PAPER • **OPEN ACCESS**

Performance evaluation of features for gene essentiality prediction

To cite this article: Olufemi Aromolaran *et al* 2021 *IOP Conf. Ser.: Earth Environ. Sci.* **655** 012019

View the article online for updates and enhancements.

This content was downloaded from IP address 165.73.192.252 on 24/05/2021 at 13:03

Performance evaluation of features for gene essentiality prediction

Olufemi Aromolaran[1,2], Jelili Oyelade[1,2], Ezekiel Adebiyi[1,2]

[1]Department of Computer and Information Science, Covenant University, Ota, Nigeria

[2]Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Nigeria

Corresponding email: olufemi.aromolaran@stu.cu.edu.ng

Abstract. Essential genes are subset of genes required by an organism for growth and sustenance of life and as well responsible for phenotypic changes when their activities are altered. They have been utilized as drug targets, disease control agent, etc. Essential genes have been widely identified especially in microorganisms, due to the extensive experimental studies on some of them such as *Escherichia coli* and *Saccharomyces cerevisiae*. Experimental approach has been a reliable method to identify essential genes. However, it is complex, costly, labour and time intensive. Therefore, computational approach has been developed to complement the experimental approach in order to minimize resources required for essentiality identification experiments. Machine learning approaches have been widely used to predict essential genes in model organisms using different categories of features with varying degrees of accuracy and performance. However, previous studies have not established the most important categories of features that provide the distinguishing power in machine learning essentiality predictions. Therefore, this study evaluates the discriminating strength of major categories of features used in essential gene prediction task as well as the factors responsible for effective computational prediction. Four categories of features were considered and *k*-fold cross-validation machine learning technique was used to build the classification model. Our results show that ontology features with an AUROC score of 0.936 has the most discriminating power to classify essential and non-essential genes. This studyconcludes that more ontology related features will further improve the performance of machine learning approach and also sensitivity, precision and AUPRC are realistic measures of performance in essentiality prediction.

Keywords: Essential genes, Essential proteins, Classification features, Machine-learning

1.0 Introduction

A gene is defined as an essential gene if its total loss of function results in a total loss of fitness of the organism[1]. The knowledge obtained from the discovery of essential genes accelerates

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

the discovery of drug targets [2,3], guides the engineering of new organisms, provides knowledge about the basic requirements for a cell and proffers insights to the correlations between genotype and phenotype. For instance, deleting just one gene that codes for an essential function in an organism is sufficient to cause lethality or infertility[4]. In comparison to non-essential genes, essential genes are expected to be conserved in biological evolution[3,5,6], e.g. genes found in bacteria such as, *dnaB, rpoA, and dcd* etc.[7]. Due to the time consumption and costly nature of experimental analysis, only few microorganisms have been extensively studied, and their essential and non-essential genesets have become models for poorly or under studied organisms. Some of the model organisms include *Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Pseudomonas, and Bacillus subtilis*. In view of the complexities and drawbacks of the *in vitro* approach, computational techniques have been developed to predict gene essentiality [8–10]with the approach gaining huge popularity in recent years [11–13].

From peer reviewed publications, there are three major computational approaches available for gene essentiality prediction, these are homology mapping[14,15], constraint-based[16,17]and machine learning approach[13,18]. A computational prediction is especially useful when the organism is either unculturable, such as *Pneumocystis carinii*, or difficult to perform gene disruption on, such as *Aspergillus fumigatus* [19].

1.1 Computational Approaches for Predicting essential genes

Homology mapping is the earliest computational approach used to determine essential genes [14]. This requires comparison between sequences of two organisms (a model and a target) to determine their similarity based on defined percentage identity threshold (e-value). If a sequence from target organism shows high similarity to a sequence of essential gene from a model organism, then the target sequence is labelled to be essential. This is premised on the biological theory that states that “structure determines function and vice versa”.

The comparative genomic analysis includes the use of homology properties such as gene-duplication data and phyletic gene age to predict essential genes. This approach has been used to predict essential genes in bacterial species such as *Mycoplasma* [20], *Liberibacter*[21], also in *P. falciparum* [22]and *Brucella spp*. [15].

Constraint Based approach uses Genome-scale metabolic network to elucidate the biology of metabolic pathways within an organism. The properties of the metabolic network can be analyzed using constraint-based methods such as flux balance analysis (FBA), which predicts the fluxes of metabolites at a steady state by applying mass balance constraints to a

stoichiometric model [23–26]. The concept of predicting essential genes using FBA is to simulate the knockout of a gene and evaluate the effect or impact on the network [27]. The use of FBA is better suited for studying *conditional* essential genes because a condition can be represented as an objective function and the significance of a gene can be determined by *in silico* deletion of the gene and the lethality is determined if there is optimal production of predefined biosynthetic precursors. Conditional essential genes are genes that are only essential in a given context. An example is immune response condition in an organism, genes responsible for immune response might not be essential if there is no disease condition in the organism. However, they become essential when the organism is in a diseased condition.

The ability of a computer system to use statistical technique to "learn" and "improve" with data in order to accurately predict outcomes without being explicitly programmed is known as **Machine learning**[28]. This approach involves constructing and training one or more classifiers with training data which is composed of features of known essential genes and non-essential genes. The trained classifier is then applied to predict the essentiality of genes in the target organism. For instance, Yu *et al*. [29] generated fractal features from genomic sequence of different 27 bacteria species and applied them to five classifiers to predict essential genes. It can be inferred that making accurate predictions requires "good" data and efficient machine learning technique. Machine learning techniques can be supervised, unsupervised or reinforcement learning. However, for essential gene prediction it often requires classification which is one of the supervised learning methods. A simple illustration of the process flow of collecting raw heterogeneous data from different sources to generate relevant features used to train a classifier and subsequently make predictions is shown in Figure 1.

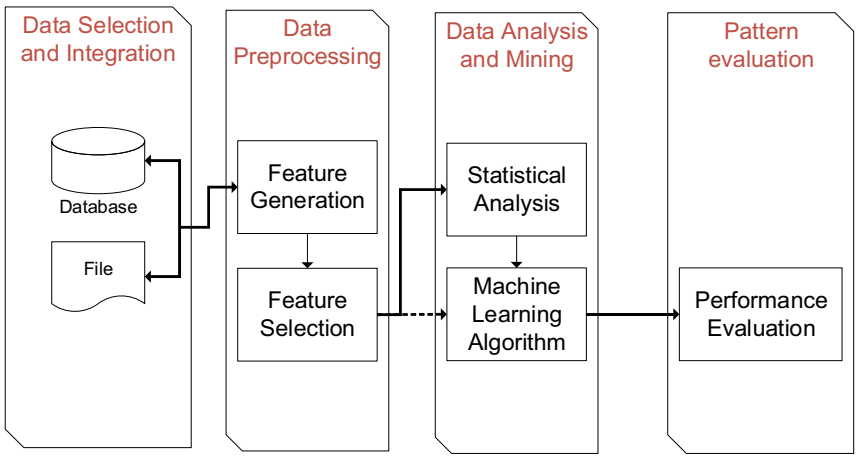

Figure 1. Simple illustration of application of machine learning to predict essential genes.

Data mining tools and machine learning (ML) algorithms have been used for classification. Open source tools such as RapidMiner, WEKA, R, and Orange provide rich functionality for data analysis and visualization.

2.0 Materials and Methods

Our comprehensive assembly of essential gene information for both *S. cerevisiae* and *Schizosaccharomyces pombe* was obtained from Database of Essential Genes (DEG) [30]and Online GEne Essentiality (OGEE) databases [31]. A total of 1037 essential genes and 4543 non-essential were obtained for *S. cerevisiae* and 1346 essential genes and 3689 non-essential obtained for *S. pombe*. This leads to an imbalance dataset available for the classification model development

2.1. Feature Generation

A large set of initial features was generated based on four different categories including protein sequence, gene sequence, topological features derived from protein interaction and gene sets enrichment from Gene Ontology, shown in Figure 2.

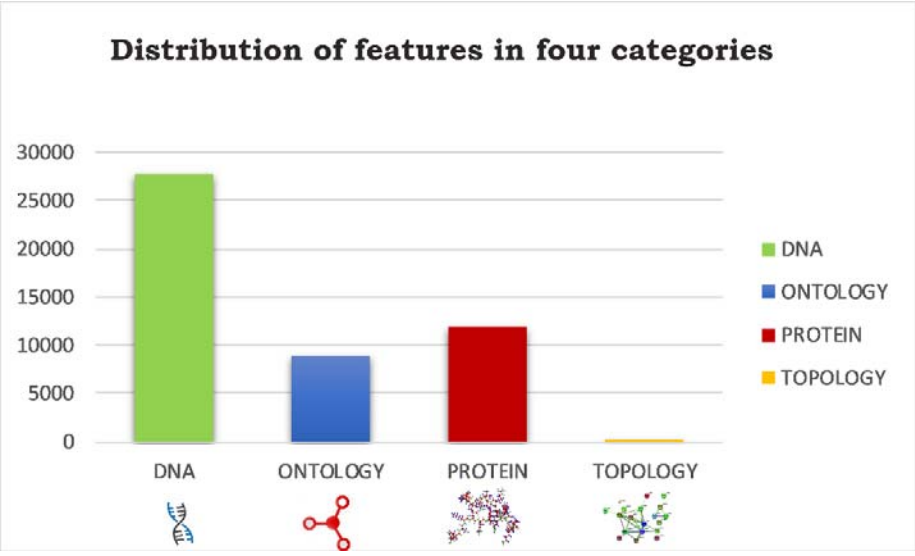

Figure 2. Multiple features were extracted from the four categories of features for gene essentiality prediction. DNA sequence category has highest number of features.

Protein and DNA sequences were obtained from Ensemble database using *Biomart* tool [32]. The protein and gene sequence features were encoded in various numerical representations

characterizing the nucleotide and amino acid sequences and compositions of the query genes were calculated using *seqinR*[33],*protr*[34], *CodonW*[35]and *rDNAse*[36].

With *seqinR* the number and fraction of individual amino acids and other simple protein sequence information including the number of residues, the percentage of physico-chemical classes and the theoretical isoelectric point were calculated. Most protein sequence features were obtained using *protr* including autocorrelation, CTD, conjoint triad, quasi-sequence order and pseudo amino acid composition. *CodonW* was used to calculate simple gene characteristics like length and GC content but also frequency of optimal codons and effective number of codons. With *rDNAse*, gene descriptors like auto covariance or pseudo nucleotide composition and *kmer* frequencies (n=2-7) were calculated. To predict the subcellular localization of the query protein, *Deeploc*[37], a tool that predictsthe probability of a gene being expressed in all the twelve subcellular compartments described for eukaryotic cells (Membrane, Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell membrane, Endoplasmic reticulum, Plastid, Golgi apparatus, Lysosome/Vacuole and Peroxisome) was used. *Interproscan*provides functional analysis of proteins by scanning sequences against Interpro's predictive models, provided by several different databases thereby classifying them into families and predicting domains [38].

Topology features were computed from protein-protein interaction (PPI), however, there are other sources of data where topology features can be extracted such as transcription profiles and metabolic pathways. The PPI network was assembled for both *S. cerevisiae* and *S. pombe* using the PPI information from *STRING* database [39]. An undirected graph was generated and topology features (including degree, degree distribution, betweenness, closeness and clustering coefficient) were calculated using *Networkx*[40]and *graphrole*[41,42]packages in python.

The Ontology category comprises gene ontology terms and orthology features such as KEGG orthologyamong others. They provide information about the enrichment of a given gene or gene set in a pathway or genome. In this study, 8846 and 8974 Gene Ontology (GO) terms were collected for *Saccharomyces cerevisiae* and *S. pombe* respectively, including biological process, cellular localization and molecular function from *gProfiler*[43]. To numerically encode the GO terms, an enrichment test was performed employing Fisher's exact test and the log of the *P-values* from the test represents the score for each gene per GO term.

2.2. Data normalization and feature selection

The numerical representation of each feature category was *z-score* transformed separately. ElasticNetCV, a cross-validation version of ElasticNet which iteratively cross validates the

partitioned data to select the optimal parameters for feature selection. The major parameters optimized are the *alpha* and *l1 ratio* with value range of $0 \leq alpha|li\_ratio \leq 1$. The *l1_ratio* parameter corresponds to *alpha* in the glmnet R package while parameter *alpha* corresponds to the *lambda* parameter in glmnet. ElasticNet uses a modification of Least Absolute Shrinkage and Selection Operator (LASSO) by adding Ridge regression into the optimization criterion. ElasticNet was used from the "sklearn" package in Python [44].

2.3. Sub-sampling, Machine Learning training and performance evaluation

To overcome class imbalances when training the classifiers, Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE creates synthetic, non-duplicated samples of the minority class balancing the total number of samples of the two classes [45]. For each sample of the minority class, SMOTE calculates the *k* nearest neighbors of the same class and randomly creates multiple synthetic samples between the observation and the nearest neighbors depending on the number of additional samples needed. Random Forests (RF), Artificial Neural Networks (NNET) and Extreme Gradient Boosting (XGB) from the *sklearn* package [44]were used as classification algorithms. Default parameters were used for the methods except for RF where the *n_estimator* parameter was set to 300. Stratified randomized 5-fold cross validation (CV) was performed to improve generalizability; where 80% of the data was used for feature selection and training of the classifiers, and 20% for testing.

In this study, four evaluation metrics were used to estimate the performance of the classification models, the metrics include; Precision, Sensitivity, Area under the receiver operating characteristic curve (AUROC)and Area under precision recall curve (AUPRC). Precision computes the rate of positive predicted value which estimates the reliability of the positive predictions of the model, also AUPRC estimates precision over the range of all possible values of recall. Similarly, sensitivity estimates the quality of positive prediction from the total predictions made by the model and AUROC quantifies True Positive rate over the range of all possible False Positive rates. These two metrics are important for essentiality predictions which aim to mainly identify or distinguish positive samples.

3. Results and Discussion

3.1 Gene Ontology features outperforms other categories of features

A total of 48535 features that spans across four categories were generated, namely; DNA sequence (27727 features), Protein sequence (11937 features), Network topology (25 features)

and Ontology (8846 features). Essential gene information was obtained from DEG and OGEE databases. Feature selection was performed to reduce the complexity of the model and a 5-fold cross-validation ML protocol was applied in which the imbalances in the class labels were corrected based on training data. Finally, the overall performance was estimated using the validation dataset.

Three ML algorithms were applied for the classification of essential genes i.e. a neural network (NNET), random forests (RF) and Extreme Gradient Boosting (XGB). In general, all three approaches yielded very similar performance results, but NNET performed slightly better than RF and XGB without model optimization (Figure 3). Gene ontology feature category outperformed other categories with AUROC of 0.936, AUPRC of 0.814 for *S. cerevisiae* and AUROC of 0.808, AUPRC of 0.633 for *S. pombe*. Followed by gene ontology is topology features with AUROC of 0.764, AUPRC of 0.470 for *S. cerevisiae* and AUROC of 0.715, AUPRC of 0.486 for *S. pombe*. DNA sequence category performs least with AUROC of 0.607, AUPRC of 0.261 for *S. cerevisiae* and AUROC of 0.549, AUPRC of 0.314 for *S. pombe*as shown in Table 1. DNA sequence category showed very weak ability to distinguish essential genes from non-essential genes.

Table 1: Accuracy metrics for the performance evaluation of essential gene classification.

		S. cerevisiae				*S. pombe*			
		Topology	**DNA**	**Protein**	**Ontology**	**Topology**	**DNA**	**Protein**	**Ontology**
RF	**AUROC**	0.749	0.67	0.696	0.908	0.675	0.608	0.647	0.824
	AUPRC	0.451	0.323	0.321	0.723	0.444	0.351	0.376	0.642
	Precision	0.409	0.545	0.385	0.706	0.402	0.41	0.467	0.579
	Sensitivity	0.477	0.017	0.041	0.601	0.49	0.051	0.063	0.561
XGB	**AUROC**	0.755	0.663	0.703	0.887	0.702	0.606	0.638	0.812
	AUPRC	0.458	0.317	0.343	0.672	0.479	0.358	0.379	0.631
	Precision	0.342	0.473	0.471	0.594	0.395	0.381	0.427	0.59
	Sensitivity	0.675	0.059	0.126	0.669	0.641	0.169	0.207	0.579
NNET	**AUROC**	0.764	0.607	0.712	0.936	0.715	0.549	0.632	0.808
	AUPRC	0.47	0.261	0.39	0.814	0.486	0.314	0.395	0.633
	Precision	0.341	0.311	0.403	0.753	0.4	0.321	0.379	0.578
	Sensitivity	0.691	0.211	0.422	0.732	0.659	0.302	0.506	0.6

Three ML approaches were used (neural network [NNET], Random Forests, [RF], and Extreme Gradient Boosting, [XGB]). Four performance metrics were used to evaluate the models

on two different organisms. The performance was measured for the test sets and ontology category distinctively has better results highlighted in bold border compared to the results from other categories. The result of the NNET machine is further presented in Figure3 where Ontology features has highest Area Under the Curve for both Receiver operating characteristic and Precision-Recallcurve.

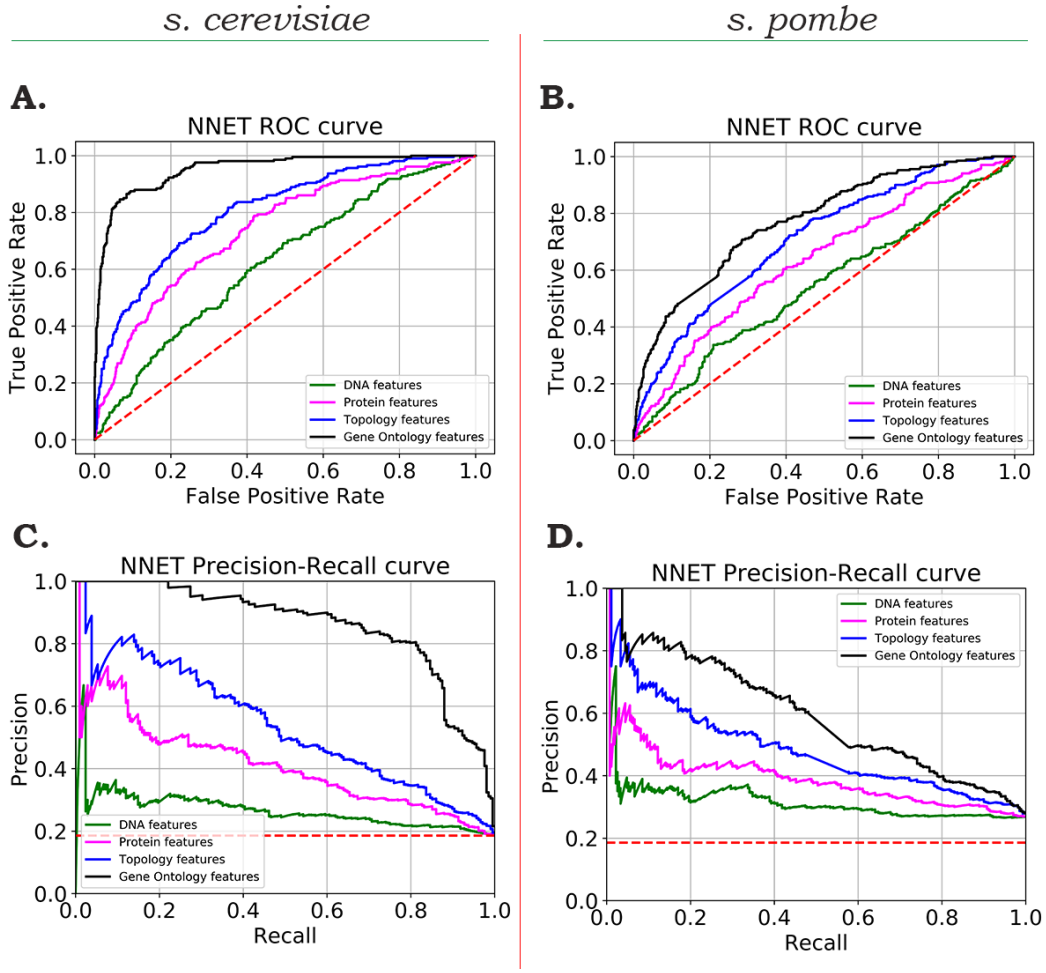

Figure 3: Receiver operating characteristic curve (**A**) and (**B**) for *S. cerevisiae*. and *S. pombe* respectively. Precision-Recall curve (**C**) and (**D**) for *S. cerevisiae*. and *S. pombe* respectively.

3.2 Analysis of features with high discriminative power

The 10 most important features and their correlation to essentiality are shown in Figure 4. All the top features in protein and topology categories are positively correlated to essentiality which implies that the higher the values of this features for a given gene the higher the probability of the gene to be an essential gene.Plaimas*et al*. [6] used network topology features to predict essential genes and reported similar trend as shown in Figure 4d. The positive correlation of ontology features to essentiality shown in Figure 4b implies that genes that are enriched (*p* value

< 0.05) in this ontology terms have high probability of being essential. The highest importance score in DNA (Figure 4a) and protein (Figure 4c) sequence categories are approximately 0.002 and 0.007 respectively, which is abysmally poor compared to 0.12 importance provided by a derivative of degree centrality in topology category. Strikingly, a topology feature appears to have the highest importance more than any of the ontology features.

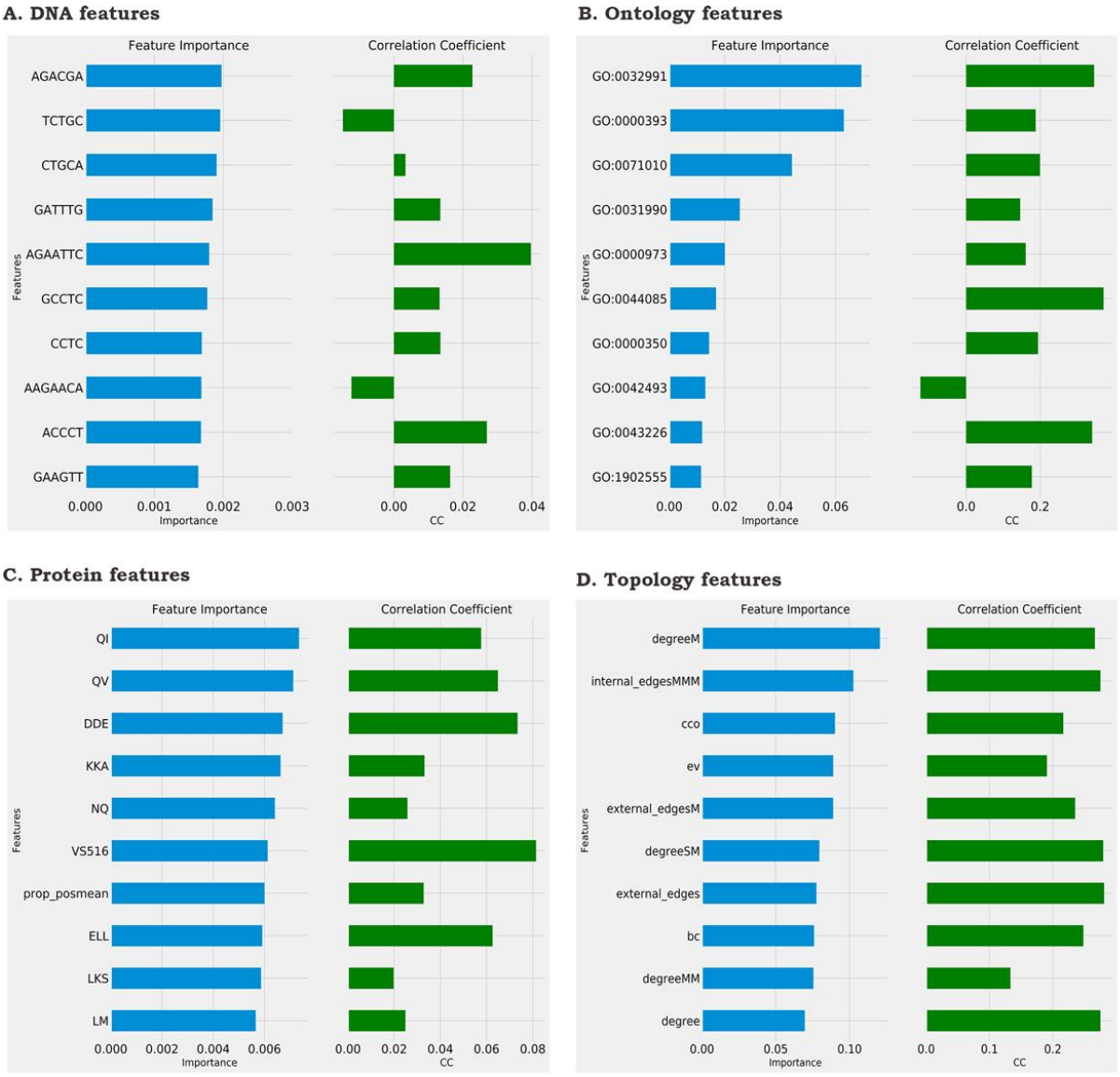

Figure 4: Top tenfeatures that contributed substantially to the predictions from each category.

Features were ranked based on their discriminative power. The blue bars represent the ranking while the green bars indicate the direction and correlation (positive or negative) of the feature to essentiality.

Previous studies have shown the informative power of ontology-based features. Zhang *et al*.[48] incorporated orthology information with network topology features and reported their method obtained about 66% improvements over the 5 traditional centrality measures (Betweenness centrality, closeness centrality (CC), eigenvector centrality (EC), and subgraph centrality and Degree centrality), which highlights the effect of ontology-based features in the model performance. Wei *et al*. [49] included orthology features with phylogeny features to develop a gene essentiality prediction tool (GETOP) that achieved AUROC of 0.918 intra-organism prediction and AUROC scores between 0.569 and 0.959 in the cross-organism predictions for 19 organisms. Chen *et al*. [12] predicted essential genes using the information about enrichments of gene sets defined by Gene Ontology and KEGG Orthology to encode each gene into a vector in which each component represented the relationship between the gene and one GO term or KEGG pathway. They achieved Matthews correlation coefficient of 0.951.

4. Conclusion

Machine learning approach using several categories of features has significantly contributed to essentiality prediction in model organisms. However, no previous studies have identified the most effectual category which can provide discriminating power to classify essential genes in both model and non-model organisms. In this study, four major categories of features in *S. cerevisiae* and *S. pombe*were compared in order to determine the most informative feature category which can enhance ML prediction of essential genes. Gene ontology feature category outperforms other feature categories considered. This study hereby proposes that more numerical representation of functional (Gene Ontology) terms should be engineered such as the enrichment test, which will further improve prediction performance of essential genes.

References

[1] Hart, T.; Brown, K.R.; Sircoulomb, F.; Rottapel, R.; Moffat, J. Measuring Error Rates in Genomic Perturbation Screens: Gold Standards for Human Functional Genomics. *Mol. Syst. Biol.*, **2014**, *10*, 733.

[2] JING, M.A. Metabolic Network Based Gene Essentiality Analysis, **2012**.

[3] Qin, C.; Sun, Y.; Dong, Y. A New Method for Identifying Essential Proteins Based on Network Topology Properties and Protein Complexes. *PLoS One*, **2016**, *11*, e0161042.

[4] Mobegi, F.M.; van Hijum, S.A.F.T.; Burghout, P.; Bootsma, H.J.; de Vries, S.P.W.; van der

Gaast-de, C.E.; Simonetti, E.; Langereis, J.D.; Hermans, P.W.M.; de Jonge, M.I. From Microbial Gene Essentiality to Novel Antimicrobial Drug Targets. *BMC Genomics*, **2014**, *15*, 958.

[5] Hurst, L.D.; Smith, N.G.C. Do Essential Genes Evolve Slowly? *Curr. Biol.*, **1999**, *9*, 747–750.

[6] Plaimas, K.; Eils, R.; König, R. Identifying Essential Genes in Bacterial Metabolic Networks with Machine Learning Methods. *BMC Syst. Biol.*, **2010**, *4*, 56.

[7] Gil, R.; Silva, F.J.; Peretó, J.; Moya, A. Determination of the Core of a Minimal Bacterial Gene Set. *Microbiol. Mol. Biol. Rev.*, **2004**, *68*, 518–537.

[8] Chen, Y.; Xu, D. Understanding Protein Dispensability through Machine-Learning Analysis of High-Throughput Data. *Bioinformatics*, **2005**, *21*, 575–581.

[9] Gustafson, A.M.; Snitkin, E.S.; Parker, S.C.; DeLisi, C.; Kasif, S. Towards the Identification of Essential Genes Using Targeted Genome Sequencing and Comparative Analysis. *BMC Genomics*, **2006**, *7*, 265.

[10] Seringhaus, M.; Paccanaro, A.; Borneman, A.; Snyder, M.; Gerstein, M. Predicting Essential Genes in Fungal Genomes. *Genome Res.*, **2006**, *16*, 1126–1135.

[11] Li, Z.; Li, B.-Q.; Jiang, M.; Chen, L.; Zhang, J.; Liu, L.; Huang, T. Prediction and Analysis of Retinoblastoma Related Genes through Gene Ontology and KEGG. *Biomed Res. Int.*, **2013**, *2013*.

[12] Chen, L.; Zhang, Y.-H.; Wang, S.; Zhang, Y.; Huang, T.; Cai, Y.-D. Prediction and Analysis of Essential Genes Using the Enrichments of Gene Ontology and KEGG Pathways. *PLoS One*, **2017**, *12*, e0184129.

[13] Campos, T.L.; Korhonen, P.K.; Gasser, R.B.; Young, N.D. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. *Comput. Struct. Biotechnol. J.*, **2019**.

[14] Mushegian, A.R.; Koonin, E. V. A Minimal Gene Set for Cellular Life Derived by Comparison of Complete Bacterial Genomes. *Proc. Natl. Acad. Sci.*, **1996**, *93*, 10268–10273.

[15] Yang, X.; Li, Y.; Zang, J.; Li, Y.; Bie, P.; Lu, Y.; Wu, Q. Analysis of Pan-Genome to Identify the Core Genes and Essential Genes of Brucella Spp. *Mol. Genet. genomics*, **2016**, *291*, 905–912.

[16] Salleh, A.H.M.; Mohamad, M.S.; Deris, S.; Illias, R.M. Identifying Minimal Genomes and Essential Genes in Metabolic Model Using Flux Balance Analysis. In *Asian Conference*

on Intelligent Information and Database Systems; Springer, **2013**; pp. 414–423.

[17] Gatto, F.; Miess, H.; Schulze, A.; Nielsen, J. Flux Balance Analysis Predicts Essential Genes in Clear Cell Renal Cell Carcinoma Metabolism. *Sci. Rep.*, **2015**, *5*, 10738.

[18] Aromolaran, O.; Beder, T.; Oswald, M.; Oyelade, J.; Adebiyi, E.; Koenig, R. Essential Gene Prediction in Drosophila Melanogaster Using Machine Learning Approaches Based on Sequence and Functional Features. *Comput. Struct. Biotechnol. J.*, **2020**.

[19] Deng, J.; Deng, L.; Su, S.; Zhang, M.; Lin, X.; Wei, L.; Minai, A.A.; Hassett, D.J.; Lu, L.J. Investigating the Predictability of Essential Genes across Distantly Related Organisms Using an Integrative Approach. *Nucleic Acids Res.*, **2011**, *39*, 795–807.

[20] Liu, W.; Fang, L.; Li, M.; Li, S.; Guo, S.; Luo, R.; Feng, Z.; Li, B.; Zhou, Z.; Shao, G. Comparative Genomics of Mycoplasma: Analysis of Conserved Essential Genes and Diversity of the Pan-Genome. *PLoS One*, **2012**, *7*, e35698.

[21] Fagen, J.R.; Leonard, M.T.; McCullough, C.M.; Edirisinghe, J.N.; Henry, C.S.; Davis, M.J.; Triplett, E.W. Comparative Genomics of Cultured and Uncultured Strains Suggests Genes Essential for Free-Living Growth of Liberibacter. *PLoS One*, **2014**, *9*, e84469.

[22] Rout, S.; Warhurst, D.C.; Suar, M.; Mahapatra, R.K. In Silico Comparative Genomics Analysis of Plasmodium Falciparum for the Identification of Putative Essential Genes and Therapeutic Candidates. *J. Microbiol. Methods*, **2015**, *109*, 1–8.

[23] Kauffman, K.J.; Prakash, P.; Edwards, J.S. Advances in Flux Balance Analysis. *Curr. Opin. Biotechnol.*, **2003**, *14*, 491–496.

[24] Orth, J.D.; Thiele, I.; Palsson, B.Ø. What Is Flux Balance Analysis? *Nat. Biotechnol.*, **2010**, *28*, 245.

[25] Papp, B.; Pal, C.; Hurst, L.D. Metabolic Network Analysis of the Causes and Evolution of Enzyme Dispensability in Yeast. *Nature*, **2004**, *429*, 661–664.

[26] Raman, K.; Chandra, N. Flux Balance Analysis of Biological Systems: Applications and Challenges. *Brief. Bioinform.*, **2009**, *10*, 435–449.

[27] Basler, G. Computational Prediction of Essential Metabolic Genes Using Constraint-Based Approaches. In *Gene Essentiality*; Springer, **2015**; pp. 183–204.

[28] Sakr, S.; Elshawi, R.; Ahmed, A.M.; Qureshi, W.T.; Brawner, C.A.; Keteyian, S.J.; Blaha, M.J.; Al-Mallah, M.H. Comparison of Machine Learning Techniques to Predict All-Cause Mortality Using Fitness Data: The Henry Ford exercIse Testing (FIT) Project. *BMC Med. Inform. Decis. Mak.*, **2017**, *17*, 174.

[29] Yu, Y.; Yang, L.; Liu, Z.; Zhu, C. Gene Essentiality Prediction Based on Fractal Features

and Machine Learning. *Mol. Biosyst.*, **2017**, *13*, 577–584.

[30] Luo, H.; Lin, Y.; Gao, F.; Zhang, C.T.; Zhang, R. DEG 10, an Update of the Database of Essential Genes That Includes Both Protein-Coding Genes and Noncoding Genomic Elements. *Nucleic Acids Res.*, **2014**, *42*, D574–D580.

[31] Chen, W.-H.; Lu, G.; Chen, X.; Zhao, X.-M.; Bork, P. OGEE v2: An Update of the Online Gene Essentiality Database with Special Focus on Differentially Essential Genes in Human Cancer Cell Lines. *Nucleic Acids Res.*, **2016**, gkw1013.

[32] Smedley, D.; Haider, S.; Ballester, B.; Holland, R.; London, D.; Thorisson, G.; Kasprzyk, A. BioMart–biological Queries Made Easy. *BMC Genomics*, **2009**, *10*, 22.

[33] Charif, D.; Lobry, J.R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In *Structural approaches to sequence evolution*; Springer, **2007**; pp. 207–232.

[34] Xiao, N.; Cao, D.-S.; Zhu, M.-F.; Xu, Q.-S. protr/ProtrWeb: R Package and Web Server for Generating Various Numerical Representation Schemes of Protein Sequences. *Bioinformatics*, **2015**, *31*, 1857–1859.

[35] Peden, J. CodonW. *Univ. Nottingham*, **1997**.

[36] Zhu, M.; Dong, J.; Cao, D.-S. rDNAse: R Package for Generating Various Numerical Representation Schemes of DNA Sequences. **2016**.

[37] Almagro Armenteros, J.J.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning. *Bioinformatics*, **2017**, *33*, 3387–3395.

[38] Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G. InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics*, **2014**, *30*, 1236–1240.

[39] Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P. STRING v11: Protein–protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.*, **2018**, *47*, D607–D613.

[40] Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Los Alamos National Lab.(LANL), Los Alamos, NM (United States), **2008**.

[41] Henderson, K.; Gallagher, B.; Li, L.; Akoglu, L.; Eliassi-Rad, T.; Tong, H.; Faloutsos, C. It's Who You Know: Graph Mining Using Recursive Structural Features. In *Proceedings*

of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; **2011**; pp. 663–671.

[42] Henderson, K.; Gallagher, B.; Eliassi-Rad, T.; Tong, H.; Basu, S.; Akoglu, L.; Koutra, D.; Faloutsos, C.; Li, L. Rolx: Structural Role Extraction & Mining in Large Graphs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*; **2012**; pp. 1231–1239.

[43] Reimand, J.; Kull, M.; Peterson, H.; Hansen, J.; Vilo, J. G: Profiler—a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments. *Nucleic Acids Res.*, **2007**, *35*, W193–W200.

[44] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **2011**, *12*, 2825–2830.

[45] Chawla, N. V; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.*, **2002**, *16*, 321–357.

[46] Olson, D.L.; Delen, D. *Advanced Data Mining Techniques*; Springer Science & Business Media, **2008**.

[47] Matthews, B.W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.*, **1975**, *405*, 442–451.

[48] Zhang, X.; Xiao, W.; Hu, X. Predicting Essential Proteins by Integrating Orthology, Gene Expressions, and PPI Networks. *PLoS One*, **2018**, *13*, e0195410.

[49] Wei, W.; Ning, L.-W.; Ye, Y.-N.; Guo, F.-B. Geptop: A Gene Essentiality Prediction Tool for Sequenced Bacterial Genomes Based on Orthology and Phylogeny. *PLoS One*, **2013**, *8*, e72343.