature subsets of various sizes are as listed in Table II. The second column shows all selected features. The subsequent columns represent feature subsets of sizes 4, 5, ..., 18 preceded by a prefix character 'N'. In each feature subset, a star(*) mark is used to indicate which feature is included.

In order to compare our results with those in the previous studies, all performance measures are calculated with a 10-fold cross-validation experiment. To account for the randomness, we follow Hwang's method to perform every 10-fold cross-validation 10 times. In addition to the precision, recall, F-measure and MCC values, the prediction values are finally imported into another software developed by Tobias's *et al.* [19] [17] in order to calculate average ROC curves and AUC

TABLE II
THE FEATURES WHICH ARE SELECTED IN THE SUBSETS WITH VARIOUS SIZES.

| | Feature | N4 | N5 | N6 | N7 | N8 | N9 | N10 | N11 | N12 | N13 | N14 | N15 | N16 | N17 | N18 | TOT |
|----|--------------------------|----|----|----|----|----|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1 | PR (phyletic retention) | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | 15 |
| 2 | EI (essentiality index) | * | * | * | * | * | * | * | * | * | * | * | * | * | * | * | 15 |
| 3 | Cytoplasm | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | 14 |
| 4 | Nucleus | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | 13 |
| 5 | Bit string of DSS | | | * | * | * | * | * | * | * | * | * | * | * | * | * | 12 |
| 6 | Occurrence of A.A. I | | | | * | * | * | * | * | * | * | * | * | * | * | * | 12 |
| 7 | Occurrence of A.A. W | | | | | * | * | * | * | * | * | * | * | * | * | * | 11 |
| 8 | Endoplasmic reticulum | | | | | | * | * | * | * | * | * | * | * | * | * | 10 |
| 9 | Occurrence of A.A. G | | | | | | | * | * | * | * | * | * | * | * | * | 6 |
| 10 | Other process | | | | | | | | | | * | * | * | * | * | * | 5 |
| 11 | Occurrence of A.A. S | | | | | | | | * | * | * | * | * | * | * | * | 5 |
| 12 | Average PSSM of A.A. R | | | | | | | * | * | * | * | * | * | * | * | * | 5 |
| 13 | KLK (clique level) | * | | | | | | | * | * | * | * | * | * | * | * | 5 |
| 14 | Average hydrophobic | | | | | | | | * | * | * | * | * | * | * | * | 4 |
| 15 | B.C. related to PRs | | | | | | | | * | * | * | * | * | * | * | * | 3 |
| 16 | Cell cycle | | | | | | | | * | * | * | * | * | * | * | * | 3 |
| 17 | Average HYD around C 2 | | | | | | | | * | * | * | * | * | * | * | * | 3 |
| 18 | Average PSSM of A.A. P | | | | | | | * | * | * | * | * | * | * | * | * | 3 |
| 19 | Indegree related to T.R. | | | | | | | | | | | | * | * | * | * | 2 |
| 20 | B.C. T.R. interactions | | | | | | | | | | | | * | * | * | * | 2 |
| 21 | Signal transduction | | | | | | | | | | | | | * | * | * | 2 |
| 22 | Other localization | | | | | | | | | | | | | * | * | * | 2 |
| 23 | DMNC | | * | * | | | | | | | | | | * | * | * | 2 |
| 24 | Occurrence of A.A. E | | | | | | | | | | * | | | * | * | * | 2 |
| 25 | Occurrence of A.A. T | | | | | * | * | | | | | | | * | * | * | 2 |
| 26 | Average PSSM of A.A. E | | | | | | | | | | * | * | * | * | * | * | 2 |
| 27 | Cysteine count | | | | * | | | | | | | | | | | | 1 |
| 28 | Average HY. around C 1 | | | | | | | | | | * | * | * | * | * | * | 1 |
| 29 | Average PSSM of A.A. A | | | | | | | | | | * | * | * | * | * | * | 1 |
| 30 | Average PSSM of A.A. Q | | | | | | | | | * | * | * | * | * | * | * | 1 |
| 31 | Average PSSM of A.A. Y | * | | | | | | | | | | | | | | | 1 |

DSS: double screening scheme, A.A.: amino acid, B.C. betweenness centrality, T.R.: transcriptional regulation
HYD: hydrophobicity, PR: physical interaction, A...Y: amino acid abbreviation, TOT: total

Procedure: Feature-Selection

input : S_0 : set of all available features, where

$|S_0| = n$

k : number of cross-validations

r : maximal number to retry

output: Selected feature subset S

begin

if $t \geq n$ **then** $m = n$ **else** $m = n - t$;

for $p = 1$ **to** m **do**

for $i = 1$ **to** r **do**

$X = \phi$;

foreach $j \in \{1, 2, \dots, |S|\}$ **do**

$X = X \cup S(j)$;

end

foreach $s \in X$ **do**

 Use k -fold CV to calculate $C(S - s)$;

end

$Q = \operatorname{argmax}_{s \in X} \{C(S - s) - C(S) | C(S - s) - C(S) \geq v\}$;

if $Q \neq \phi$ **then** $S = Q$ and exit for;

if $i = r$ **then** stop and output S ;

end

end

end

values. The reported performance measures, including AUC, Precision, Recall, F1, and MCC values, are averaged over 10 10-fold cross-validations. In this paper, we use predictors with Hwang's 10 features as a benchmark because it yields distinguished results in terms of performance and feature number. Predictors corresponding to the remaining feature subsets are compared with Hwang's by *bootstrap cross-validation* method with 200 bootstrap samples for both imbalanced and balanced data sets.

B. Significant Tests

Table III shows the average values of five performance measures in 10 experiments for given various feature subsets with all data involved for 10-fold cross-validations. The first two rows show Hwang's and Acencio's results while those produced by our method are listed in the subsequent rows. Values enclosed by parentheses in the first column represent the numbers of features. Significance tests are performed by the *bootstrap cross-validations* over 200 bootstrap samples. The first plus(+) or minus(-) symbol following each value represents significantly better or worse than Hwang's result. For example, the AUC corresponding to N6 is significantly worse than that of Hwang while the recall value is significantly better than that of Hwang. The second symbols are used to indicate the significance between two neighboring rows among feature subsets with a prefix name 'N'. For example, the AUC

value of N7 is significantly better than that of N6 and its recall value is significantly better than N8. The underlined value represents the best performance measure in each column. The last row shows results with the full set of 90 features. All measures are quite close to those of Hwang, except for most significant in the AUC value.

This table shows that Hwang's predictors outperform Acencio's in all performance measures, even the feature size is considerably smaller. For those of N8, its AUC value is as good as Hwang's while the rest of measure values are dominantly above Hwang. When the feature size is increased to above 8, the improvement to Hwang's in most results, except for precision values, are consistently significant.

Table IV shows the average values of performance measures in 10 balanced experiments. It is shown that Hwang's predictors outperform Acencio's in most performance measures. It is observed that feature subsets with sizes above 5 seems good enough to compete with Hwang's while more than 12 features are required to achieve significance.

After we inspect the feature subsets in Table II, we can find that the most prominent features indeed come in diverse sources, including sequence, topology and protein processes. Among these 31 features, cytoplasm, endoplasmic reticulum, nucleus, bit string of double screening scheme, amino acid occurrence I, amino acid occurrence W, PR (phyletic retention) and EI (essentiality index) are selected more than 10 times. PR and EI are even included in all feature subsets and thus be considered as the most important factors to identify essential proteins. It turns out that N9 and N8 are the feature subsets covering most of these 8 features, which outperforms Hwang's results in all performance measures except for AUC values. In addition to the small sizes in the feature subsets, these two subsets contain two amino acid occurrences, which are relatively easier to extract. Just like N8 and N9, predictors built by the feature subsets of 8 or more features are consistently surpass Hwang's in almost all performance measures. By analyzing Table III and Table IV, we recommend to use N6, N9 and N16. Because by choosing one additional feature, they achieve more significant items than N5, N9 and N15 in performance measures.

V. CONCLUSION AND FUTURE WORK

In this paper, we incorporate several kinds of protein properties, including sequence, topology and protein processes. There are totally 45 properties and 90 features. To deal with data imbalance problem, we propose a modified sequential backward feature selection strategy. Then, we apply the SVM classification to identify essential proteins. We build several SVM models for both imbalanced data set and balanced data set. To compare with some previous results, we adopt several kinds of performance measures and carry out significant tests. The F-measure and MCC of our best results on the imbalanced data set are 0.549 and 0.495, respectively. For the balanced data set, we obtain 0.77 and 0.545, respectively. From our experimental results, we discover that the performance of our models is better than others starting from the N9 model. For

getting high accuracy, we suggest the N16 model (16 features). If one would prefer the feature set with small size, we suggest to take the N9 model (9 features). We also list the features that are significant in performance. These features may be crucial to discriminate essential proteins. Since the obtained feature set are quite small in size, we hope this would facilitate interpretation in the future.

There are some possible ways to further improve the prediction accuracy. Some features related to the protein primary structure may be helpful to identify essentiality. For example, Lin *et al.* [15] derived bi-gram and spaced bi-gram features from protein sequences and demonstrated good performance in protein fold classification. Besides, Gupta *et al.* [8] proposed to transform protein sequences into signals and adopt conventional signal processing techniques to extract features. The experimental results also show the effectiveness in protein class classification. In addition to involving more features, another possible way to improve the performance is to incorporate other tools or to construct hybrid predictors. According to the literature [13], the possibility for improvement arises from the fact that the combined classifiers must be distinct enough. Otherwise, some negative effects would impose on the classifier ensemble. We will devote our efforts into these issues in the future.

REFERENCES

- [1] M. L. Acencio and N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information," *BMC Bioinformatics*, vol. 10, no. 1, pp. 290–307, 2009.
- [2] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.
- [3] K. M. Cadigan, U. Grossniklaus, and W. J. Gehring, "Functional redundancy: The respective roles of the two sloppy paired genes in drosophila segmentation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 14, pp. 6324–6328, 1994.
- [4] C.-S. Chin and M. P. Samanta, "Global snapshot of a protein interaction network: a percolation based approach," *Bioinformatics*, vol. 19, pp. 2413–2419, 2003.
- [5] C.-H. Chin, C.-W. Ho, and M.-T. Ko, "Prediction of essential proteins and functional modules from protein-protein interaction networks," Ph.D. dissertation, National Central University, Chung-Li, Taiwan, 2010.
- [6] L. M. Cullen and G. M. Arndt, "Genome-wide screening for gene function using RNAi in mammalian cells," *Immunology and Cell Biology*, vol. 83, no. 3, pp. 217–223, 2003.
- [7] W. J. Fu, R. J. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 9, pp. 1979–1986, 2005.
- [8] R. Gupta, A. Mittal, and K. Singh, "A time-series-based feature extraction approach for prediction of protein structural class," *EURASIP Journal on Bioinformatics and Systems Biology*, 2008.
- [9] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- [10] Y.-C. Hwang, C.-C. Lin, J.-Y. Chang, H. Mori, H.-F. Juan, and H.-C. Huang, "Predicting essential genes based on network and sequence analysis," *Molecular BioSystems*, vol. 5, no. 12, pp. 1672–1678, 2009.
- [11] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.
- [12] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, 2001.

TABLE III
PERFORMANCE COMPARISON FOR IMBALANCED DATA SET.

| | AUC | Precision | Recall | F-measure | MCC |
|-------------|---------------|-----------|---------------|---------------|---------------|
| Hwang(10) | 0.775 | 0.743 | 0.343 | 0.469 | 0.432 |
| Acencio(23) | 0.707 (-) | 0.675 (-) | 0.121 (-) | 0.204 (-) | 0.228 (-) |
| N4 | 0.744 (-) | 0.782 (+) | 0.327 (-) (-) | 0.461 (-) (-) | 0.439 |
| N5 | 0.727 (-) | 0.741 (-) | 0.387 (+) | 0.509 (+) | 0.461 (+) |
| N6 | 0.730 (-) | 0.752 | 0.395 (+) | 0.518 | 0.472 |
| N7 | 0.761 (+) | 0.767 | 0.386 (+) | 0.513 (+) | 0.473 |
| N8 | 0.772 | 0.755 | 0.371 (-) | 0.498 (+) (-) | 0.457 (-) |
| N9 | 0.782 | 0.749 | 0.382 (+) (+) | 0.506 (+) | 0.462 (+) |
| N10 | 0.781 | 0.751 | 0.399 (+) | 0.521 (+) | 0.474 (+) |
| N11 | 0.786 (+) | 0.752 | 0.402 (+) | 0.524 (+) | 0.476 (+) |
| N12 | 0.798 (+) | 0.759 | 0.409 (+) | 0.532 (+) | 0.485 (+) |
| N13 | 0.789 (+) | 0.748 | 0.433 (+) (+) | 0.549 (+) | 0.495 (+) |
| N14 | 0.802 (+) | 0.749 | 0.397 (+) (-) | 0.519 (+) (-) | 0.471 (+) (-) |
| N15 | 0.801 (+) | 0.763 | 0.406 (+) | 0.530 (+) | 0.485 (+) (+) |
| N16 | 0.814 (+) (+) | 0.762 | 0.401 (+) | 0.525 (+) | 0.480 (+) |
| N17 | 0.814 (+) | 0.761 | 0.407 (+) | 0.530 (+) | 0.484 (+) |
| N18 | 0.811 (+) | 0.751 | 0.411 (+) | 0.531 (+) | 0.482 (+) |
| N90 | 0.829 (+) (+) | 0.738 | 0.355 (-) | 0.479 (-) | 0.438 (-) |

*With the polynomial kernel function, Hwang's precision, recall and MCC values are 0.77, 0.23 and 0.36, respectively.

TABLE IV
PERFORMANCE COMPARISON FOR THE BALANCED DATA SET.

| | AUC | Precision | Recall | F-measure | MCC |
|-------------|-----------|-----------|---------------|-----------|-----------|
| Hwang(10) | 0.822 | 0.778 | 0.720 | 0.748 | 0.516 |
| Acencio(23) | 0.808 (-) | 0.696 (-) | 0.734 | 0.714 (-) | 0.414 (-) |
| N4 | 0.811 (-) | 0.777 | 0.716 | 0.745 | 0.512 |
| N5 | 0.824 (+) | 0.778 | 0.735 (+) | 0.756 | 0.527 |
| N6 | 0.827 (+) | 0.778 | 0.739 | 0.758 | 0.530 |
| N7 | 0.831 | 0.779 | 0.733 | 0.755 | 0.526 |
| N8 | 0.826 | 0.786 | 0.721 | 0.752 | 0.527 |
| N9 | 0.833 (+) | 0.791 | 0.735 (+) | 0.762 | 0.541 |
| N10 | 0.834 | 0.789 | 0.736 | 0.761 | 0.540 |
| N11 | 0.831 | 0.784 | 0.737 | 0.760 | 0.535 |
| N12 | 0.829 | 0.779 | 0.732 (+) | 0.755 | 0.526 |
| N13 | 0.834 (+) | 0.788 | 0.730 | 0.758 | 0.535 |
| N14 | 0.836 (+) | 0.777 | 0.743 (+) | 0.759 (+) | 0.530 |
| N15 | 0.843 (+) | 0.784 | 0.748 (+) | 0.766 (+) | 0.542 (+) |
| N16 | 0.842 (+) | 0.777 | 0.756 (+) | 0.767 (+) | 0.540 (+) |
| N17 | 0.847 (+) | 0.778 | 0.763 (+) | 0.770 (+) | 0.545 (+) |
| N18 | 0.840 (+) | 0.779 | 0.740 (+) (-) | 0.759 | 0.531 |
| N90 | 0.839 (+) | 0.760 (-) | 0.753 (+) | 0.757 (+) | 0.516 |

*With the polynomial kernel function, Hwang's precision, recall, F-measure and MCC values are 0.763, 0.713, 0.737 and 0.492, respectively.

- [13] L. Kuncheva, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.
- [14] C.-Y. Lin, C.-B. Yang, C.-Y. Hor, and K.-S. Huang, "Disulfide bonding state prediction with svm based on protein types," *Bio-Inspired Computing: Theories and Applications*, pp. 1436–1442, 2010.
- [15] C. Y. Lin, K.-L. Lin, C.-D. Huang, H.-M. Chang, C. Y. Yang, C.-T. Lin, C. Y. Tang, and D. F. Hsu, "Feature selection and combination criteria for improving predictive accuracy in protein structure classification," *IEEE Transaction on nanobioscience*, vol. 6, no. 2, pp. 186–196, 2007.
- [16] N. Pržulj, D. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, pp. 340–348, 1998.
- [17] R. Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: <http://www.R-project.org>
- [18] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.
- [19] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, pp. 3940–3941, 2005.
- [20] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.
- [21] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.
- [22] S. Wuchty and P. F. Stadle, "Centers of complex networks," *Journal of Theoretical Biology*, pp. 45–53, 2003.
- [23] Z.-J. Yang, C.-B. Yang, and C.-T. Tseng, "Prediction for essential proteins with the support vector machine," in *Proc. of National Computer Symposium, Workshop on Algorithms and Bioinformatics*, 2011, pp. 26–33.
- [24] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics," *PLoS Computational Biology*, vol. 3, pp. 713–720, 2007.