Vietnam J Comput Sci (2014) 1:169–178
DOI 10.1007/s40595-014-0018-5

REGULAR PAPER

Alternating decision tree algorithm for assessing protein interaction reliability

Min Su Lee · Sangyoon Oh

Received: 21 November 2013 / Accepted: 10 March 2014 / Published online: 25 March 2014

Abstract This paper presents a machine learning approach for assessing the reliability of protein–protein interactions in a high-throughput dataset. We use an alternating decision tree algorithm to distinguish true interacting protein pairs from noisy high-throughput data using various biological attributes of interacting proteins. The alternating decision tree algorithm is used both for identifying discriminating biological features that could be used for assessing protein interaction reliability and for constructing a classifier to identify true positive interacting pairs. Experimental results show that the proposed approach has a good performance in distinguishing true interacting protein pairs from noisy protein–protein interaction data. Moreover, our alternating decision tree classifier supplemented with domain knowledge may be helpful to understand the biological conditions in connection with interacting protein pairs.

Keywords Alternating decision tree algorithm · Machine learning · Protein–protein interaction · Reliability

1 Introduction

Machine learning algorithms have been successfully applied to many bioinformatics problems. One of the key issues in bioinformatics is the analysis of protein–protein interactions (PPIs) [1], which is the fundamental basis of cellular operations. PPI knowledge is essential in predicting unknown functions of proteins [2–6], in clarifying biological pathways [7–10], and in understanding biological mechanisms in the disease [11]. Conventional studies on PPI have examined each interacting pair separately in terms of the physical and chemical properties of proteins. Thanks to advanced molecular-level technologies, large amounts of PPI data have been collected through high-throughput experiments by testing for physical interactions among multiple proteins. They include genome-scale Yeast Two-Hybrid assays (Y2H) [12–14] and protein complex identification methods through mass spectrometry [15,16].

While vast amounts of data obtained from high-throughput experiments allow for efficient identification of different kinds of PPI information, they are prone to higher false positive rates than small-scale studies [17–22]. Some studies have reported that approximately half of the interactions obtained from high-throughput data may be false positives [17,19]. This necessitates an additional experimental or computational method to estimate the reliability of each PPI precisely. To do this, selecting relevant properties of PPI as circumstantial evidences, and adopting efficient computational methodologies are important. Several circumstantial features have previously been used to identify true PPIs from high-throughput experimental data in yeast [17,23–27].

The intersection of multiple high-throughput PPI datasets can be effective in obtaining more reliable PPIs. If an interaction is detected from two distinct experiments, the interaction can be regarded as more reliable. However, different experimental methods often generate different levels of information. For example, mass spectrometry detects which proteins are part of a stable complex, but does not necessarily indicate which proteins in the complex have a direct interaction. Mass spectrometry might also fail to uncover transient or weak interactions. Y2H might not detect interactions that are

M. S. Lee
Computational Omics Lab, School of Informatics and Computing, Indiana University, Bloomington, IN, USA

S. Oh (✉)
Department of Computer Engineering, Ajou University, Suwon, Republic of Korea
e-mail: syoh@ajou.ac.kr

dependent on post-translational modifications or that should be stabilized by the presence of another protein. Moreover, since PPI is very sensitive to experimental conditions, PPI data produced at different research groups are substantially different although the same technologies are used [12,13]. Due to these limitations in high-throughput technologies, the coverage of intersections is very small even with the immense amount of PPI datasets [17].

The conserved interactions between different species are named interlog [23]. If two proteins interact in one species, their ortholog proteins are more likely to interact with each other. This property has been used to enhance the confidence of prediction in high-throughput data [23,28].

Interaction network topology is another means of identifying true interactions. With the protein interaction network, the interaction generality measure (IG2) based on the topological properties [29], and statistical and topological correlation between the paired proteins [30] have been studied to assess the reliability of an interaction. Since proteins with more than one interaction partner are rare cases, these methods have a low sensitivity (i.e., a low true positive rate) although they have a high specificity (i.e., a low false positive rate).

Most of these methods use only one criterion at once and need an entire genome-scale PPI dataset to assess the reliability of each PPI pair. Each biological feature may also have false and missing values. Moreover, it is very difficult for biologists to define the proper cutoff values of confidence scores to distinguish between true positives and false positives. Since different biological features may cover different subsets of interacting pairs, a combination of various existing methods would be more effective.

Recently, computational methods for assessing the reliability of putative protein interaction have also been studied based on supervised machine learning techniques, such as Bayesian network [31,32], and maximum likelihood estimation [25]. Patil et al. [31] used a combination of three genomic features—Pfam domains, Gene Ontology annotations and sequence homology—to assign reliability to the protein–protein interaction. And Bayesian network approach was used to compute the likelihood ratio to be real interactions. Deng et al. [20] used another three attributes which are the distribution of gene expression correlation coefficients, the reliability based on gene expression correlation coefficient, and the accuracy of protein function predictions. And maximum likelihood method was used to estimate the reliability of protein interaction datasets based on the three attributes. Lin et al. [32] proposed a Bayesian network-based approach which assigns likelihood scores to individual protein pairs based on interlogs and their genomic features derived from microarray data and gene ontology.

In this paper, we present a new evaluation system for PPI datasets that can distinguish true interacting protein pairs from noisy datasets. This work is inspired by our previous work that performs comparative study of classification methods for protein interaction verification system [33]. Through the empirical comparative study, K-nearest neighborhood and decision tree algorithm are the two top performance producers among other methods. We adopted a decision tree algorithm among those two methods since it can produce an interpretable output classifier with good performance. In this system, we use an alternating decision tree algorithm [27] that dynamically selects discriminating features among various attributes related with PPIs and trains an interpretable classifier that can distinguish true interacting protein pairs with confidence scores from noisy datasets. The system may help not only to identify relevant circumstantial evidences among various biological features for assessing reliability of PPIs, but also to understand the characteristics of true interacting protein pairs based on the alternating decision tree. The statistical evaluation of the system using tenfold cross-validation shows that the system performs well in terms of various performance measures. Specifically, the average rates of accuracy, sensitivity, precision, F-measure, and MCC (Matthew's correlation coefficient) are 97.13, 96.91, 97.33, 97.12, and 94.26 %, respectively.

The contributions of this study can be summarized as follows: Firstly, our proposed protein interaction evaluation system shows a good performance in distinguishing true interacting protein pairs from a noisy PPI dataset compared with similar approaches. Secondly, we use a novel negative example generation method for assessing protein interaction reliability. This helps to derive more reliable and high performance output model. Third, by applying ADTree algorithm, interpretable prediction model can be derived which shows the biological conditions of interacting protein pairs with confidence score. Also, missing values in query data can be more naturally handled by considering the reachable decision nodes in ADTree.

2 Materials and methods

In this section, we present a protein interaction evaluation system to assess the reliability of PPI data obtained from high-throughput experiments. To separate true positives and false positives from the putative PPI dataset, we have developed a classification model based on an alternating decision tree algorithm. Figure 1 shows the basic system architecture of the classification model. In Fig. 1, the solid arrows indicate the training process to construct a classifier, and the dotted arrows show the classifying process of querying high-throughput protein interaction data. Our evaluation system consists of a PPI database, a PPI annotation database, a computation module of circumstantial evidences for each PPI pair, an alternating decision tree algorithm for selecting relevant genomic features and training a classification model,

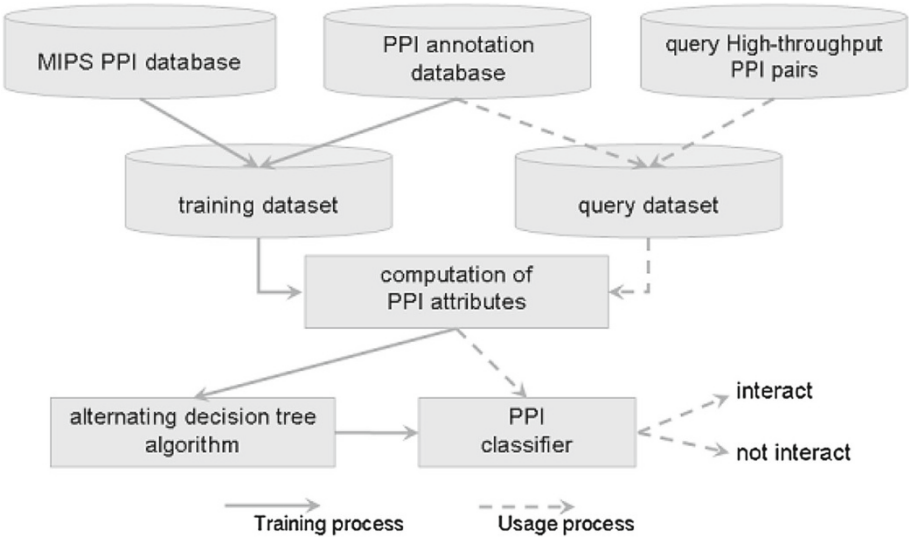

Fig. 1 System architecture for classifying high-throughput protein interaction pairs into positives and negatives

and a PPI classifier generated by the algorithm. The system first trains from a collection of protein pairs that consist of positive and negative PPI examples and their genomic features based on an alternating decision tree algorithm. The trained classifier can be used to distinguish true positive PPI pairs from the noisy query dataset based on the values of genomic evidences of each pair.

2.1 Training dataset

Genome-scale PPI data are currently available for *S. cerevisiae* [12,13,15,16], *H. pylori*, *C. elegans*, *D. melanogaster*, and *H. sapiens*. Since the number of protein–protein interaction pairs for yeast is larger than others, we evaluated our system using the yeast PPI data. The proposed system first trains with a set of reliable positive and negative protein interaction pairs. The quality of training dataset is a critical point to construct a reliable prediction model. An ideal training dataset should be independent from the discriminating features, sufficiently large for reliable statistics, and free of systematic biases [34,35]. Hence, the selection of a reliable training dataset is a definitely important procedure for constructing a robust PPI evaluation system.

Positive protein pairs were extracted from the MIPS database [36,37]. Manually collected MIPS database has been regarded as a trusted standard dataset. The MIPS database contains PPI pairs and protein complexes in yeast that provides annotations for experimental methods in PPIs. We used a protein complex dataset as a reliable positive protein pair by decomposing it into binary interactions [34]. The MIPS protein complex dataset consists of known protein complexes based on data collected from the literature, and most of these are derived from small-scale studies.

Unlike positive protein interaction pairs, negative protein interaction pairs are harder to define, because there is no experimentally verified set of non-interacting proteins. In fact, it is almost impossible to validate non-interacting protein pairs theoretically and empirically through rigorous wet lab experiments. Hence, we created negative datasets synthetically based on co-localization enrichment of interaction between two proteins. Interactions are strongly enriched between proteins that co-localize, but the degree of enrichment varies widely by compartment. Huh et al. [38] determined the subcellular localizations of each interacting protein pair and the fraction of the total number of interactions occurring for each localization pair. They compared the interaction between specific compartments to a randomized interaction set, showing that some compartments interact preferentially with others. We used this localization enrichment data among 22 subcellular locations to derive non-interacting proteins. To do this, we first generated random PPI pairs using proteins that appeared in positive datasets. Secondly, since proteins are located in several subcellular compartments by shuttling or transporting, we found all combinations of subcellular locations of two proteins among randomly generated PPI pairs. After that, we selected negative protein pairs whose minimum enrichment value for all possible co-localization cases was zero based on Huh et al.'s experimental results. Since previous researches have reported that approximately half of the interactions obtained from high-throughput data may be false positives [17,19], we finally adjusted the number of instances of negative protein pairs to that of corresponding positive protein pairs in order to conform the distribution of a sample to that of a population. As a result, our training dataset consisted of 8,250 positives and 8,250 negatives.

We think that this strategy of deriving negative datasets is more sophisticated than Jansen et al.'s [34] approach.

Jansen et al. derived negative protein pairs by selecting PPI pairs in different subcellular compartments based on localization attributes, including nucleus, mitochondria, cytoplasm, membrane, and secretory pathway [34,39].

This plot demonstrates overlapping proportion of each PPI attributes to positive and negative instances according to the type of attributes. 'pos' means overlap between each positive instances in a dataset and PPI attributes, and 'neg' means overlaps between each negative instances in a dataset and PPI attributes.

2.2 Biological attributes for evaluating PPIs

As we described in Sect. 1, each biological attribute, by itself, is only a weak predictor of protein interactions. Assessing the reliability of a protein interaction pair can be improved by integrating different biological attributes because the task depends on the existence of circumstantial evidence that supports it. When multiple distinct attributes all support a candidate interaction pair, the confidence of PPI increases. Different attributes may cover different subsets of interacting protein pairs, and in this case, attribute integration can increase the coverage.

Hence, we collected open biological features of two proteins and integrated seven biological attributes that can be used as indicators of putative interacting proteins. Table 1 gives some explanation of the biological attributes, which include the name of attribute, description, and the range of attribute value.

Interacting proteins whose transcripts are co-expressed are more likely to be credible [18,40,41]. Hence we computed Pearson's correlation of mRNA expression levels between two proteins (attribute 1 in Table 1) using publicly available time-series expression datasets (Rosetta compendium and yeast cell cycle). Since two interacting proteins should be present in a similar amount, the absolute expression levels of two proteins (attribute 2 in Table 1) and the absolute amount of two proteins (attribute 3 in Table 1) were used [42].

Another important property of interacting proteins is that two proteins in the same biological process are more likely to interact. Attributes 4 and 5 in Table 1 are this functional similarity of two proteins in terms of biological ontology such as MIPS Functional catalog [43,44] and Gene Ontology about biological processes [45]. The similarity measure between two proteins was quantified by computing the frequencies of a set of functional terms that two proteins share based on semantic similarity measure [46]. In general, a lower frequency means a higher specificity of the functional term for the two proteins. Hence the semantic similarity of shared functional terms can be inferred by the frequency of the term.

Attributes 6 and 7 are related to the essentiality of two proteins in a cell. Attribute 6 computes a marginal essentiality of two proteins. Marginal essentiality is a measure of the importance of a non-essential gene in a cell that could be derived from topological characteristics of protein interaction networks [47]. Attribute 7 is based on the hypothesis that if two proteins are in the same protein complex or pathway, they are likely to share the essentiality in the cell because they should perform the same function together [36].

We considered some additional biological attributes for assessing PPIs such as co-regulation and interlog. However, we filtered them out since the training dataset rarely contains those attributes. Figure 2 shows the proportion of each attribute that appeared in the training dataset with respect to the positive and negative instances. Since these attributes can be computed only when both proteins have a corresponding annotation, the proportion of the attributes that appeared in the datasets was rather smaller.

Figure 3a shows an example of a decision tree in a graph, and Fig. 3b shows an alternating decision tree. Here, A and B are names of attributes, 1 and m are values of attributes, and Class +1 and Class −1 are

Table 1 Description of biological attributes

	Name	Description	Range
1	mRNA co-expression	Pearson correlation of mRNA expression levels between two proteins using microarray dataset	$-1 \leq x \leq 1$
2	Absolute mRNA expression	Similarity of expression levels of two proteins	$0 \leq x < 16$
3	Absolute protein abundance	Similarity of abundance levels of two proteins	$0 \leq x < 10$
4	MIPS functional similarity	Specificity of common MIPS functional category of two proteins	$0 \leq x < 7$
5	GO functional similarity	Specificity of common GO biological process category of two proteins	$0 \leq x < 7$
6	Marginal essentiality	Quantitative measure of a non-essential gene to a cell based on topological property within the interaction network	$-20 < x < 0$
7	Co-essentiality	Whether two proteins are both essential or not	0: lethal 1: viable 2: lethal/viable

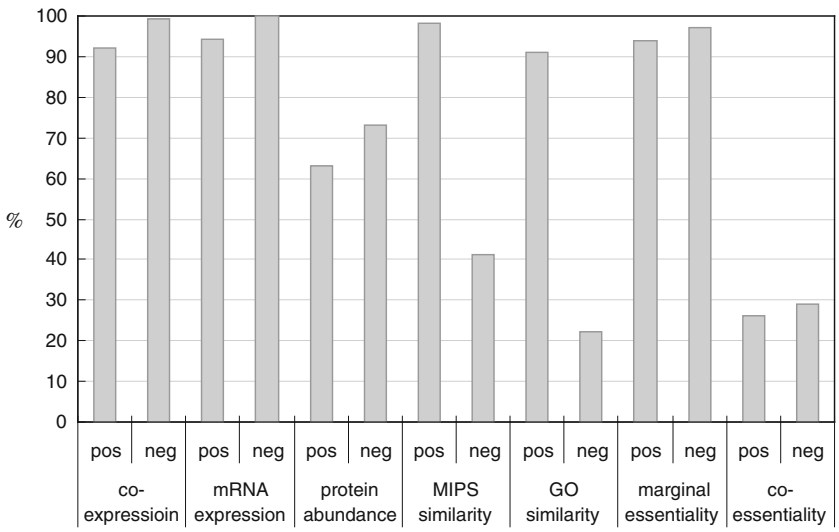

Fig. 2 Overlaps between dataset and attributes

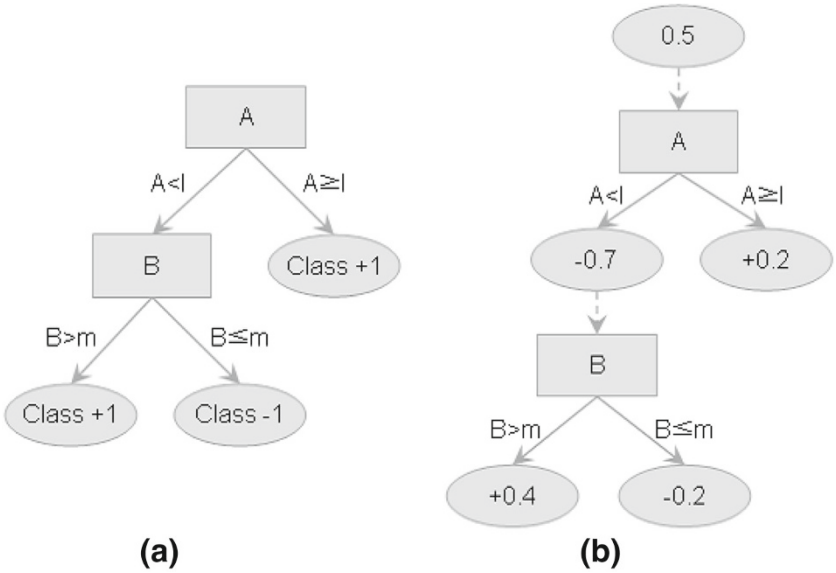

Fig. 3 Examples of Tree-based classifiers

target classes of this classification task. In an alternating decision tree, unlike a decision tree, the classification result is the sign of the sum of the predictions along the path, instead of the label of the leaf.

2.3 Alternating decision tree algorithm

We applied a kind of decision tree algorithm to classify PPI pairs into positive and negative interactions. Decision tree algorithms have been successfully used to predict categorical class labels, such as positive and negative (Fig. 3a). A decision tree algorithm constructs a tree-like classification model which consists of nodes, branches and leaves. Each node in the tree specifies some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. Each leaf node represents one of the target classes. To build a decision tree, it basically chooses an attribute that provides the maximum degree of discrimination for target classes. The information gain is a good measure to determine how well a given attribute separates the training instances according to their target class. The attribute with the highest information gain is chosen as the attribute for the current node, and the instances are partitioned accordingly. A resulting decision tree is composed of hierarchical if-then rules for values of attributes. Then, a decision tree algorithm classifies instances by evaluating their values of attributes from the root to some leaf node, and provides the classification results of the instance.

Decision tree algorithms have several advantages. Since decision trees use an intuitive white box model, it easily predicts the target class of query data using Boolean logic. Especially, decision trees are simple to understand and interpret. Moreover, decision tree algorithms are robust and scalable for large data and they can train a classification model in reason-

able processing time. Hence the decision tree model is effective not only to filter noisy PPI pairs from high-throughput experimental PPI data, but also to gain important insights that describe circumstantial evidences of PPIs.

However, general decision tree algorithms also have some disadvantages. Sometimes, the number of nodes generated from a decision tree algorithm is too large to interpret. Moreover, the performance of general decision tree algorithms is not usually better than that of functional or statistical machine learning algorithms, such as neural networks, support vector machines, and Bayesian networks. Recently, decision tree-based algorithms have strictly improved and become sophisticated in combination with other methods such as boosting [27] or logistic regression [48].

An alternating decision tree [27] is a combination of a decision tree and boosting that generates classification rules often smaller in number of nodes and easier to interpret. Especially, an alternating decision tree gives a measure of confidence that is called classification margin.

An alternating decision tree algorithm can be defined as a sum of simple rules. It uses a generalized representation for classification rules that consists of alternating layers of prediction nodes (represented by ellipses in Fig. 3b) and splitter nodes (represented by rectangles in Fig. 3b). The values in a root node represent the initial probability for assigning the target class according to the training dataset. Alternating decision tree classifiers are then built according to a particular structure using boosting wherein simple rules are successively added to the alternating decision tree classifier until the unit classifier of the tree exhibits satisfactory performance. Since boosting iteration adds three nodes (one splitter node and two prediction nodes) to the tree, more boosting iterations will result in larger and potentially more accurate trees. Unlike original decision trees: an instance is mapped into a path along the tree from the root to one of the leaves and output is the label of the leaf, the classification result of an alternating decision tree became the sign of the sum of the predictions along the multi-path associated with the given instances. When some feature values are unknown, the alternating decision tree algorithm only considers the reachable decision nodes [27]. Since the algorithm can handle missing values in a dataset more naturally, the alternating decision tree algorithm can be applied to analyzing our PPI dataset which includes lots of missing values. Also, the alternating decision tree algorithm has shown competitive performances and has produced smaller and intuitive classification rules than general decision tree algorithms.

3 Experimental results and discussion

To learn a prediction model for assessing protein interaction reliability, we first prepared training interaction pairs which are labeled with a target class, as described in Sect. 2.1. Then, seven biological attributes are integrated into the training pairs. Figure 2 demonstrates the proportion of each attribute appeared in the training dataset with respect to the positive and negative instances. We used the same number of positive and negative interactions for the training dataset of the alternating decision algorithm. We set the number of boosting interaction as 10 by considering both the accuracy and complexity of the system. The resulting alternating decision tree consisted of 10 splitter nodes and 21 prediction nodes. An exhaustive search method was used to build the alternating decision tree. Once an alternating decision tree is built, the system classifies any input PPI pairs into positives and negatives by searching possible paths along the tree and selecting the most confident prediction.

We computed the performances of the trained alternating decision tree classifier with tenfold cross-validation. Tenfold cross-validation means that the available examples are partitioned into ten disjoint subsets. The cross-validation procedure then runs ten times, and each time the procedure uses one of the ten subsets as the test set and the others for training sets. We used various performance criteria to evaluate the effectiveness of our system with the composition of training dataset. These criteria are calculated based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The performance criteria that we used are as follows:

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})\times 100}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}$$

$$\text{Sensitivity} = \frac{\text{TP}\times 100}{\text{TP}+\text{FN}} = \text{Recall}$$

$$\text{Specificity} = \frac{\text{TN}\times 100}{\text{FP}+\text{TN}}$$

$$\text{Precision} = \frac{\text{TP}\times 100}{\text{TP}+\text{FP}}$$

$$F\text{-measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2\times\text{TP}\times 100}{2\times\text{TP}+\text{FP}+\text{FN}}$$

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

The accuracy is the proportion of correctly classified examples among total examples. Sensitivity is the proportion of examples that were classified as positive class, among all examples that are truly positive class. It is equivalent to Recall. The specificity is the proportion of true negatives among all examples that are negative class. The precision is the proportion of the examples that are truly positive class among all those which were classified as positive class. The F-measure is a single measure that characterizes recall and precision. MCC is the Matthew's correlation coefficient. The Matthew's correlation coefficient ranges from -1 to 1. A value of MCC = 1 indicates the best prediction, and a

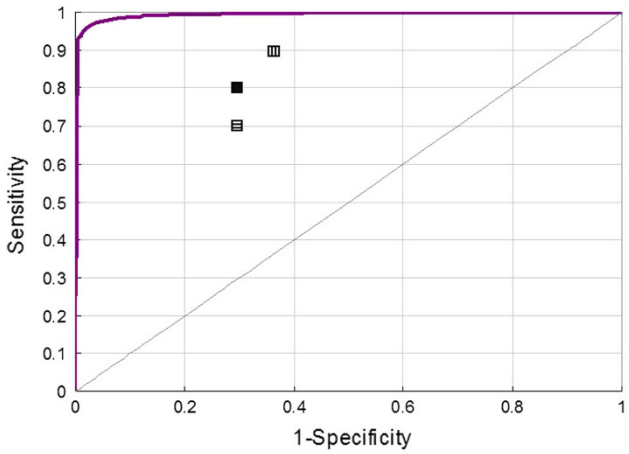

Fig. 4 Comparison of prediction power with similar studies using ROC *plot*

value of MCC = −1 indicates the worst possible prediction. A value of MCC = 0 would be expected for a random prediction scheme. The statistical evaluation of the system through tenfold cross validation shows that the proposed system performs well with average rates of 97.13 % of accuracy, 96.91 % of sensitivity, 97.35 % of specificity, 97.33 % of precision, 97.12 % of F-measure, and 94.26 % of MCC. These measures of discriminative power indicate that the model is not biased.

While traditional decision tree classifiers yield binary class labels, the alternating decision tree classifiers produce confidence measures representing the degree to which class the instance belongs to. Therefore, the result of alternating decision tree can be represented by ROC (receiver operating characteristic) curve as show in Fig. 4. ROC curve (Receiver operating characteristic curve) is a graphical plot of the sensitivity versus (1-specificity) for a binary classifier system as its discrimination threshold varies [48]. The ROC curve depicts relative trade-offs between sensitivity (benefits) and 1-specificity (costs). The ROC curve for a good classifier will be as close as possible to the upper-left corner of the chart.

We compared the performance of our system with other related studies using ROC plot (Fig. 4). The rectangle with horizontal stripe denotes the performance of Deng et al.'s [20] study, the rectangle with vertical stripe shows the performance of Patil et al.'s [31] study, and the black rectangle shows the performance of Lin et al.'s [32] study. Deng et al. used a maximum likelihood estimation method to assess protein interaction reliability using genomic features of gene expression correlation and protein function. Patil et al. proposed a Bayesian network based filtering method for high-throughput protein interaction data using biological features of protein domain, functional similarity, and homology. Lin et al. used Bayesian network-based integrative framework which assigns likelihood scores to each protein pairs based on genomic features of their interlogs. The genomic features were derived from microarray data and gene ontology. Since each of studies uses different sets of training data and attributes, it is hard to compare their performance directly. Especially, the negative training datasets are very different among studies. Because there is no experimentally verified set of non-interacting pairs, negative datasets are usually generated synthetically. Our negative data generation method which utilizes co-localization enrichment information of interacting proteins may help to improve system performance. The result reflects that the relevancy and quality of training dataset determines the reliability and performance of output model. This ROC plot also shows robust performance of the alternating decision tree classifier compared with other similar studies which uses Bayesian network and maximum likelihood estimation.

The ROC curve shows the performance of alternating decision tree classifier. The rectangle with horizontal stripe denotes the performance of Deng et al.'s [20] study, the rectangle with vertical stripe shows the performance of Patilet al.'s [31] study, and the black rectangle shows the performance of Lin et al.'s [32] study.

The generated alternating decision tree shows the classification rules for assessing PPI reliability (Fig. 5). Since the alternating decision tree algorithm dynamically selects relevant attributes and disregards non-informative attributes, the output model can describe about the influence of combinatorial effects of attributes. The generated tree does not include any splitter node using 'absolute protein abundance' or 'co-essentiality' attribute which contains many missing entries. On the other hand, functional similarity measures (which are MIPS similarity and GO similarity) are highly used. And mRNA expression level and co-expression information was also used. In general, lower indices correspond to more influential nodes that were added earlier in the boosting process. From the alternating decision tree classifier, we can observe the following characteristics:

- Rule I: The most discriminative attribute was MIPS functional similarity. If the frequency of shared MIPS functional term between two proteins is less than 4.338, the interacting protein pair must be a positive pair.
- Rule II: If the Pearson's correlation coefficient of mRNA expression data of two interacting proteins is more than 0.326, the interacting protein pair must be a positive pair.
- Rules III, VII, and VIII: Like MIPS functional similarity, the smaller frequency of the shared GO functional term indicates the higher reliability of protein interactions.
- Rules IV and VI: Since marginal essentiality and absolute mRNA expression produce low confidential predictions, they could be regarded as relatively weak evidences for assessing protein–protein interaction reliability.

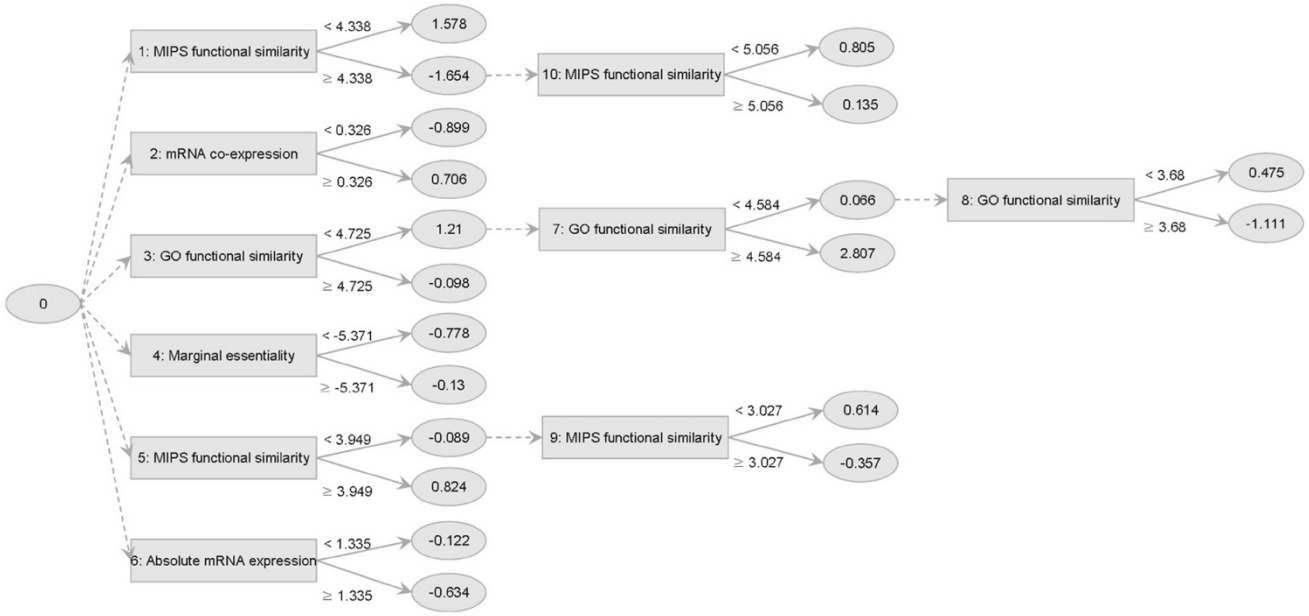

Fig. 5 An alternating decision tree classifier for assessing PPI reliability

Table 2 Reliability assessment of pure high-throughput PPI dataset

	Experimental method	# of detected interaction pairs	% of predicted positives
Ito et al. [13]	Yeast two hybrid assay	4,390	15.52
Garvin et al. [15]	Tandem-affinity purification and mass spectrometry	16,358	31.53

- Rule X: If the frequency of shared MIPS functional term is more than 5.056, the protein pair will be predicted as a negative interaction pair with high confidence. The confidence of prediction is calculated by summing the prediction values [0 + (−1.654) + (+0.135) = −1.519] along the path. The sign of the resultant sum means the target class, and the absolute value represents the confidence of the prediction for the instance.

Since all types of the biological annotations in a protein are not always available, some feature values are frequently unknown. The proposed classification scheme based on alternating decision tree algorithm relieves this problem by considering only the reachable nodes whose associated predictions are large. As you can see from the above rule set, attributes that we used as the domain knowledge mutually make up for each other to cover the lack of information.

We investigated the alternating decision tree classifier with two sets of unlabeled high-throughput experimental datasets: The one was detected by Y2H assay by Ito et al. [13] and the other was obtained by mass spectrometry by Garvin et al. [15]. We filtered out some PPI pairs whose seven attributes are not annotated. The evaluation results are summarized in Table 2. The number of assessed positives in Ito et al.'s Y2H dataset and Garvin et al.'s mass spectrometry dataset is relatively smaller than previous reports [18,19], because our system predicts true interacting protein pair with high confidence. However, the percentile of reliable interacting pairs in Ito's dataset is very similar with Deng et al.'s [20] report which was 15 % sensitivity and specificity. As might have been expected, the percentile of predicted positives of Garvin et al.'s data is larger than that of Ito et al.'s data.

4 Conclusion

In this paper, we presented an assessment scheme for the reliability of candidate interacting proteins in a PPI dataset. This scheme is based on the alternating decision tree algorithm and utilizes the domain knowledge of PPIs. Since the quality of training data determines the reliability and robustness of classifiers, we carefully derived a negative dataset based on co-localization enrichment measure. As a result, we constructed an evaluation system for assessing protein interaction reliability using positive datasets obtained from known protein complex and negative datasets derived based on co-localization characteristics. The experimental results show that applying an alternating decision tree algorithm supplemented with various biological attributes provides excellent performance overall in distinguishing true interacting

protein pairs from a noisy PPI dataset. Moreover, our alternating decision tree classifier is helpful to understand the biological conditions of interacting protein pairs with confidence score. The classifier may also be helpful in predicting new candidate interaction protein pairs.

Studies of biological networks should start with reliable interaction data. A number of reliable PPI datasets assessed by this system can be used as a valuable resource for proteomics research.

References

1. Rivas, J., Fontanillo, C.: Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. PLOS Comput. Biol. **6**(6), e1000807 (2010)
2. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein–protein interaction networks. Nat. Biotechnol. (2003). doi:10.1038/nbt825
3. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics (2003). doi:10.1093/bioinformatics/btg1026
4. Samanta, M.P., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. PNAS (2003). doi:10.1073/pnas.2132527100
5. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. PNAS **100**(21), 12123–12128 (2003). doi:10.1073/pnas.2032324100
6. Deng, M., Tu, Z., Sun, F., Chen, T.: Mapping gene ontology to proteins based on protein–protein interaction data. Bioinformatics (2004). doi:10.1093/bioinformatics/btg500
7. Steffen, M., Petti A., Aach J., D'haeseleer, P., Church, G: Automated modelling of signal transduction networks. BMC Bioinfo. **3**, 34 (2002)
8. Blow, N.: Systems biology: untangling the protein web. Nature **460**, 415–418 (2009)
9. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Mering, C., Jensen, L.: STRING v9.1: protein–protein interaction networks, with increased coverage and integration. Nucl. Acids Res. **41**(D1), D808–D815 (2013)
10. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L.: Reactome: a database of reactions, pathways and biological processes. Nucl. Acids Res. **39**(suppl1), D691–D697 (2011)
11. Wu, G., Feng, X., Stein, L.: A human functional protein interaction network and its application to cancer data analysis. Gen Biol. **11**(5), R53 (2010)
12. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleish, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.: A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature (2000). doi:10.1038/35001009
13. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakai, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. PNAS (2001). doi:10.1073/pnas.061034498
14. Tong A.H., Drees B, Nardelli G, Bader G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C., Fields, S., Boone, C., Cesareni, C.: A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science (2002). doi:10.1126/science.1064987
15. Gavin, A.C., Bosche, M., Krause, R., et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature (2002). doi:10.1038/415141a
16. Ho, Y., Gruhler, A., Heilbut, A., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature (2002). doi:10.1038/415180a
17. von Mering, C., Krause, R., Snel, B., Cornell, M., et al.: Comparative assessment of large-scale data sets of protein–protein interactions. Nature (2002). doi:10.1038/nature750
18. Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D.: Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell Proteomics **1**(5), 349–356 (2002)
19. Sprinzak, E., Sattath, S., Margalit, H. J.: How reliable are experimental protein–protein interaction data? J. Mol. Biol. **327**(5), 919–923
20. Deng, M., Sun, F., Chen, T.: Assessment of the reliability of protein–protein interactions and protein function prediction. Pac. Symp. Biocomput. 140–151 (2003)
21. Legrain, P., Wojcik, J., Gauthier, J.M.: Protein–protein interaction maps: a lead towards cellular functions. Trends Gen. (2001). doi:10.1016/S0168-9525(01)02323-X
22. Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M., Matthews, J.M.: Protein interactions: is seeing believing? Trends Biochem. Sci. (2007). doi:10.1016/j.tibs.2007.09.006
23. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., et al.: Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or "Interologs". Gen. Res. **11**(21):2120–2126 (2001)
24. Chatr-Aryamontri, A., Ceol, A., Licata, L., Cesareni, G.: Protein interactions: integration leads to belief. Trends Biochem. Sci. (2008)
25. Liu, Y., Liu, N., Zhao, H.: Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics (2005). doi:10.1093/bioinformatics/bti492
26. Ben-Hur, A., Noble, W.S.: Kernel methods for predicting protein–protein interactions. Bioinformatics (2005). doi:10.1093/bioinformatics/bti1016
27. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceeding of the sixteenth international conference on data mining, pp. 124–133. (1999)
28. Sato, T., Yamanishi, Y., Kanehisa, M., Toh, H.: The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics (2005). doi:10.1093/bioinformatics/bti564
29. Saito, R., Suzuki, H., Hayashizaki, Y.: Construction of reliable protein–protein interaction networks with a new interaction generality measure. Bioinformatics (2003). doi:10.1093/bioinformatics/btg070
30. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nat. Biotech. (2004). doi:10.1038/nbt924
31. Patil, A., Nakamura, H.: Filtering high-throughput protein–protein interaction data using a combination of genomic features. BMC Bioinfo. (2005). doi:10.1186/1471-2105-6-100

32. Lin, X., Liu, M., Chen, X.: Assessing reliability of protein–protein interactions by integrative analysis of data in model organisms. BMC Bioinfo. (2009). doi:10.1186/1471-2105-10-S4-S5
33. Lee, M.S., Park, S.S.: Comparative analysis of classification methods for protein interaction verification system. Lecture Notes in Computer Science, vol. 4243, Advances in Information Systems, pp. 227–236 (2006)
34. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., et al.: A bayesian networks approach for predicting protein–protein interactions from genomic data. Science (2003). doi:10.1126/science.1087361
35. Jasen, R., Greenbaum, D., Gerstein, M.: Relating whole-genome expression data with protein–protein interactions. Gen. Res. (2002). doi:10.1101/gr.205602
36. Mewes, H.W., Ruepp, A., Theis, F., Rattei, T., Walter, M., Frishman, D., Suhre, K., Spannagl, M., Mayer, K. F. X., Stümpflen, V., Antonov, A.: MIPS: curated databases and comprehensive secondary data resources in 2010. Nucl. Acids Res. (2011)
37. Guldender, U., Munsterkotter, M., Oesterheld, M., Pagel, P., et al.: MPact: the MIPS protein interaction resource on yeast. Nucl. Acids Res. (2006). doi:10.1093/nar/gkj003
38. Huh, W.K., Falvo, J.V., Gerke, L.C., et al.: Global analysis of protein localization in budding yeast. Nature (2003). doi:10.1038/nature02026
39. Lu, L.J., Xia, Y., Yu, H., Rives, A., et al.: Protein interaction prediction by integrating genomic features and protein interaction network analysis. In: Azuaje, F., Dopazo, J. (eds.) Data Analysis and Visualization in Genomics and Proteomics, pp. 61–81. John Wiley & Sons (2005)
40. Ge, H., Liu, Z., Church, G.M., Vidal, M.: Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet. (2001). doi:10.1038/ng776
41. Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., et al.: Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell **9**(5), 1133–1143 (2002)
42. Greenbaum, D., Jansen, R., Gerstein, M.: Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics **18**(4):585–596 (2002)
43. Tetko, I.V., Rodchenkov, I.V., Walter, M.C., Rattei, T., Mewes, H.W.: Beyond the "Best" Match: machine learning annotation of protein sequences by integration of different sources of information. Bioinformatics **24**(5), 621–628 (2008)
44. Ashburner, M., et al.: The gene ontology consortium. Nat. Gen. **25**, 25–29 (2000)
45. Lord, P.W., Stevens, R.D., Goble, C.A.: Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics (2003). doi:10.1093/bioinformatics/btg153
46. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., Gerstein, M.: Genomic analysis of essentiality within protein networks. Trends Gen. (2004). doi:10.1016/j.tig.2004.04.008
47. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. (2005). doi:10.1007/s10994-005-0466-3
48. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. HP Laboratories Technical report HPL-2003-4, Palo Alto (2004)