Molecular Genetics and Genomics (2021) 296:905–918
https://doi.org/10.1007/s00438-021-01789-8

ORIGINAL ARTICLE

Predicting gene phenotype by multi-label multi-class model based on essential functional features

Lei Chen[1,2] · Zhandong Li[3] · Tao Zeng[4] · Yu-Hang Zhang[5] · Hao Li[3] · Tao Huang[6] · Yu-Dong Cai[1]

Received: 26 January 2021 / Accepted: 13 April 2021 / Published online: 29 April 2021

Abstract

Phenotype is one of the most significant concepts in genetics, which is used to describe all the characteristics of a research object that can be observed. Considering that phenotype reflects the integrated features of genotype and environment factors, it is hard to define phenotype characteristics, even difficult to predict unknown phenotypes. Restricted by current biological techniques, it is still quite expensive and time-consuming to obtain sufficient structural information of large-scale phenotype-associated genes/proteins. Various bioinformatics methods have been presented to solve such problem, and researchers have confirmed the efficacy and prediction accuracy of functional network-based prediction. But general functional descriptions have highly complicated inner structures for phenotype prediction. To further address this issue and improve the efficacy of phenotype prediction on more than ten kinds of phenotypes, we first extract functional enrichment features from GO and KEGG, and then use node2vec to learn functional embedding features of genes from a gene–gene network. All these features are analyzed by some feature selection methods (Boruta, minimum redundancy maximum relevance) to generate a feature list. Such list is fed into the incremental feature selection, incorporating some multi-label classifiers built by RAkEL and some classic base classifiers, to build an optimum multi-label multi-class classification model for phenotype prediction. According to recent researches, our method has indeed identified many literature-supported genes/proteins and their associated phenotypes, and even some candidate genes with re-assigned new phenotypes, which provide a new computational tool for the accurate and effective phenotypic prediction.

Keywords Phenotype · Multi-label classification · Network embedding · Functional enrichment · RAkEL · Feature selection

Lei Chen and Zhandong Li contributed equally to this work.

✉ Tao Huang
tohuangtao@126.com

✉ Yu-Dong Cai
cai_yud@126.com

Lei Chen
chen_lei1@163.com

Zhandong Li
lizd591@jlenu.edu.cn

Tao Zeng
zengtao@sibs.ac.cn

Yu-Hang Zhang
zhangyh825@163.com

Hao Li
lihao@jlenu.edu.cn

[1] School of Life Sciences, Shanghai University, Shanghai 200444, People's Republic of China

[2] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China

[3] College of Food Engineering, Jilin Engineering Normal University, Changchun 130052, People's Republic of China

[4] CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

[5] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[6] Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China

Introduction

Phenotype is one of the most significant concepts in genetics (Studies et al. 2007; Lopes et al. 2013). Generally, phenotype is used to describe all the characteristics of a research object that can be observed (Studies et al. 2007). Covering both macroscopical and microcosmic structures, phenotype is not only defined by whether such characteristics can be "seen" by the researchers, but also includes all the biochemical and physiological features (Wojczynski and Tiwari 2008). Corresponding to phenotype, genotype reflects the detailed inner genetic characteristics of an organisms, which is usually represented by the sequence and modification of DNA (Glatt et al. 2007). Generally, phenotype is affected by both genotype and environment, making it a more complicated biological concept (Glatt et al. 2007; Wojczynski and Tiwari 2008).

With the development of next generation sequencing techniques (Davey et al. 2011; Sommer et al. 2013), the genotype of a single organism can be easily detected, sequentially monitored and even predicted according to genetic rules. However, as for phenotype, considering that phenotype reflects the integrated features of genotype and environment factors (Lopes et al. 2013), it is hard to define phenotype characteristics, even difficult to predict unknown phenotypes. For centuries, various bioinformatics methods have been presented, providing a group of potential computational approaches to solve such problem. Generally, all such approaches focused on either the biochemical and biophysical structures (structural features) or the functional network (functional features) of the target protein or large molecule to predict their respective phenotypes (Glatt et al. 2007; Wojczynski and Tiwari 2008; Lopes et al. 2013). For instance, in 2010, researchers have identified the clinical phenotype of various fabry disease associated proteins by their specific mutant structures (Saito et al. 2010). And early in 2007, researchers confirmed the efficacy and prediction accuracy of network-based prediction in *Saccharomyces cerevisiae*, providing a reliable application of functional network-based prediction (McGary et al. 2007).

According to recent publications (Jiang et al. 2016; Zitnik and Leskovec 2017), both structural and functional feature-based phenotype prediction are effective and reliable in phenotype associated studies (McGary et al. 2007; Saito et al. 2010; Sommer et al. 2013). However, restricted by current biological techniques, it is still quite expensive and time-consuming to obtain sufficient structural information of large-scale phenotype-associated genes/proteins. Therefore, up to now, functional feature-based phenotypic prediction will be the most effective and accurate to comprehensively analyze phenotypes. In different functional feature-based studies, the biological functions have different descriptions with different research perspective to phenotype studies. Here, we introduced the most famous groups of gene/protein function descriptors for further analysis: Gene Ontology (GO) (Consortium 2018) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2015) terms, as new candidates of phenotypic functional features. As known, the concept of GO can describe biological functions regardless the diversity of molecular levels and multiplicity of species generically, and KEGG is utilized for multi-omics biological functions and related bioinformatics researches. Therefore, both of these two feature groups can properly supply general functional descriptions for functional prediction on phenotypes. Considering the GO and KEGG terms have highly complicated inner structures for prediction, we also applied node2vec (Grover and Leskovec 2016) to learn new embedding features from a protein–protein network (PPI), which has been described as an effective algorithmic framework (Grover and Leskovec 2016; Yan et al. 2016) to learn useful feature representations from highly structured networks (e.g. PPI) for downstream tasks (e.g. phenotype prediction) (Yang et al. 2019). In addition, we formulate the phenotype prediction as a multi-label classification (Pan et al. 2019) in this work because a gene/protein may be associated with multiple phenotypes.

In brief, we first extracted functional enrichment features from GO and KEGG, and learned functional embedding features of genes from a gene–gene network by node2vec. Then these fused feature representations were fed into a multi-step feature selection to determine optimal features, which were further fed into a multi-label multi-class classification model for final phenotype prediction. According to recent studies, our method has indeed identified many literature-supported genes/proteins and their associated phenotypes, which provides a new computational tool for the accurate and effective phenotypic prediction.

Materials and methods

Datasets

We employed the proteins of budding yeast *Saccharomyces cerevisiae* model organism used in one previous study (Chen et al. 2016), which were retrieved from CYGD (ftp://ftpmips.gsf.de/yeast/) (Güldener et al. 2005). The original data contained some proteins without sequences and phenotypic annotations, after excluding which, 1462 proteins were accessed and investigated in this study. These proteins are assigned one or more following types of phenotypic annotations: (I) conditional phenotypes; (II) cell cycle defects; (III) mating and sporulation defects; (IV) auxotrophies, carbon,

and nitrogen utilization defects; (V) cell morphology and organelle mutants; (VI) stress response defects; (VII) carbohydrate and lipid biosynthesis; (VIII) nucleic acid metabolism defects; (IX) sensitivity to amino acid analogs and other drugs; (X) Sensitivity to antibiotics; (XI) sensitivity to immunosuppressants. The distribution of 1462 proteins on 11 types can be found in the previous study (Chen et al. 2016), where 853 proteins were assigned exact one type of phenotypic annotation, 374 were labeled exact two types, and the rest proteins had more than two types of phenotypic annotation. Accordingly, the problem for predicting protein phenotypic annotation is a multi-label multi-class classification problem.

Feature representation

GO term and KEGG pathway are two widely used materials in bioinformatics. For each gene/protein, its relationship to GO terms and KEGG pathways can be encoded into a vector for representing the protein. Here, we used the enrichment scores (Carmona-Saez et al. 2007) to indicate such relationship. Such way to encode proteins/genes is quite popular (Li et al. 2013, 2019; Chen et al. 2017b, 2019). Compared with the one-hot way to encode proteins/genes, which is quite sensitive to the relationship to some GO terms or KEGG pathways, the enrichment scores are much more robust because they were always continuous numbers. In addition, we also abstracted the relationship to other proteins for a given protein to represent the protein via a network embedding algorithm.

GO and KEGG enrichment features

Given a protein p, let G_p be a set consisting of it and its interacting proteins in STRING. Its GO enrichment score to one GO or KEGG term was computed in the following way.

GO enrichment score The GO enrichment score of p on a GO term GO_j was defined as the $-\log 10$ of the hypergeometric test P value on G_p and the set G_{GO} containing proteins annotated by GO_j. Its calculation formula is as follows:

$$\mathrm{GES}_j = -\log_{10}\left(\sum_{k=m}^{n}\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right), \quad (1)$$

where N was the total number of proteins in yeast, M was the number of proteins in G_{GO}, n was the number of proteins in G_p and m was the number of proteins both in G_p and G_{GO}. 5523 GO terms yielded 5523 GO enrichment scores for each protein.

KEGG enrichment score The KEGG enrichment score of p on a KEGG pathway P_j was defined in a similar way. In detail, it was the $-\log 10$ of the hypergeometric test P value on G_p and the set G_{pathway} containing proteins annotated by P_j, which was computed by

$$\mathrm{PES}_j = -\log_{10}\left(\sum_{k=m}^{n}\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right), \quad (2)$$

where N and n were same as those in Eq. 1, M stood for the number of proteins in G_{pathway}, m stood for the number of proteins both in G_p and G_{pathway}. 106 KEGG pathways produced 106 KEGG enrichment scores for each protein.

The GO and KEGG enrichment scores were termed as functional enrichment features.

Embedding features learned from a protein–protein interaction network

In recent years, some network embedding algorithms have been applied to tackle various biological problems (Luo et al. 2017; Zhao et al. 2019; Che et al. 2020; Zhou et al. 2020a; Zhu et al. 2021). These algorithms can overview a node in a system level and abstract its locations into various numbers. Here, one powerful network embedding algorithm, Node2vec (Grover and Leskovec 2016), was employed to encode each investigated protein.

To apply such network embedding algorithm, a protein network was necessary. This study used the protein–protein interaction (PPI) information reported in STRING (https://string-db.org/, version 10) (von Mering et al. 2003) to construct the protein network. We downloaded the file '4932.protein.links.v10.0.txt.gz', which contained all PPI information for yeast. The constructed network defined 6418 yeast proteins as nodes and two proteins were adjacent if and only if they can interact with each other. The number of edges in such network was 939,998. For convenience, the constructed protein network was denoted as N_p.

The node2vec (Grover and Leskovec 2016) was applied on N_p to obtain the feature vector of each node in N_p. It extends the Skip-gram architecture (Mikolov et al. 2013) of word2vec to the network version by employing the random walk algorithm on a network. For each node, it generates some sequences of nodes in terms of the random walk algorithm. Each sequence of nodes is termed as a sentence and each node is a word. After that, a feature vector is produced based on word2vec. For the detailed description of node2vec, please refer to (Grover and Leskovec 2016). In this study, the node2vec program was downloaded from https://snap.stanford.edu/node2vec/. Default parameters were adopted. Furthermore, the dimension of the output vector

was set to 500. For convenience, these features were called functional embedding features.

As a result, each protein was represented by a vector with collecting functional enrichment and embedding features. Totally, 6129 (=5523 + 106 + 500) features constituted the vector for each gene/protein.

Boruta feature filtering

Boruta feature filtering is able to select all relevant features to the output labels fast. Boruta is based on the random forest (RF) classifier. Boruta consists of the following steps: (1) create copies of original data and shuffle the feature values (called shadow features) of the copies data, and the original and shuffled data are combined to train a RF, which measures the feature importance; (2) for each feature, the *Z* score is calculated, it is standardization of the feature importance score from the RF; (3) select the maximum *Z* score from the shadow features as MZSF; (4) tag the original features whose *Z* score is greater than MZSF as important, and tag the feature whose *Z* score is smaller than MZSF as unimportant; (5) repeat the above processes until all features are tagged.

In this study, the Boruta program retrieved from https://github.com/scikit-learn-contrib/boruta_py is adopted. Default parameters are used for convenience.

mRMR feature selection

The minimum redundancy maximum relevance (mRMR) method (Peng et al. 2005) is a mutual information (MI)-based method for evaluating the importance of each feature. This procedure is implemented by calculating the MI values between features and output labels, and also between features themselves. To indicate the importance of each feature, a feature list is produced by the mRMR method, in which important features have high ranks. This study uses the mRMR program downloaded from http://penglab.janelia.org/proj/mRMR/. Also, default parameters are adopted.

Incremental feature selection (IFS)

IFS is a feature selection with an integrated supervised classifier (Liu and Setiono 1998). Based on the ranked features from mRMR, a series of feature subsets are constructed with a step interval as 1. For instance, the first feature subset has the top 1 feature, and the second feature subset has the top 2 features, and so on. For each feature subset, a classifier is trained on the samples consisting of the features from this feature subset, and the performance is evaluated using tenfold cross-validation (Kohavi 1995). After evaluating on all the generated feature subsets, the feature subset is selected as optimal feature subset when it achieves the highest performance.

Multi-label multi-class classifier RAkEL

In this study, we formulate the phenotype prediction as a multi-label multi-class classification problem. RAkEL (Tsoumakas et al. 2011) is a multi-label classification framework, which breaks the initial labels into several small subsets and is based on label powerset (LP) framework. Previously, LP considers each combination of labels in the training set as class values for single-label classification and train one base classifier on the new transformed data. However, LP cannot handle the data with a large set of labels and some classes with a few training samples, which is time-intensive. RAkEL improves LP by breaking the original labels into several label sets, each label set has a corresponding LP classifier. To date, several multi-label classification models have been set up with this method in tackling different biological problems (Saleema et al. 2012; Weng et al. 2018; Che et al. 2020; Jia et al. 2020a; Zhou et al. 2020a, b; Zhu et al. 2021). In this study, we use the implemented RAkEL in MEKA, which set the parameters $m = 10$, $k = 10$, and three base classifiers are used for multi-class classification respectively. These classifiers have wide applications in bioinformatics (Pan et al. 2010, 2021; Chen et al. 2017a; Jia et al. 2020b; Liang et al. 2020; Liu et al. 2021; Zhang et al. 2021a, b).

IBk

IBk is a *K*-nearest neighbors classifier, which automatically selects the *K* value based on cross-validation. IBk only uses specific instances with a low storage requirement, and its main output is a concept description that consists of multiple stored instances and the past performance during the training process. IBk has three main components: (1) similarity function, which calculates the similarity between a training instance *s* and instances in the concept description; (2) classification function, which is sued to classify the instance *s* and the instances in the concept description; (3) concept description updater, which updates the classification performance in concept description and decides which instance should be kept in the concept description.

RF

RF is a meta classifier consisting of multiple decision trees, and each tree is grown from a bootstrap sample set with a feature subset randomly selected from original features. RF has been widely used in analyzing biological data and demonstrate impressive performance in many studies and applications.

Support vector machine (SVM)

SVM tries to find a hyperplane with the maximum margin between two classes; it can handle both linear and non-linear data. Especially for non-linear data, it uses kernel trick to map the original nonlinear data in a low-dimensional space to a new linear data in a high-dimensional space. SVM needs find those support vectors on the margin between two classes, and these vectors are further used for classifying new samples.

SMOTE

In this work, the analyzed data were imbalance. Thus, the SMOTE (Chawla et al. 2002) is applied to produce new samples for the minor class iteratively until the sample number of the minor class is equivalent to that of the major class, so that, the new balanced data can help promote the construction efficiency of the classification models. We adopt the tool "SMOTE" from Weka in this work.

Performance metrics

In this study, we train multi-label multi-class classifier to predict the phenotypes of genes. Thus, each gene will be predicted to have multiple phenotypes. We mainly use two metrics to measure the prediction performance. One is the exact match, in which the predicted labels must exactly be the same as the true labels. The other is accuracy, which is calculated based on the joint and union set of true and predicted labels as follows:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i \cap y_i^*\right|}{\left|y_i \cup y_i^*\right|}, \tag{3}$$

where y_i is the true label set for sample i, y_i^* is its predicted label set, and N is the total number of samples. Evidently, the higher the exact match/accuracy is, the higher the performance of the classifier is.

In addition, another measurement, hamming loss, is also employed, which can be computed by

$$\text{Hamming loss} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i \Delta y_i^*\right|}{m}, \tag{4}$$

where m is the number of labels ($m = 11$ in this study) and Δ represents the symmetric difference operation of sets.

Results

In this study, the essential features are extracted from GO terms, KEGG pathways and PPI network for each gene. Several advanced computational techniques are adopted to build the multi-label multi-class classification model. The whole pipeline of our analytic method is shown in Fig. 1.

Results of Boruta and mRMR methods

We first extract the enrichment features for each gene from GO and KEGG, and use node2vec to learn the embedding features for each gene (coding proteins). We combine the two sources of features as the final extracted features. Before adopting feature selection methods to analyze features, we construct a new dataset, where each sample has only one label. For example, if a sample has two labels, it will be deemed as two samples with different labels in the new dataset. Such new dataset is fed into Boruta feature selection to extract important features, resulting in 299 features, which are given in Supplementary Material S1. Among these 299 features, embedding features are most, followed by GO enrichment features and KEGG enrichment features (see Fig. 2a). Thus, the embedding features are most relevant to the identification of gene phenotype. Finally, the selected relevant features are further fed into mRMR method to rank. The ranked feature list is also given in Supplementary Material S1. The rank distribution of three feature types is illustrated in Fig. 2b. Evidently, embedding features occupy

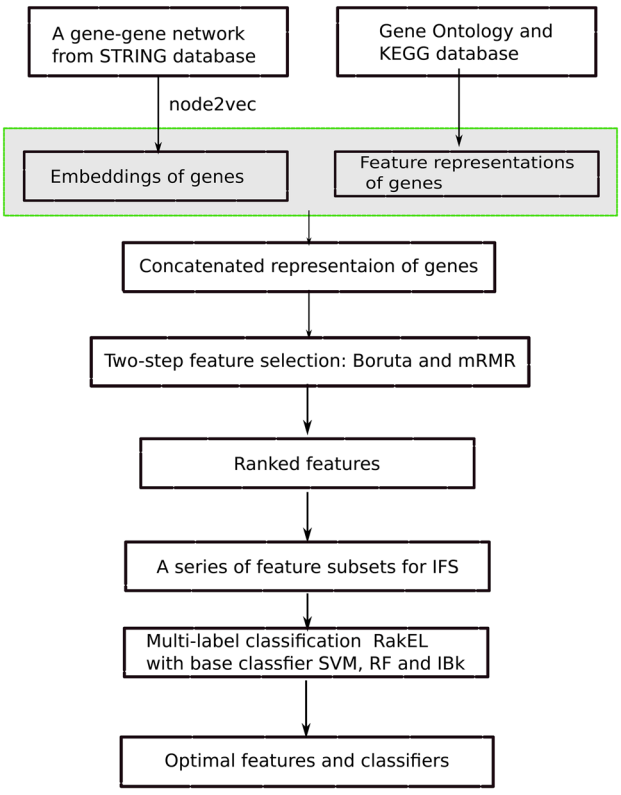

Fig. 1 Flowchart of the proposed multi-label multi-class classification models for predicting gene phenotypes

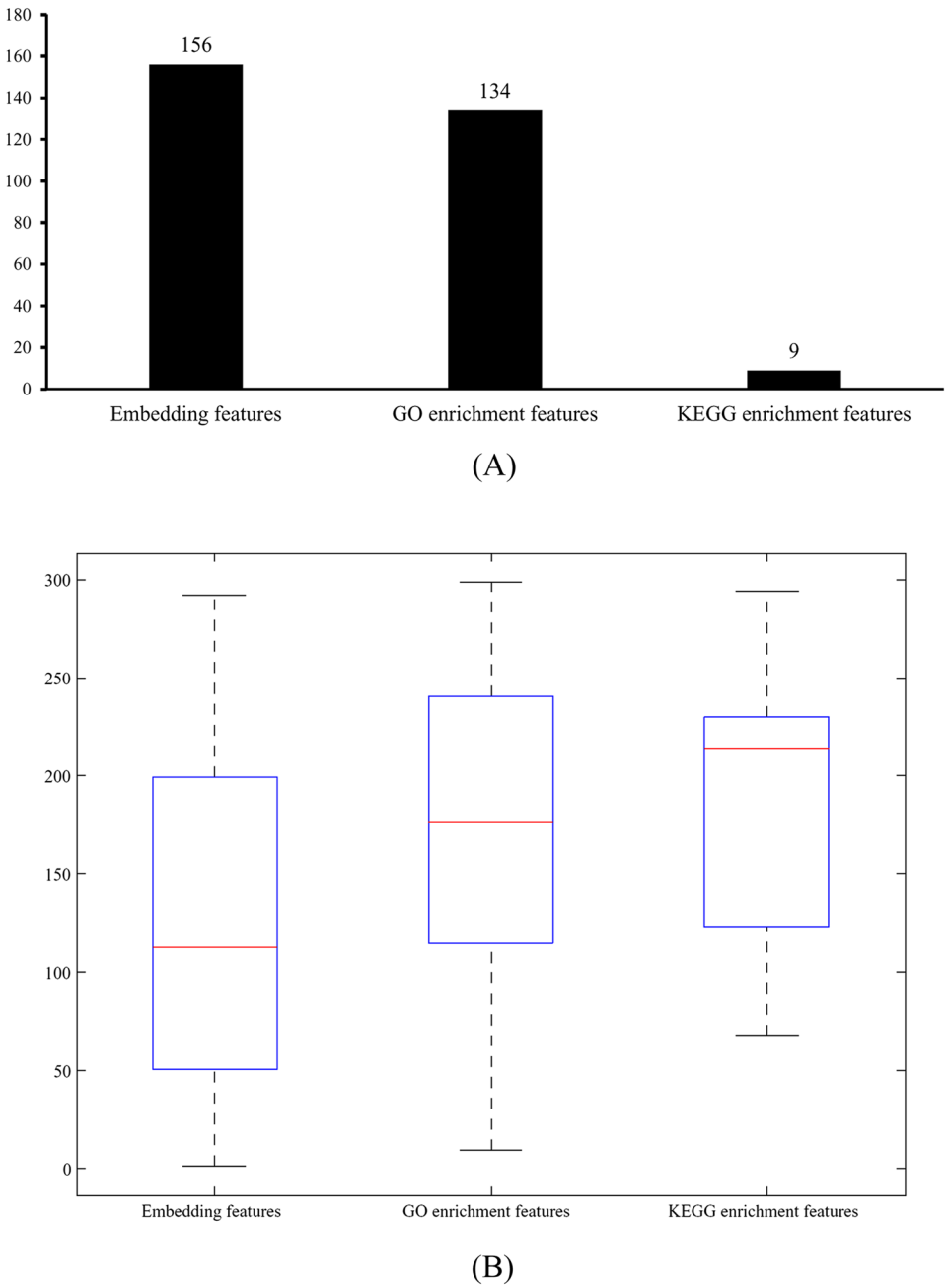

Fig. 2 Analysis of the features selected by Boruta and evaluated by mRMR method on three feature types. **a** Number of selected features on three feature types; **b** Rank distribution of three feature types

most high rank features, further confirming the importance of embedding features.

Results of the IFS method

Based on the feature list obtained in section "Results of Boruta and mRMR methods", we run IFS with RAkEL using three base classifiers (IBk, RF and SVM), to detect optimal features for distinguishing gene phenotypes. A series of feature subsets for IFS are generated. For each feature subset, we train and evaluate the RAkEL on the samples consisting of features from such subset. The evaluation results are counted as accuracy, exact match and hamming loss, as listed in Supplementary Material S2. Accuracy and exact match are selected as the key measurements to assess the performance of each classifier. Accordingly, two curves are plotted for each of base classifier, where one is for accuracy and the other one is for exact match, as shown in Figs. 3 and 4, respectively.

From Fig. 3, we can see that the highest accuracies for RF, IBk and SVM are 0.5290, 0.5195 and 0.4047, respectively. These values are obtained using top 217, 184 and 299 features. The hamming loss values of these classifiers are listed in Table 1. Clearly, RAkEL using RF as the base

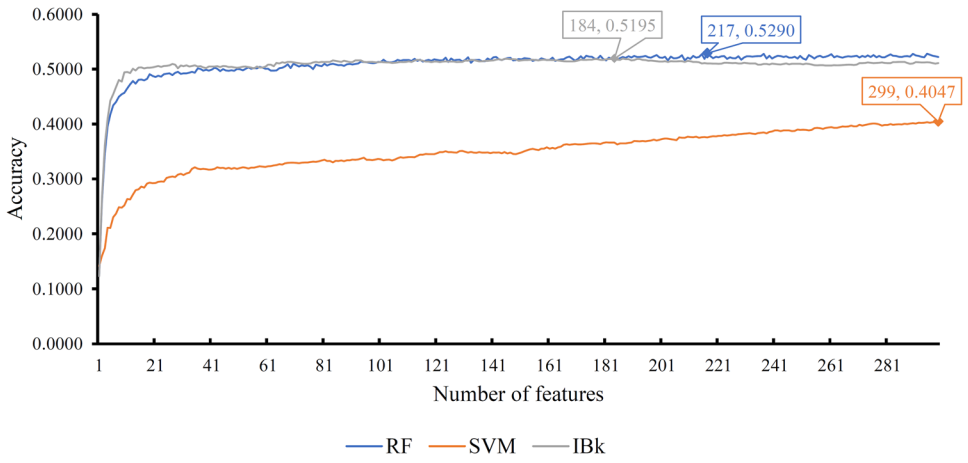

Fig. 3 Accuracy of RAkEL with three base classifiers (RF, SVM and IBk) using different number of features. The RAkEL with RF and top 217 features yields the highest accuracy of 0.5290

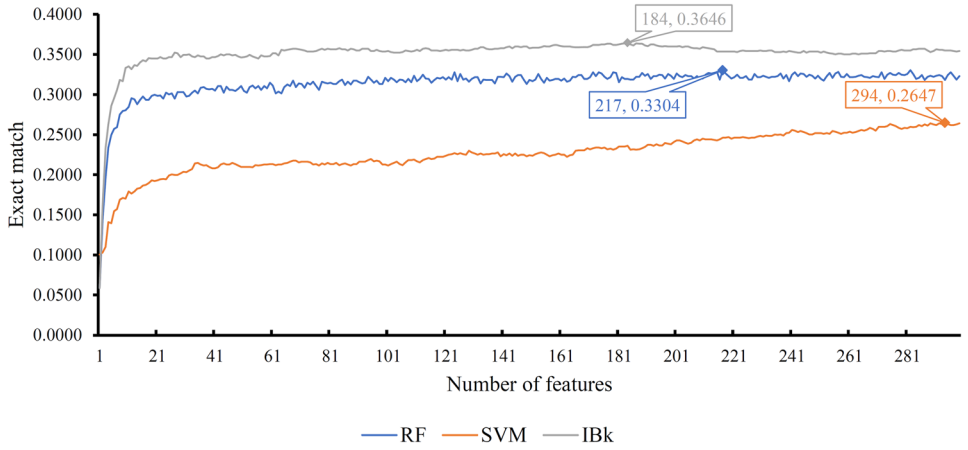

Fig. 4 Exact match of RAkEL with three base classifiers (RF, SVM and IBk) using different number of features. The RAkEL with IBk and top 184 features yields the highest exact match of 0.3646

Table 1 Accuracy and hamming loss of RAkEL with different base classifiers

Classifier	Number of optimal features	Accuracy	Hamming loss
IBk	184	0.5195	0.1077
RF	217	0.5290	0.1214
SVM	299	0.4047	0.1415

classifier and top 217 features yields the highest accuracy. Such classifier is deemed as the optimum classifier based on accuracy. For the 217 features used in this classifier, 123 are embedding features, 89 are GO enrichment features and five are KEGG enrichment features, as shown in Fig. 5a. Embedding features still occupy most, followed by the GO enrichment features and KEGG enrichment features. Furthermore, the rank distribution of three feature types among these 217 features is also investigated, as shown in Fig. 5b. Similar to the features selected by Boruta and evaluated by mRMR method, embedding features are more important than other two feature types.

From Fig. 4, it can be observed that the highest exact match values for three base classifiers are 0.3304, 0.3646 and 0.2647, respectively. They are obtained using top 217, 184 and 294 features. The corresponding hamming loss values are listed in Table 2. Clearly, the RAkEL using IBk as the base classifier and top 184 features produces the highest exact match. Thus, such classifier is deemed as the optimum classifier based on exact match. Among the features used in this classifier, 110 features are embedding features, 71 are GO enrichment features and three are KEGG enrichment features, as shown in Fig. 6a. Likewise, embedding features are still most. Moreover, we investigated the rank distribution of three feature types, as shown in Fig. 6b. Again, the ranks of embedding features are highest.

With the above arguments, we can build two optimum classifiers. One uses the RF as the base classifier and the other one adopts IBk as the base classifier. To indicate the robustness of such two classifiers, we further evaluate their performance with tenfold cross-validation 100 times.

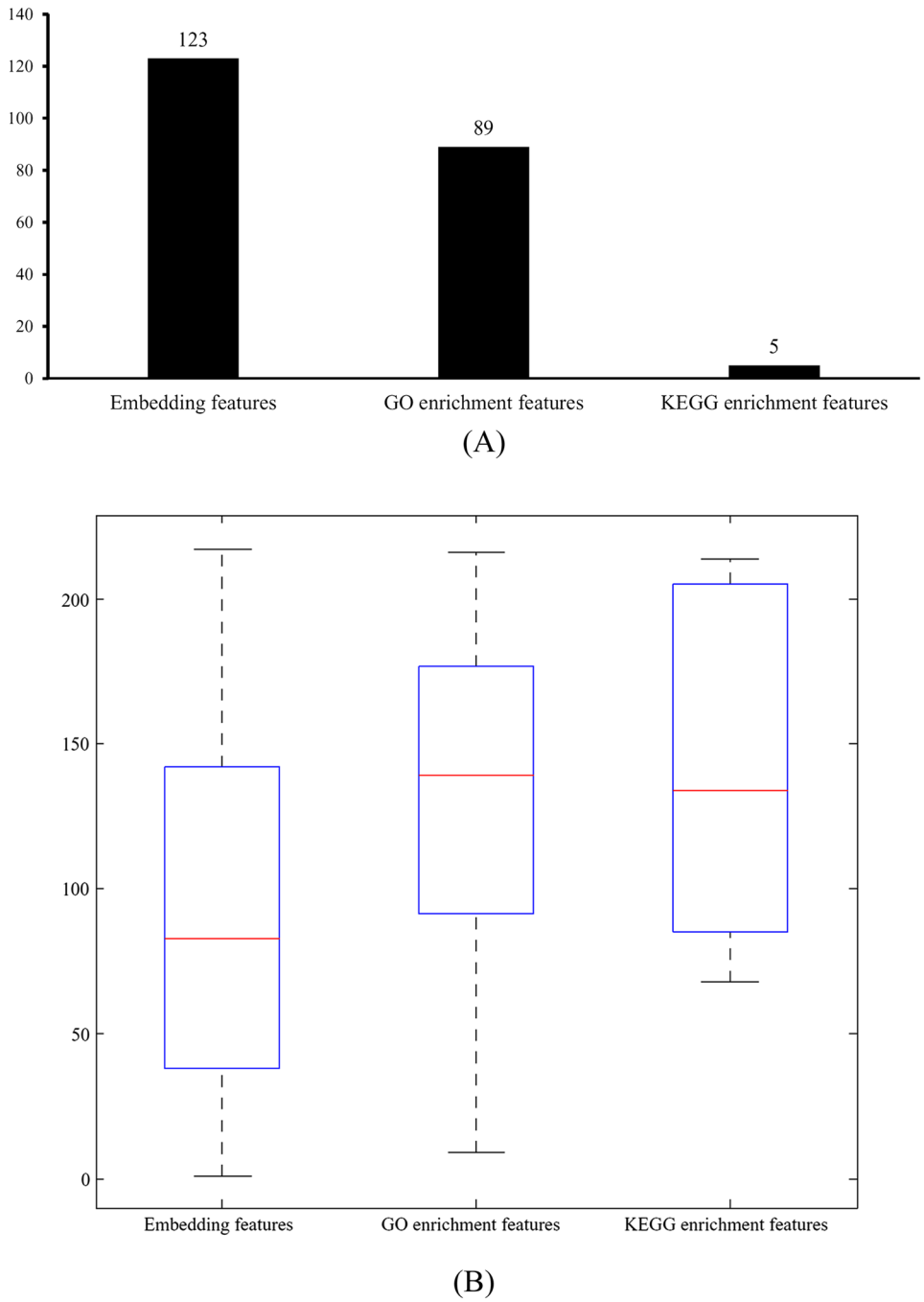

Fig. 5 Analysis of the features used in the optimum RAkEL classifier based on accuracy. **a** Number of selected features on three feature types; **b** Rank distribution of three feature types

Table 2 Exact match and hamming loss of RAkEL with different base classifiers

Classifier	Number of optimal features	Exact match	Hamming loss
IBk	184	0.3646	0.1077
RF	217	0.3304	0.1214
SVM	294	0.2647	0.1424

Obtained three measurements: accuracy, exact match and hamming loss, are shown in Figs. 7 and 8, respectively. It can be observed that each measurement varies in a small interval, suggesting that these two classifiers are quite stable.

Potential novel phenotypic annotations of some genes

As mentioned above, two classifiers are proposed for predicting phenotypes of proteins/genes. The tenfold

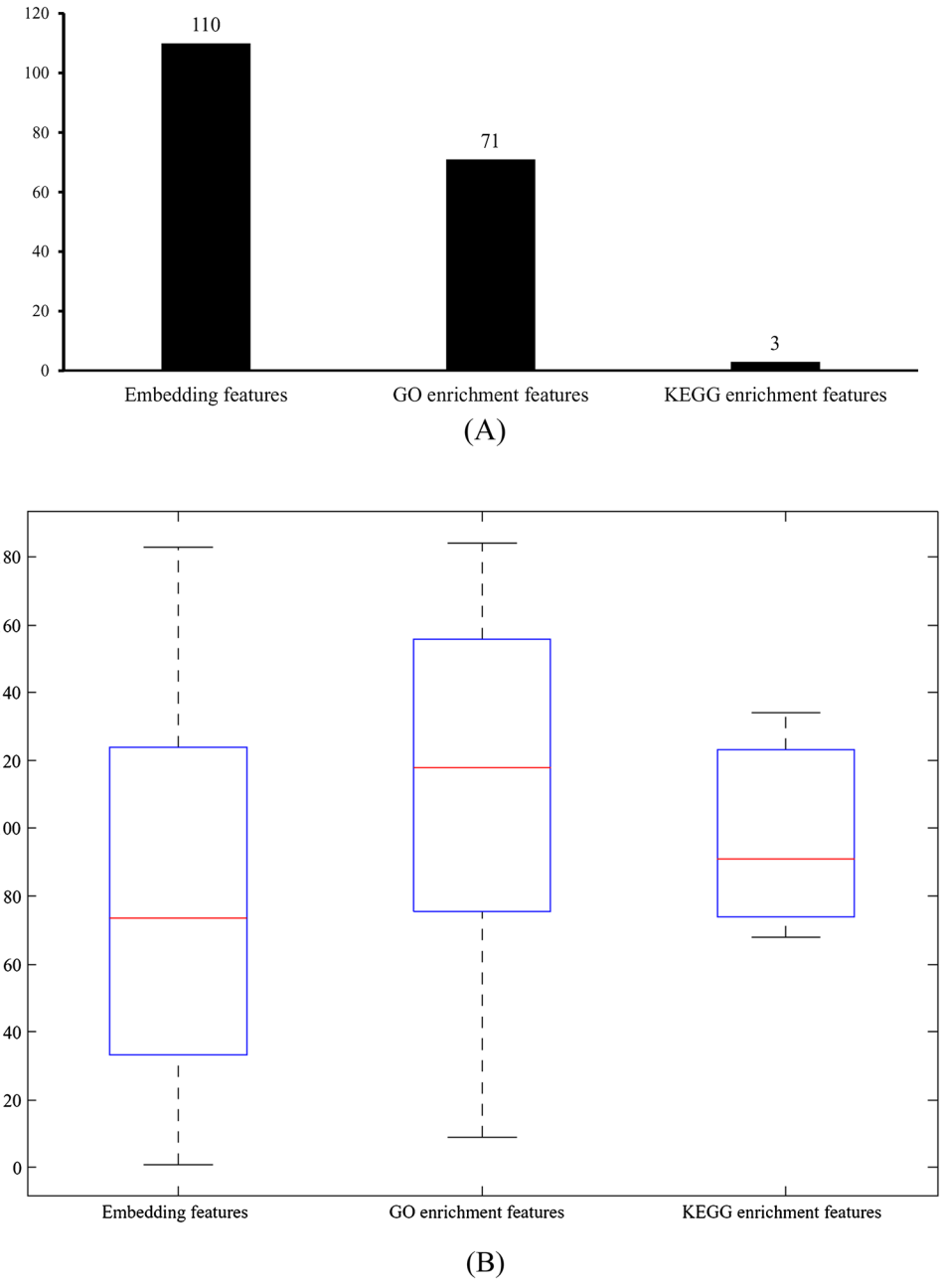

Fig. 6 Analysis of the features used in the optimum RAkEL classifier based on exact match. **a** Number of selected features on three feature types; **b** Rank distribution of three feature types

cross-validation results of each classifier are picked up for detailed analysis.

For the RAkEL using IBk as the base classifier, the cross-validation results indicate that the predicted phenotypes of 1047 genes (71.61%) are all members of their true phenotypes. For the rest 415 genes, the incorrectly predicted phenotype with the maximum likelihood is picked up, which is provided in Supplementary Material S3. In section "IBk-based gene phenotype prediction", some of them will be discussed.

As for the cross-validation results of RAkEL using RF as the base classifier, the predicted phenotypes of 944 genes (64.57%) are all correct. We also picked up the incorrectly predicted phenotype with the maximum likelihood for each of the rest 518 genes, which is also available in Supplementary Material S3. Some of them will be analyzed in section "RF-based gene phenotype prediction".

Discussion

As we have mentioned above, we encode genes with their proper functional annotations (GO, KEGG and PPI). Further using novel machine learning models, we identify the

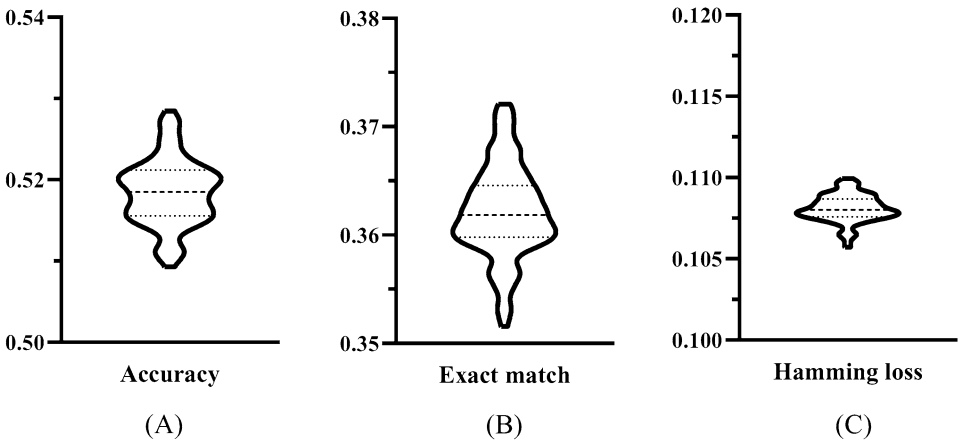

Fig. 7 Violin plot to show the performance of the optimum RAkEL classifier based on exact match under tenfold cross-validation 100 times. **a** Accuracy; **b** Exact match; **c** Hamming loss

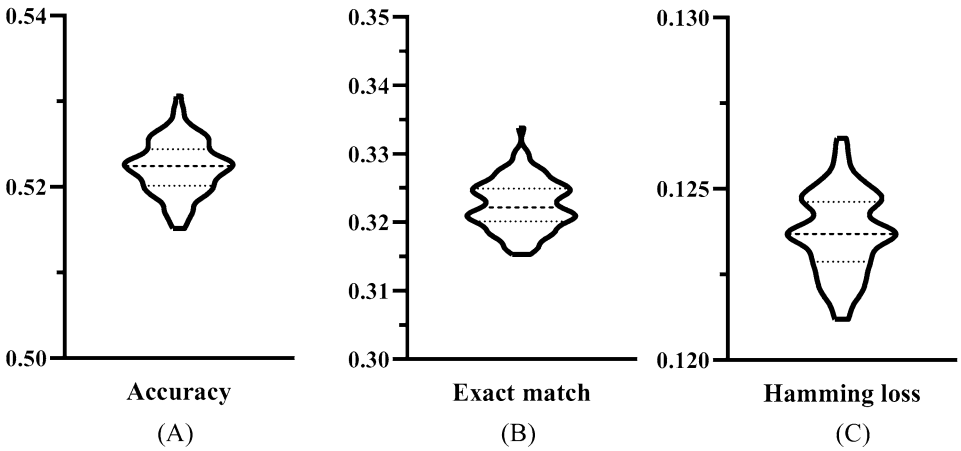

Fig. 8 Violin plot to show the performance of the optimum RAkEL classifier based on accuracy under tenfold cross-validation 100 times. **a** Accuracy; **b** Exact match; **c** Hamming loss

functional clustering patterns of genes. In this study, we use two base classifiers to build the multi-label classifiers: IBk and RF. According to the prediction results, some genes are clustered into the so-called incorrected directed classes, such genes and their cluster re-assignments can be confirmed to be reasonable at the biological function level, which are supported by Saccharomyces Genome Database and recent publications. These findings can help discover novel phenotypes of proteins/genes and can be further confirmed by solid experiments.

IBk-based gene phenotype prediction

When we screened out specific genes processed by RAkEL using IBk as the base classifier, predicted phenotypic annotations of most genes (71.61%) are absolutely true annotations. As for the remaining miss directed genes, actually, they are not simply clustered into at least one incorrect cluster but be re-assigned to the alternative cluster due to their functional complexity. The most likely incorrect clusters of these genes were picked up. Some of them are listed in Table 3.

The first gene is **YAL010C**, also named as MDM10 and participating in the biological regulation of ERMES and the SAM complex (König 2012). According to the existing datasets, such gene would be clustered in 1 and 5 clusters (conditional phenotypes, cell morphology and organelle mutants). Previous studies has already confirmed the contribution of YAL010C on conditional phenotypes and cell morphology (Sogo and Yaffe 1994). However, our presented computational method, clustered such gene into cluster 4, auxotrophies, carbon, and nitrogen utilization defects. According to recent publications, early in 2003, researchers confirmed MDM10 regulated the amino acid utilization in *Aspergillus nidulans*, another typical fungus (Koch et al. 2003). Therefore, it is quite reasonable to have different functional annotation with new phenotype of YAL010C considering its biological complexity.

The next re-assigned gene is **YAL035W.** Also named as FUN12, such gene has been widely reported to participate as GTPase promoting Met-tRNAiMet binding (Alone et al. 2008; Kim et al. 2018). Initially, such gene would be clustered into class 1, indicating its specific biological functions in conditional phenotypes (Haruki et al. 2008). However,

Table 3 Latent novel phenotypic annotations of some genes identified by RAkEL with IBk

Gene	True phenotypic annotations	Most likely predicted phenotypic annotations
YAL010C	Conditional phenotypes; cell morphology and organelle mutants	Auxotrophies, carbon, and nitrogen utilization defects
YAL035W	Conditional phenotypes	Sensitivity to amino acid analogs and other drugs
YAL047C	Conditional phenotypes; cell morphology and organelle mutants	Cell cycle defects
YAL054C	Cell morphology and organelle mutants	Sensitivity to antibiotics
YAL058W	Cell morphology and organelle mutants	Sensitivity to amino acid analogs and other drugs

with our newly present method, such gene has been clustered into class 9 (sensitivity to amino acid analogs and other drugs). According to recent publications, early in 2006, researchers have already identified such gene as an eukaryotic ribosomal complexes associated protein interacting with certain exogenous amino acid analogs and were shown to be associated with related drug sensitivity (Fleischer et al. 2006), corresponding with our prediction. Therefore, the prediction of YAL035W in re-assigned functional clusters may be caused by multi-functional capacity of such gene.

As the following gene, **YAL047C** has also been clustered into a different cluster comparing with previous information. In the prediction result, acting like a receptor for gamma-tubulin small complex, such gene has been widely reported to contribute to microtubule formation and stabilization (Luban et al. 2005), which would be initially clustered into cluster 1 and 5 just like YAL010C (Corbacho et al. 2005; Nguyen et al. 2018). By contrast, such gene has been clustered into class 2 which describes cell cycle defects, newly discovered in this work.

For the following gene as **YAL054C**, according to SGD, it has also been known as FUN44 and ACS1 widely reported to participate in histone acetylation-associated biological processes (Yukawa et al. 2009; Li et al. 2010a). Originally, such gene has been confirmed to participate in class 5-associated biological processes (cell morphology and organelle mutants) (White 1999). In our prediction list, YAL054C has been functionally clustered into class 10, describing sensitivity to antibiotics. Early in 2003, a system study (Palsson et al. 2003) on the composition and methods for yeast metabolism confirmed that our candidate gene YAL054C may actually participate in antibiotics associated processes. Apart from this independent study, the direct evidence for the interrelationship between YAL054C and antibiotics biological processes is still remained for further validation at different molecular levels.

What is more, we also observed a specific gene named as **YAL058W**. According to recent publications, such gene has been participating in ER membrane folding and glycoprotein quality control (Li et al. 2010b), which might originally be classified into class 5 (cell morphology and organelle mutants) (Seeley et al. 2002). Meanwhile, according to our new computational analysis, such gene has been classified into class 9 (sensitivity to amino acid analogs and other drugs). According to related publications (Caro et al. 1997; Li et al. 2010b), YAL058W has been widely reported to be actually sensitive to amino acids, validating the efficacy and accuracy of our prediction.

RF-based gene phenotype prediction

Similar to above genes predicted by RAkEL using IBk as the base classifier, we also predicted various genes with accurate functional cluster distribution by RAkEL using RF as the base classifier. Among all the genes, 944 genes (64.57%) were predicted a part of their true functional annotations (clustering results). As for the remaining genes, they may also be re-assigned into different clusters due to the complexity of their biological functions. We also selected the most likely predicted cluster for each of these genes. Some of the top candidate genes of the re-assigned prediction of RF are just the same as those of IBk (like YAL010C, YAL047C, YAL035W and YAL054C), indicating the robust of our prediction based on novel machine learning models. Here, we discussed other two genes, listed in Table 4.

In our optimal prediction list from RF, gene **YAL002W** could initially be clustered into class 1 (conditional phenotypes) (Horazdovsky et al. 1996) and class 5 (cell morphology and organelle mutants) (Zhou et al. 2009). Relied on RF, such gene has been re-clustered into another effective biological group (Class 6), describing stress response defects. Gene YAL002W also named as VPS8 has been

Table 4 Latent novel phenotypic annotations of some genes identified by RAkEL with RF

Gene	True phenotypic annotations	Most likely predicted phenotypic annotations
YAL002W	Conditional phenotypes; cell morphology and organelle mutants	Stress response defects
YAL023C	Cell morphology and organelle mutants	Carbohydrate and lipid biosynthesis

widely reported to participate in membrane-binding processes of the CORVET complex (Peplowska et al. 2007), and such gene has been widely reported to contribute to the regulation of heat stress responses in multiple species including yeast (Huisinga and Pugh 2004; Le Breton and Mayer 2016). Therefore, considering the complicated biological contribution of YAL002W, it's quite reasonable to have such different phenotype cluster assignment in our prediction results.

Another gene named as **YAL023C** also has different cluster assignment under the RF model. Initially, such gene could be clustered to class 5, describing cell morphology and organelle mutants, on the basis of recent publications (Karpova et al. 1998; Mouyna et al. 2010) and SGD annotation. However, by RF prediction, YAL023C has been re-clustered to a new class (carbohydrate and lipid biosynthesis defects). Although the direct relationship between our candidate gene and such phenotype has not been identified, there are some publications (Lussier et al. 1995; Novotná et al. 2004; Villa-García et al. 2011) confirming that YAL023C is associated with basic membrane functions of yeast.

Although both computational methods work well for the prediction of gene phenotypes, in this study, IBk method may work better and may be more suitable for further application in such research field. All in all, both computational methods can group most candidate genes into their respective functional clusters correctly. However, due to the complexity of gene functions, some genes have been re-clustered to another clusters/classes. According to our discussion above, we confirmed that most of such re-assigned genes indeed participate in the newly predicted phenotype at biological functional level, validating the efficacy and accuracy of our function-based gene phenotype prediction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00438-021-01789-8.

Author contributions TH and YDC designed the study, supervised the project and finalized the manuscript. LC, ZL and TZ did the experiments. ZDL, YHZ and HL analyzed the results. LC drafted the manuscript.

Funding This research was funded by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), National Key R&D Program of China (2017YFC1201200), Shanghai Municipal Science and Technology Major Project [2017SHZDZX01], National Key R&D Program of China [2018YFC0910403], National Natural Science Foundation of China [31701151], Shanghai Sailing Program [16YF1413800], the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) [2016245], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

Availability of data and material The proteins and their phenotypic annotations were retrieved from CYGD (ftp://ftpmips.gsf.de/yeast/).

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Ethics approval This article does not contain any studies with human participants performed by any of the authors.

References

Alone PV, Cao C, Dever TE (2008) Translation initiation factor 2gamma mutant alters start codon selection independent of Met-tRNA binding. Mol Cell Biol 28:6877–6888

Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8:R3

Caro LHP, Tettelin H, Vossen JH, Ram AF, Van Den Ende H, Klis FM (1997) In silicio identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of Saccharomyces cerevisiae. Yeast 13:1477–1489

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Che J, Chen L, Guo Z-H, Wang S, Aorigele C (2020) Drug target group prediction with multiple drug networks. Combin Chem High Throughput Screen 23:274–284

Chen L, Zhang YH, Huang T, Cai YD (2016) Identifying novel protein phenotype annotations by hybridizing protein–protein interactions and protein sequence similarities. Mol Genet Genom 291:913–934

Chen L, Wang S, Zhang Y-H, Li J, Xing Z-H, Yang J, Huang T, Cai Y-D (2017a) Identify key sequence features to improve CRISPR sgRNA efficacy. IEEE Access 5:26582–26590

Chen L, Zhang Y-H, Lu G, Huang T, Cai Y-D (2017b) Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. Artif Intell Med 76:27–36

Chen L, Pan X, Zhang Y-H, Liu M, Huang T, Cai Y-D (2019) Classification of widely and rarely expressed genes with recurrent neural network. Comput Struct Biotechnol J 17:49–60

Consortium GO (2018) The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res 47:D330–D338

Corbacho I, Olivero I, Hernández LM (2005) A genome-wide screen for Saccharomyces cerevisiae nonessential genes involved in mannosyl phosphate transfer to mannoprotein-linked oligosaccharides. Fungal Genet Biol 42:773–790

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Fleischer TC, Weaver CM, McAfee KJ, Jennings JL, Link AJ (2006) Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. Genes Dev 20:1294–1307

Glatt SJ, Chayavichitsilp P, Depp C, Schork NJ, Jeste DV (2007) Successful aging: from phenotype to genotype. Biol Psychiat 62:282–293

Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, San Francisco, California, USA, pp 855–864

Güldener U, Münsterkötter M, Kastenmüller G, Strack N, Van Helden J, Lemer C, Richelles J, Wodak S, Garcia-Martinez J, Perez-Ortin J (2005) CYGD: the comprehensive yeast genome database. Nucleic Acids Res 33:D364–D368

Haruki H, Nishikawa J, Laemmli UK (2008) The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. Mol Cell 31:925–932

Horazdovsky BF, Cowles CR, Mustol P, Holmes M, Emr SD (1996) A novel RING finger protein, Vps8p, functionally interacts with the small GTPase, Vps21p, to facilitate soluble vacuolar protein localization. J Biol Chem 271:33607–33615

Huisinga KL, Pugh BF (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13:573–585

Jia Y, Chen L, Zhou J-P, Liu M (2020a) iMPT-FRAKEL: A simple multi-label web-server that only uses fingerprints to identify which metabolic pathway types compounds can participate in. Open Bioinform J 13:83–91

Jia Y, Zhao R, Chen L (2020b) Similarity-based machine learning model for predicting the metabolic pathways of compounds. IEEE Access 8:130687–130696

Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 17:184

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462

Karpova TS, Moltz SL, Riles LE, Guldener U, Hegemann JH, Veronneau S, Bussey H, Cooper JA (1998) Depolarization of the actin cytoskeleton is a specific phenotype in Saccharomyces cerevisiae. J Cell Sci 111:2689–2696

Kim E, Kim JH, Seo K, Hong KY, An SWA, Kwon J, Lee SV, Jang SK (2018) eIF2A, an initiator tRNA carrier refractory to eIF2alpha kinases, functions synergistically with eIF5B. Cell Mol Life Sci 75:4287–4300

Koch KV, Suelmann R, Fischer R (2003) Deletion of mdmB impairs mitochondrial distribution and morphology in *Aspergillus nidulans*. Cell Motil Cytoskelet 55:114–124

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence. Lawrence Erlbaum Associates Ltd, pp 1137–1145

König J (2012) Untersuchungen zur anterograden Bewegung und Vererbung von Mitochondrien in *Saccharomyces cerevisiae*

Le Breton L, Mayer MP (2016) Heat shock response: a model for handling cell stress. Elife 5:e22850

Li L, Ching W, Chan Y, Mamitsuka H (2010a) On network-based kernel methods for protein–protein interactions with applications in protein functions prediction. J Syst Sci Complexity 23:917–930

Li S, Spooner RA, Allen SC, Guise CP, Ladds G, Schnöder T, Schmitt MJ, Lord JM, Roberts LM (2010b) Folding-competent and folding-defective forms of ricin A chain have different fates after retrotranslocation from the endoplasmic reticulum. Mol Biol Cell 21:2543–2554

Li Z, Li BQ, Jiang M, Chen L, Zhang J, Liu L, Huang T (2013) Prediction and analysis of retinoblastoma related genes through gene ontology and KEGG. Biomed Res Int 2013:304029

Li J, Lu L, Zhang Y, Liu M, Chen L, Huang T, Cai Y-D (2019) Identification of synthetic lethality based on a functional network by using machine learning algorithms. J Cell Biochem 120:405–416

Liang H, Chen L, Zhao X, Zhang X (2020) Prediction of drug side effects with a refined negative sample selection strategy. Comput Math Methods Med 2020:1573543

Liu HA, Setiono R (1998) Incremental feature selection. Appl Intell 9:217–230

Liu H, Hu B, Chen L, Lu L (2021) Identifying protein subcellular location with embedding features learned from networks. Curr Proteom

Lopes LR, Rahman MS, Elliott PM (2013) A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic cardiomyopathy caused by sarcomeric protein mutations. Heart 99:1800–1811

Luban C, Beutel M, Stahl U, Schmidt U (2005) Systematic screening of nuclear encoded proteins involved in the splicing metabolism of group II introns in yeast mitochondria. Gene 354:72–79

Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8:573

Lussier M, Gentzsch M, Sdicu A-M, Bussey H, Tanner W (1995) Protein O-glycosylation in yeast THE PMT2 GENE SPECIFIES A SECOND PROTEIN O-MANNOSYLTRANSFERASE THAT FUNCTIONS IN ADDITION TO THE PMT1-ENCODED ACTIVITY. J Biol Chem 270:2770–2775

McGary KL, Lee I, Marcotte EM (2007) Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. Genome Biol 8:R258

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: International conference on learning representations, Scottsdale, Arizona, USA

Mouyna I, Kniemeyer O, Jank T, Loussert C, Mellado E, Aimanianda V, Beauvais A, Wartenberg D, Sarfati J, Bayry J (2010) Members of protein O-mannosyltransferase family in *Aspergillus fumigatus* differentially affect growth, morphogenesis and viability. Mol Microbiol 76:1205–1221

Nguyen TD, Walker ME, Gardner JM, Jiranek V (2018) Appropriate vacuolar acidification in *Saccharomyces cerevisiae* is associated with efficient high sugar fermentation. Food Microbiol 70:262–268

Novotná D, Flegelová H, Janderová B (2004) Different action of killer toxins K1 and K2 on the plasma membrane and the cell wall of *Saccharomyces cerevisiae*. FEMS Yeast Res 4:803–813

Palsson BO, Famili I, Fu P, Nielsen JB, Forster J (2003) Compositions and methods for modeling *Saccharomyces cerevisiae* metabolism. In. Google Patents

Pan XY, Zhang YN, Shen HB (2010) Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. J Proteome Res 9:4992–5001

Pan X, Fan YX, Jia J, Shen HB (2019) Identifying RNA-binding proteins using multi-label deep learning. Sci China Inf Sci 62:019103

Pan X, Li H, Zeng T, Li Z, Chen L, Huang T, Cai Y-D (2021) Identification of protein subcellular localization with network and functional embeddings. Front Genet 11:626500

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238

Peplowska K, Markgraf DF, Ostrowicz CW, Bange G, Ungermann C (2007) The CORVET tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. Dev Cell 12:739–750

Saito S, Ohno K, Sese J, Sugawara K, Sakuraba H (2010) Prediction of the clinical phenotype of Fabry disease based on protein sequential and structural information. J Hum Genet 55:175–178

Saleema JS, Sairam B, Naveen SD, Yuvaraj K, Patnaik LM (2012) Prominent label identification and multi-label classification for cancer prognosis prediction. In: TENCON 2012 IEEE Region 10 conference, pp 1–6

Seeley ES, Kato M, Margolis N, Wickner W, Eitzen G (2002) Genomic analysis of homotypic vacuole fusion. Mol Biol Cell 13:782–794

Sogo LF, Yaffe MP (1994) Regulation of mitochondrial morphology and inheritance by Mdm10p, a protein of the mitochondrial outer membrane. J Cell Biol 126:1361–1373

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. BMC Genom 14:542

Studies N-NWGoRiA, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF Jr, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007) Replicating genotype–phenotype associations. Nature 447:655–660

Tsoumakas G, Katakis I, Vlahavas I (2011) Random k-labelsets for multilabel classification. IEEE Trans Knowl Data Eng 23:1079–1089

Villa-García MJ, Choi MS, Hinz FI, Gaspar ML, Jesch SA, Henry SAJMg, Genomics (2011) Genome-wide screen for inositol auxotrophy in *Saccharomyces cerevisiae* implicates lipid metabolism in stress response signaling. Mol Genet Genom 285:125–149

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31:258–261

Weng H, Liu Z, Maxwell A, Li X, Zhang C, Peng E, Li G, Ou A (2018) Multi-label symptom analysis and modeling of TCM diagnosis of hypertension. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 1922–1929

White A-M (1999) Identification of cell surface assembly mutants in *Saccharomyces cerevisiae*. In. Massachusetts Institute of Technology

Wojczynski MK, Tiwari HK (2008) Definition of phenotype. Adv Genet 60:75–105

Yan JC, Cho MS, Zha HY, Yang XK, Chu SM (2016) Multi-graph matching via affinity optimization with graduated consistency regularization. IEEE Trans Pattern Anal Mach Intell 38:1228–1242

Yang S, Tian J, Zhang H, Yan J, He H, Jin Y (2019) TransMS: knowledge graph embeddding for complex relations by multidirectional semantics. IJCAI

Yukawa M, Yo K, Hasegawa H, Ueno M, Tsuchiya E (2009) The Rpd3/HDAC complex is present at the URS1 cis-element with hyperacetylated histone H3. Biosci Biotechnol Biochem 73:378–384

Zhang Y-H, Li H, Zeng T, Chen L, Li Z, Huang T, Cai Y-D (2021a) Identifying transcriptomic signatures and rules for SARS-CoV-2 infection. Front Cell Dev Biol 8:627302

Zhang Y-H, Zeng T, Chen L, Huang T, Cai Y-D (2021b) Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles. Front Genet 11:599970

Zhao X, Chen L, Guo Z-H, Liu T (2019) Predicting drug side effects with compact integration of heterogeneous networks. Curr Bioinform 14:709–720

Zhou X, Arita A, Ellen TP, Liu X, Bai J, Rooney JP, Kurtz AD, Klein CB, Dai W, Begley TJ (2009) A genome-wide screen in Saccharomyces cerevisiae reveals pathways affected by arsenic toxicity. Genomics 94:294–307

Zhou J-P, Chen L, Guo Z-H (2020a) iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. Bioinformatics 36:1391–1396

Zhou J-P, Chen L, Wang T, Liu M (2020b) iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. Bioinformatics 36:3568–3569

Zhu Y, Hu B, Chen L, Dai Q (2021) iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network. Comput Math Methods Med 2021:6683051

Zitnik M, Leskovec J (2017) Predicting multicellular function through multi-layer tissue networks. Bioinformatics 33:i190–i198