Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 09, pp.56875-56878, September, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

MACHINE LEARNING BASED PREDICTION OF ESSENTIAL GENES OF *SACCHAROMYCES CEREVISIAE* UTILISING PROTEIN ABUNDANCE AS A FEATURE

***,[1]Partha S. Das, [2,3]Sandip Chakroborty, [1]Keshab C. Mondal, [2]Tapash C. Ghosh and [1]Bikas R. Pati**

[1]Bioinformatics Infrastructure Facility, Department of Microbiology, Vidyasagar University, Medinipur, India, Pin-721102
[2]Bioinformatics Centre, Bose Institute, Kolkata, India, Pin-700054
[3]Department of Biology, University of Nevada, Reno, USA, Pin-NV 89557

ARTICLE INFO

Article History:
Received 12th June, 2017
Received in revised form
04th July, 2017
Accepted 23rd August, 2017
Published online 29th September, 2017

Key words:

Protein abundance, Essential gene,
Machine Learning, Neural Network,
Saccharomyces cerevisiae,
Rapidminer.

ABSTRACT

Protein abundance is a measure of expression of mRNA in a cell. The code of DNA is not expressed constitutively in all conditions; hence it is important to determine the population and abundance of proteins in a cell to determine cellular functions. Essential protein products are mandatory for functioning of a live cell. Thus patterns of abundance of essential and non-essential proteins may vary. A machine learning based framework has been applied here using neural n*et al*gorithm which could predict essential genes of S. cerevisiae using protein abundance as a feature using 77.65 percent accuracy.

Citation: Partha S. Das, Sandip Chakroborty, Keshab C. Mondal, Tapash C. Ghosh and Bikas R. Pati, 2017. "Machine learning based prediction of essential genes of *Saccharomyces cerevisiae* utilising protein abundance as a feature", *International Journal of Current Research*, 9, (09), 56875-56878.

INTRODUCTION

Proteins play a crucial role in our lives as they are required for almost all cellular processes. The deluge of sequence information found in biological databases gives an idea about the different mRNAs coded by the genome but it does not reflect the post-translational modifications and also the expression level during the normal phase of cellular life. The rapid development of high throughput methods in proteome studies has made possible to enlist the complete plethora of proteins found in cells in different phases of the life cycle (Schena *et al*., 1995; Eisen and Brown, 1999). The techniques include DNA Microarrays tagging the ORFs with high-affinity epitopes and expression from its natural chromosomal location (Ghaemmaghami *et al*., 2003), use of high throughput flow cytometry and GFP tagged yeast strains (Newman *et al*., 2006) to calculate the protein abundance level in a single cell, etc. *Saccharomyces cerevisiae* remains one of the most well studied organisms with a detailed profiling of protein concentrations in the nearly proteome wide level.

***Corresponding author: Partha S. Das,**
Bioinformatics Infrastructure Facility, Department of Microbiology, Vidyasagar University, Medinipur, India, Pin-721102.

Ghaemmaghami *et al*. (Akthar *et al*., 2012) reported that nearly 80% of the proteome is expressed during the normal growth conditions. Thus the abundance of proteins crucial for life may show a significant variance with the proteins expressed only during special conditions. The genome of an organism comprises of the complete set of genes that it is capable of encoding. But all of the genes are not transcribed and translated under any given condition. The flexibility and survivability that is exhibited by an organism to environmental perturbations is partially conferred by the genes that are constitutively expressed under all the conditions, and partially by a subset of genes that are induced under the defined conditions. An essential gene is defined here as a gene necessary for growth to a fertile adult (Kemphues and Kenneth, 2005). Thus essential genes of an organism may constitute its minimal gene set, which is the smallest possible group of genes that is adequate to sustain a functioning cellular life form under the most favourable conditions (Koonin, 2000; Juhas *et al*., 2011). The removal or effective mutation of only one of these genes is sufficient to confer a lethal phenotype on an organism irrespective of the presence of remaining genes. Therefore, the functions of essential genes are crucial for survival of an organism and may be viewed as a foundation of life. Identification of essential genes is important not only for the

understanding of the minimal requirements for cellular life, but also for practical purposes. Since most antibiotics target essential metabolic processes, essential proteins of microbial cells are being viewed as very effective targets for antimicrobial drugs (Juhas *et al*., 2012). It has been reported that PPI networks constructed using affinity purification methods for yeast and *Eschericia coli* demonstrate a correlation between protein degree, or number of interactions, and cellular abundance (Ghaemmaghami *et al*., 2003). Ning *et al*. (2010) also reported that there is a strong correlation between hub proteins and essential proteins. Essential proteins being crucial for cellular functioning occur in larger complexes and core proteins are involved in more number of biological processes than attachment proteins (Chakraborty *et al*., 2013). There have been several attempts for in-silico prediction of essential genes. The initial works were mostly based on sequence features of genes and proteins with or without homology comparison (del Rio *et al*., 2009). Advances in machine learning techniques enabled us to analyse more complex biological data and create predictive models for determination of essentiality of a candidate gene or protein. Later, with accumulation of data derived from experimental small-scale studies and high-throughput techniques it was possible to construct networks of gene and proteins interaction (De Las Rivas *et al*., 2012) and then investigate whether the topological properties of these networks would beuseful for predicting essential genes (Acencio and Lemke , 2009; Joy *et al*., 2005; Paladugu *et al*., 2008; Saha *et al*., 2006). Recently, a comprehensive review of use of topological features of biological networks used under a machine learning framework to predict essential genes have been published by Zhang, Acencio and Lemke (Zhang *et al*., 2016). Various other machine learning based methods like use of codon usage bias (Henry *et al*., 2007), disorderness of proteins (Das *et al*., 2016) etc. as classifier have also been demonstrated.

The protein expressions are regulated by on various factors in a cell and their spatial distribution also varies. But it is often observed that the different proteins maintain an innate and specific range of abundance levels in a cell (Rapidminer). In a growing yeast cell, the absolute abundance levels may range from 32 to 500000 copies per cell, with the rarest proteins being low abundant (Akthar *et al*., 2012). Ivanic *et al*. (2009) did an analysis of correlation of protein abundance with high throughput protein-protein interaction studies and came to a conclusion that essential proteins also show high abundance compared to their non-essential counterparts. Thus, we felt it will be interesting to assess the extent of relationship of protein abundance to its essentiality for the organism. A direct correlation based analysis of abundance with essentiality is not possible as Greenbaum *et al*. (Greenbaum *et al*., 2003), opined that the abundance data are often quite complex and noisy to ascertain their expression features. Machine learning approaches have effectively demonstrated that it can significantly classify and segregate between noisy data (Zhu *et al*., 2004). So in this work an attempt has been made to predict essentiality of proteins (and their corresponding genes) using abundance data as classifier in a machine learning framework and test if this could be used as a sole parameter to predict essentiality of a protein. The various methods mentioned above including topological properties and codon usage bias include many parameters to be used in the machine learning framework. Calculation of topological features of proteins in biological network is also acumbersome task. Thus the prediction of essentiality just using a single parameter, i.e. protein abundance may be an easy and simple process to implement.

MATERIALS AND METHODS

Gene sequence and related information of*S. cerevisiae* were downloaded from Ensembl (www.ensmbl.org) using R Programming environment (Aken *et al*., 2016). BiomaRt package of R (Durinck *et al*., 2010; Smedley *et al*., 2009) was used to extract the data from the Ensembl server (Biomart). The information about essential genes was downloaded from the Database of Essential Genes (DEG version) (Zhang *et al*., 2009). This information was used to segregate the yeast proteins among essential and non-essential types. There is no database of non-essential genes. Thus from the entire genepool of *S. cerevisiae*, the genes not included under essential gene database were regarded as non-essential. The protein abundance data were obtained from PaxDb: Protein Abundance Across Organisms (Wang *et al*., 2012). For machine learning framework, Rapidminer Studio (version 5.3.015, community edition), (Rapidminer) a widely accepted open source software environment for predictive analytics was used. The dataset employed here included 2564 *S. cerevisiae* proteins, out of which 577 were essential and the rest i.e 1987 were non-essential ones. As per the requirement of Rapidminer (Akthar *et al*., 2012), the data were formatted and arranged in a csv file for further analysis. The essential design of the system was reading the data from csv file and then assigning roles. The names of the yeast proteins were used as unique identifier (ID) and the case whether the particular protein was essential or not was used as label or outcome. The metadata are presented in Figure 1.

The data were channelized through a ten-fold cross-validation. The cross validation is a statistically accepted measure for evaluation of the performance of a machine learning algorithm. The X-Validation operator in Rapidminer is a nested operator. It has two sub-processes: a training sub-process and a testing sub-process. The training sub-process is used for training a model. The trained model is then applied in the testing sub-process. The trained model is then applied in the testing sub-process. The performance of the model is also measured during the testing phase. In this cross validation, shuffled sampling was used, which first shuffled the entire data, then selected 10% of that dataset and kept in a block. From the entire dataset, ten such blocks were produced. In the first instance, the rest of the 90% data were trained with the given classifier, and then the block of 10% data were used to test the accuracy of prediction. The accuracy of the prediction was noted against the label (which was not exposed to the algorithm during prediction) and this is the performance of the prediction. The performance was recorded by the system and the process was repeated, this time the second block of the 10% data being used for testing while the rest of the data (including the first block mentioned above) was used for training. The performance was recorded again. The entire process was repeated in a loop till the tenth block of the 10% data was used for testing. The averages of ten performances were used to conclude the overall performance of the classifier. This method thus eliminates chances of over fitting and biases in the training and performance measurements.

Figure 2 and 3 describes the arrangement of the operators in Rapidminer.

ExampleSet (2564 examples, 2 special attributes, 1 regular attribute)

Role	Name	Type	Statistics	Range	Missings
id	ID	text	mode = YAL002W (1), least = '	YAL002W (1), YAL007C (1), YA	0
label	essentiality	binominal	mode = 0 (1967), least = 1 (57	0 (1967), 1 (577)	0
regular	Protein abundance	numeric	avg = 12775.771 +/- 55683.71	[41.100 ; 1255722.326]	0

Fig. 1. Metadata view of Rapidminer

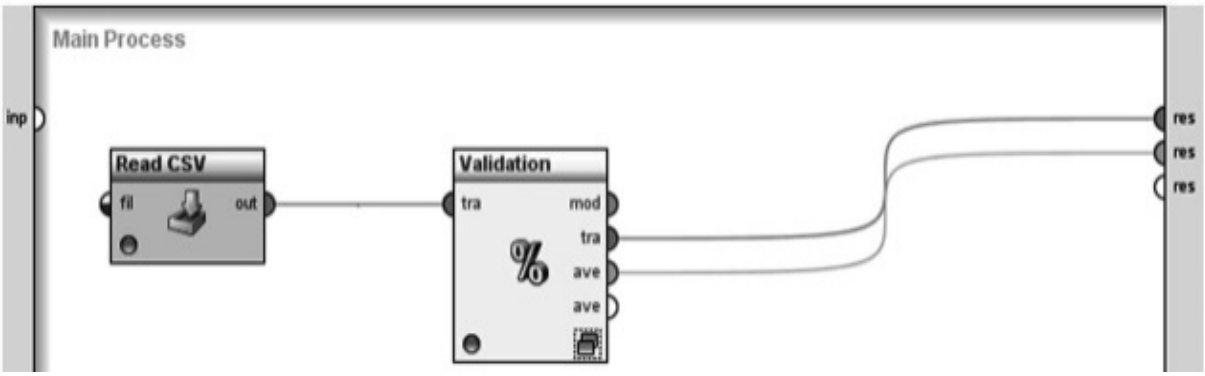

Fig. 2. Main process window of Rapidminer

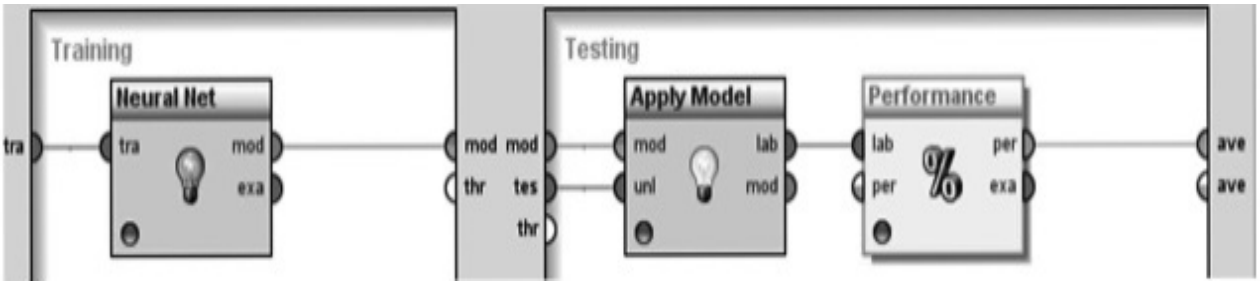

Fig. 3. Cross validation sub-process in Rapidminer

In the process of training and evaluation, neural network was used as a classifier. A neural network is a powerful computational data model that is able to capture and represent complex input/output relationships and thus this classifier was used for analysis. The parameters used for neural network were as follows: Hidden layers: 1, Training cycles: 500, Learning rate:0.3 and Momentum 0.2. This operator along with the model application and performance evaluation for each cycle is given in Figure 3, which runs in a nesting loop till the last block of the data is used for evaluation of performance.

RESULTS AND DISCUSSION

In the experiment, the Performance vector was found to be 77.65% indicating that the machine learning framework implemented here could perform with considerable amount of efficiency. The classification error was 22.35% while precision for essential genes was 100% and non-essential genes was 77.62% +/- 0.28% indicating that the system could predict the essential genes with very high accuracy while it failed to recognise non-essential genes in 22.35% cases. The abundance of non-essential proteins of a cell is generally found to be lower than the essential proteins, but it may not be true for all cases. Thus while the system could recognise the essential class of proteins with high accuracy it moderately lagged in doing so for recognising the non-essential ones.

As mentioned earlier, the non-essential genes were regarded as the subset of genes as a result of subtraction of essential genes from the entire yeast genome, and thus there could be a few candidates incorporated here as non-essential which are actually essential genes and not covered in the essential gene database, resulting in a lower predictive value. We feel as the main goal of the study is to find the essential genes or proteins, the high level of accuracy of the predictive modelling system employed here to recognise the essential ones among a pool of candidates is the main strength of the predictive modelling system.

Acknowledgements

Funding: This work was supported by the Bioinformatics Division, Department of Biotechnology, Ministry of science and technology, Government of India (grant number BT/BI/25/036/2012 (BIF-VUM)).

REFERENCES

Acencio, M. and Lemke, N. 2009. "Towards the Prediction of Essential Genes by Integration of Network Topology, Cellular Localization and Biological Process Information." *BMC Bioinformatics*, 10. doi:10.1186/1471-2105-10-290.

Aken, Bronwen L, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, *et al*., 2016. "The Ensembl Gene Annotation System." *Database* 2016 (January). doi:10.1093/database/baw093 .

Akthar, Fareed, Caroline Hahne, and Operator Reference, 2012. "RapidMiner 5," 990. http://www.rapid-i.com.

Chakraborty, Sandip, and Tapash Chandra Ghosh, 2013. "Evolutionary Rate Heterogeneity of Core and Attachment Proteins in Yeast Protein Complexes." *Genome Biology and Evolution* 5 (7): 1366–75. doi:10.1093/gbe/evt096.

Das, Partha S. Sandip Chakroborty, Mondal, Keshab C. Ghosh, Tapash C. and Pati and Bikas, R. 2016. "Protein disorderness based prediction of essential genes genes of Saccahromyces cerevisiae: a machine learning approach." International Journal of Current Research 8 (05): 31156–60.

De Las Rivas, Javier and Celia Fontanillo, 2012. "Protein-Protein Interaction Networks: Unraveling the Wiring of Molecular Machines within the Cell." *Briefings in Functional Genomics* 11 (6). England: 489–96. doi:10.1093/bfgp/els036.

del Rio, Gabriel, Dirk Koschutzki and Gerardo Coello, 2009. "How to Identify Essential Genes from Molecular Networks?" BMC Systems Biology 3 (October). England: 102. doi:10.1186/1752-0509-3-102.

Durinck, Steffen and Wolfgang Huber, 2010. "The biomaRt User's Guide." *Database*, 1–21.

Eisen, M.B. and Brown, P.O. 1999. "DNA Arrays for Analysis of Gene Expression." *Methods in Enzymology* 303. United States: 179–205.

Ghaemmaghami, Sina, Won-Ki Huh, Kiowa Bower, Russell W Howson, Archana Belle, Noah Dephoure, Erin K O'Shea, and Jonathan S. Weissman, 2003. "Global Analysis of Protein Expression in Yeast." *Nature* 425 (6959). England: 737–41. doi:10.1038/nature02046.

Greenbaum, Dov, Christopher Colangelo, Kenneth Williams, and Mark Gernstein, 2003. "Comparing Protein Abundance and mRNA Expression Levels on a Genomic Scale." *Genome Biol* 4: 117. doi:10.1186/gb-2003-4-9-117.

Henry, Ian, and Paul M. Sharp, 2007. "Predicting Gene Expression Level from Codon Usage Bias." *Molecular Biology and Evolution* 24 (1): 10–12. doi:10.1093/ molbev/msl148.

Ivanic, Joseph, Xueping Yu, Anders Wallqvist and JaquesReifman, 2009. "Influence of Protein Abundance on High-Throughput Protein-Protein Interaction Detection." *PloS One* 4 (6). United States: e5815. doi:10.1371/journal.pone.0005815.

Joy, Maliackal Poulo, Amy Brock, Donald E Ingber and Sui Huang, 2005. "High-Betweenness Proteins in the Yeast Protein Interaction Network." *Journal of Biomedicine & Biotechnology* 2005 (2): 96–103. doi:10.1155/JBB.2005.96.

Juhas, Mario, Leo Eberl and John I Glass, 2011. "Essence of Life: Essential Genes of Minimal Genomes." *Trends in Cell Biology* 21 (10). Elsevier Ltd: 562–68. doi:10.1016/j.tcb. 2011.07.005.

Juhas, Mario, Leo Eberl, and George M. Church, 2012. "Essential Genes as Antimicrobial Targets and Cornerstones of Synthetic Biology." *Trends in Biotechnology*. Elsevier Ltd. doi:10.1016/j.tibtech.2012.08. 002.

Kemphues, Kenneth, 2005. "Essential Genes."*WormBook : The Online Review of C. Elegans Biology, 1–7. doi:10.1895/wormbook.1.57.1.*

Koonin, E.V. 2000. "How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept." *Annual Review of Genomics and Human Genetics* 1. United States: 99–116. doi:10.1146/annurev.genom.1.1.99.

Newman, John R.S. Sina Ghaemmaghami, Jan Ihmels, David K. Breslow, Matthew Noble, Joseph L. De Risi, and Jonathan S. Weissman, 2006. "Single-Cell Proteomic Analysis of S. Cerevisiae Reveals the Architecture of Biological Noise." *Nature* 441 (7095). England: 840–46. doi:10.1038/nature04785.

Ning, Kang, HoongKee Ng, Sriganesh Srihari, Hon Wai Leong, and Alexey I. Nesvizhskii, 2010. "Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology." *BMC Bioinformatics* 11 (1): 505. doi:10.1186/1471-2105-11-505.

Paladugu, Sri R, Shan Zhao, Animesh Ray and AlpanRaval, 2008. "Mining Protein Networks for Synthetic Genetic Interactions." *BMC Bioinformatics* 9 (January): 426. doi:10.1186/1471-2105-9-426.

Rapidminer: www.rapidminer.com (Last accessed on 06.11.16)

Saha, Soma and Steffen Heber, 2006. "In Silico Prediction of Yeast Deletion Phenotypes." *Genetics and Molecular Research (Electronic Resource) : GMR.* 5 (1): 224–32.

Schena, Mark, Dari Shalon, Ronald W Davis and Patrick O. Brown, 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270 (5235), American Association for the Advancement of Science: 467–70. doi:10.1126/science.270.5235.467.

Smedley, Damian, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson and Arek Kasprzyk, 2009. "BioMart -- Biological Queries Made Easy." *BMC Genomics* 10 (1): 22. doi:10.1186/1471-2164-10-22.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S.P., Hengartner, M.O. and von Mering, C. 2012. "PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life." *Molecular & Cellular Proteomics* 11 (8):492–500. doi:10.1074/mcp.O111.014704

Zhang, Ren and Yan Lin, 2009. "DEG 5.0, a Database of Essential Genes in Both Prokaryotes and Eukaryotes." *Nucleic Acids Research* 37 (SUPPL. 1): 455–58. doi:10.1093/nar/gkn858.

Zhang, Xue, Marcio Luis Acencio and Ney Lemke, 2016. "Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review." *Frontiers in Physiology*. doi:10.3389/fphys.2016.00075.

Zhu, Xingquan and Xindong Wu, 2004. "Class Noise vs. Attribute Noise: A Quantitative Study." *Artificial Intelligence Review* 22 (3): 177–210. doi:10.1007/s10462-004-0751-8.

****