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

International Journal of Current Research
Vol. 8, Issue, 05, pp.31156-31160, May, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

PROTEIN DISORDERNESS BASED PREDICTION OF ESSENTIAL GENES OF *SACCAHROMYCES CEREVISIAE:* A MACHINE LEARNING APPROACH

[1]Partha Sarathi Das, [2,3]Sandip Chakroborty, [1]Keshab Chandra Mondal, [2]Tapash Chandra Ghosh and [*,1]Bikas Ranjan 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 23rd February, 2016
Received in revised form
04th March, 2016
Accepted 27th April, 2016
Published online 20th May, 2016

Key words:

Protein disorder, Intrinsically disordered proteins, Essential genes, Machine Learning, Neural Network.

ABSTRACT

Protein structure is phyletically conserved to serve its specific functions. But several proteins are found to be partially lacking a definite folded structure under specific conditions and this is known as disorderness. Intrinsically disordered proteins serve many important functions of the cell. Essential genes of an organism refer to the minimal gene set among the genome, mutation in any of which may confer a lethal or non-fertile phenotype. Essential genes show high phyletic retention and high degree of conservation. Thus comparison on the basis of disorderness of the proteins encoded by essential and non-essential genes may form a basis of segregation between these two kinds of genes. A machine learning framework using neural network as a classifier was implemented here which could successfully segregate the essential genes from non-essential ones.

Citation: Partha Sarathi Das, Sandip Chakroborty, Keshab Chandra Mondal, Tapash Chandra Ghosh and Bikas Ranjan Pati, 2016. "Protein disorderness based prediction of essential genes of *Saccahromyces cerevisiae*: A machine learning approach", *International Journal of Current Research*, 8, (05), 31156-31160.

INTRODUCTION

Protein structure is conserved to serve its function, but it is also seen that certain proteins lack a definite folded structure under certain physiological conditions like pH 7 and 25^0C (Uversky *et al*, Tompa *et al*). It does not contradict with the conservation of the overall structure of the proteins, because only a certain part of the protein is found to be in disordered state. The disorder serves to play important roles in interactions with other proteins or DNA. These proteins are also found to be 'adaptive'(Gunasekharan *et al*) RNA and protein chaperones, transcriptional regulators, signal transduction proteins, ion channels, and motor proteins are commonly found to have disordered regions (Dunker *et al*, Wright *et al*, Dyson *et al*, Demchenko *et al*). Thus disordered proteins could play crucial roles to keep the cell alive and in controlling gene expression. Studies on disordered proteins have revealed that certain amino acids occur in disordered regions, which have been termed disorder promoting (Campen *et al*, Wang *et al*) and various predictors have been designed on this theory (Campen *et al*, Ferron *et al*). Estimations of disordered fractions of many proteins from the proteomes of many species has been done with validations of the level of predictability (Dunker *et al.* (2000), Romero *et al*, Campen *et al*, Uversky *et al.*) In few interesting studies, the functional distributions of predicted disordered proteins were performed. At least 20 different proteins were studied using 710 Swiss Protein Functional keywords. It was found from that out of those 302 of the indicated functions were carried out by structured proteins while disordered proteins or disordered regions carried out around 238 of the functions (Xie *et al*). These functions by the latter group included signalling, regulation and control which is crucial for the life of the organism and also for its fertility.

Essential genes

The genome of an organism characterizes the complete set of genes that it is capable of encoding. However, not all of the

Corresponding author: Bikas Ranjan Pati,
Bioinformatics Infrastructure Facility, Department of Microbiology, Vidyasagar University, Medinipur, India Pin-721102.

genes are transcribed and translated under any defined condition. The robustness that an organism exhibits to environmental perturbations is partly conferred by the genes that are constitutively expressed under all the conditions, and partly 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). Essential genes of an organism constitute its minimal gene set, which is the smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions (Kunin *et al*, Glass *et al*). The deletion of only one of these genes is sufficient to confer a lethal phenotype on an organism regardless the presence of remaining genes. Therefore, the functions encoded by essential genes are crucial for survival and could be considered as a foundation of life itself. The identification of essential genes is important not only for the understanding of the minimal requirements for cellular life, but also forpractical purposes. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for such drugs (Sarangi *et al*).

The prediction and discovery of essential genes have been performed by experimental procedures such as single gene knockouts, RNA interference and conditional knockouts (Gustafson), but these techniques require a large investment of time and resources and they are not always feasible. Considering these experimental constraints, a computational approach capable of accurately predicting essential genes would be of great value. For prediction of essential genes, some investigators have implemented computational approaches in which most are based on sequence features of genes and proteins with or without homology comparison (Seringhaus *et al*). Sometimes the protein-protein interaction data and protein-protein networks have also been employed for detection of essentiality (Gabriel del Rio *et al*)

Implementation of machine learning techniques to predict microbial essential genes

Several attempts have been made to identify essential genes of prokaryotes through wet lab and *in-silico* techniques. In most cases the experimental basis of identifying essential genes of the organisms in the wet lab has been gene knock-out experiments (Kemphues *et al*) where a mutant was raised with a single gene "knocked out" and observation was recorded whether the mutation was lethal or if the organism was able to grow as a fertile being or not (Karp, Palsson). This is a very cumbersome task and needs huge sampling to validate the test cases. This has been successful in case of organisms like *E. coli (*Baba *et al)*, *S. cerevisae* (Smith *et al*) , *Mus musculus* (Bult *et al.*) etc. All these works have met varying degrees of success. *In-silico* techniques and machine learning methods have been attempted to predict the essential genes of the organisms mentioned above (Plaimas *et al*, Chen *et al*, Heber *et al*). The availability of the protein-protein interaction networks have made this possible (Gong *et al*).In some cases this system has been used to predict disease causing genes of prokaryotes. Various wet lab experiments have been performed to identify the essential genes of *S. cerevisiae* and a database has been created under DEG (Database of Essential Genes) (Zhang, 2009). DEG 5.0 has been hosted which contains information about essential genes of 14 different prokaryotes and 6 eukaryotes. In the database it is interesting to note that for the yeast (*Saccharomyces cerevisiae*) 1110 essential genes have been identified with high confidence. The machine learning techniques have been applied in different fields of bioinformatics (Brown *et al*, Furey *et al*) but little work has been done to identify essential genes or proteins of yeast. The speciality of the application of machine learning techniques lies to the fact that based on the data available on the cause and effect relationships, where the underlying reason is not clear or is not taught to the analytical system, still the machine learning system can create its own equations and may predict outcomes with the given variables which was not taught to it earlier. Through statistical methods like cross validation and performance measurements the accuracy etc. can be determined which proves the effectiveness of its predictability.

MATERIALS AND METHODS

Sequences of the genes of *S. cerevisiae* were downloaded from Ensemble (www.ensmbl.org) using R Programming environment. BiomaRt package of R (Smedley *et al*., Durnick *et al*.) was used to extract the data from the Ensemble server (Biomart). The information about essential genes were downloaded from the Database of Essential Genes (DEG version 5)(Zhang *et al*).This information was used to segregate the yeast proteins among essential and non-essential types. The disorderness of the proteins can be characterised by two parameters, viz. percentage of disorder and length of disorder. The length of disorder (dis_lengh) was calculated from Fold index, a web based tool based on Uversky et al which can be accessed at http://bip.weizmann.ac.il/fldbin/findex. The default parametric values were used during prediction of disorderness for both essential and non-essential proteins used for training the classifier. For machine learning framework, Rapidminer version 5.3.015 (community edition), 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 1987were non-essential ones. As per the requirement of Rapidminer, 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 subprocesses: a training subprocess and a testing subprocess. The training subprocess is used for training a model. The trained model is then applied in the testing subprocess. The performance of the model is also measured during the testing phase. In ten-fold 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.

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

Role	Name	Type	Statistics	Range	Missings
id	protein name	text	mode = YAL002W (1), least = YALC	YAL002W (1), YAL007C (1), YAL	0
label	essentialiity	binominal	mode = 0 (1987), least = 1 (577)	0 (1987), 1 (577)	0
regular	dis_length	integer	avg = 188.946 +/- 178.214	[0.000 ; 1669.000]	0
regular	%dis	real	avg = 36.299 +/- 24.810	[0.000 ; 100.000]	0

Figure 1. Metadata view of the Essential and non-essential proteins incorporated under the study

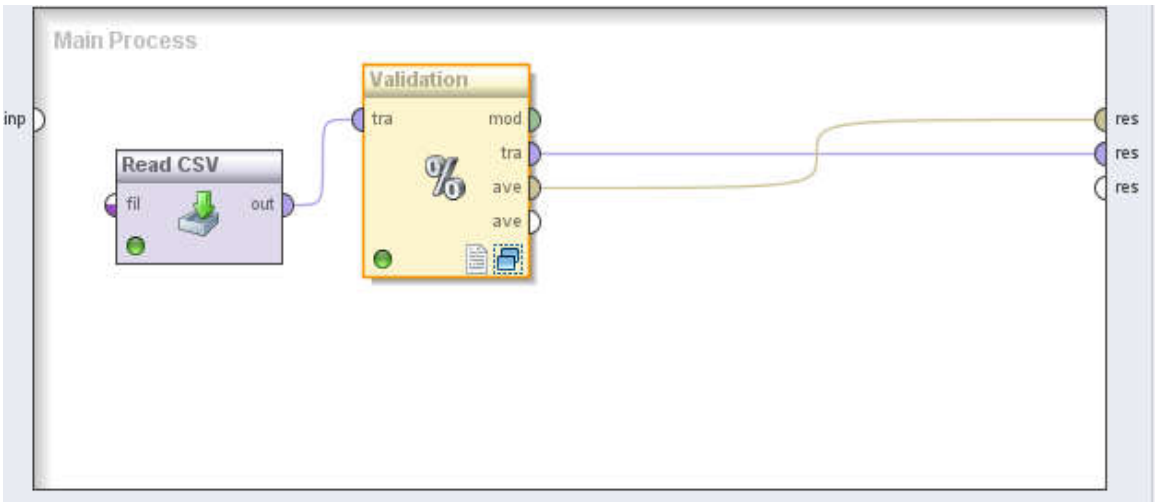

Fig. 2. The read CSV and cross validation operators of Rapidminer

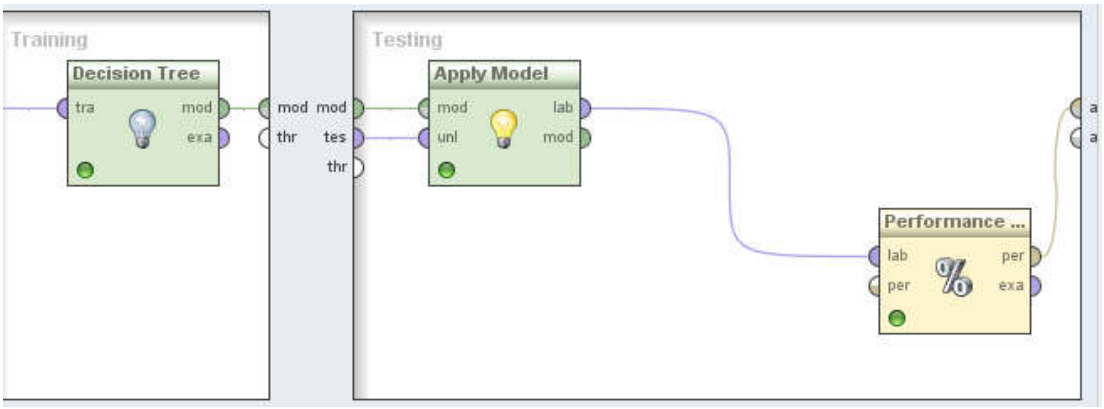

Figure 3. The nested portion of the machine learning setup

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 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 describes the arrangement of the operators. 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. The machine learning framework using neural network as a classifier could predict essential genes with 77.50 percent accuracy. The classification error was found to be 22.50% only. This proves that many essential genes possess fair amount of disorderness which enable them to perform crucial and essential functions of the cell.

RESULTS AND DISCUSSION

The measurement of percentage of disorder and length of disorder can provide a fair idea about the essentiality of the protein and this kind of machine learning framework can provide basis of automated prediction of essentiality of the protein (and its corresponding gene) in an easy and successful manner. This method provides an alternative to the cumbersome wet lab methods of detecting essential genes. Thus it could provide useful insights towards understanding the essential functions of the cell. These essential genes can also be used as an effective target for novel drugs.

REFERENCES

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L. and Mori, H. 2006. "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection," *Mol Syst Biol,* Vol. 2, 2006.

Brown, M.P., Grundy, W.N., Lin, D., *et al*. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl Acad. Sci. USA, 97, 262–7.4.

Bult, C. J., Kadin, J.A., Richardson, J.E., Blake, J. A. and Eppig, J.T. 2009. "The mouse genome database: Enhancements and updates," *Nucleic Acids Res*., Vol. 38, no. SUPPL.1, pp. 586–592, 2009.

Campen, A., Williams, R.M. ,Brown, C.J., Meng, J., Uversky, V.N. and Dunker, A.K. 2008. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder,Protein Pept Lett. 15(9): 956–963.

Demchenko, A.P. 2001. Recognition between flexible protein molecules:induced and assisted folding. J. Mol. Recognit. 14, 42–61

Dunker, A.K. *et al*. 2002. Intrinsic disorder and protein function.Biochemistry 41, 6573–6582

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C. and Brown, C.J. 2000. Proc. Genome Informatics 11, Tokyo, Japan, pp. 161-171

Durinck, S., Spellman, P., Birney, E. and Huber, W. 2009. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nature Protocols*, 4, pp. 1184–1191.

Dyson, H.J. and Wright, P.E. 2002. Coupling folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol*., 12, 54–60

Ferron, F., Longhi, S., Canard, B. and Karlin D. 2006. A Practical Overview of Protein Disorder Prediction, PROTEINS: Structure, Function, and Bioinformatics 65:1–14.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, 16, 906–14

Gong, X., Fan, S., Bilderbeck, A., Li, M., Pang, H. and Tao, S. 2008. "Comparative analysis of essential genes and nonessential genes in Escherichia coli K12," *Mol. Genet. Genomics,* Vol. 279, no. 1, pp. 87–94, 2008.

Gustafson, A.M., Snitkin, E.S., Parker, S.C.J., DeLisi, C. and Kasif, S. 2006. "Towards the identification of essential genes using targeted genome sequencing and comparative analysis.," *BMC Genomics*, vol. 7, p. 265, Jan. 2006.

Hwang, Y.C., Lin, C.C., Chang, J Y., Mori, H., Juan H.F. and Huang, H.C., "Predicting essential genes based on network and sequence analysis.," *Mol. Biosyst.,* Vol. 5, no. 12, pp. 1672–1678, 2009.

John, I. 2006. Glass, Essential genes of a minimal bacterium, PNAS January 10, 2006 vol. 103 no. 2 425-430

Karp, G. Cell and Molecular Biology: Concepts and Experiments

Kemphues, K. 2005. Essential Genes (December 24, 2005), *Worm Book*, ed. The *C. elegans* Research Community, Worm Book

Koonin, E.V. 2000. How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* 1:99-116

Palsson, B. 2006. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*, *J Bacteriol.,* 2006 December; 188(23): 8259–8271.

Plaimas, K., Eils, R. and König, R. 2010. "Identifying essential genes in bacterial metabolic networks with machine learning methods.," *BMC Syst. Biol.,* Vol. 4, p. 56.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. Dunker, A.K. 2001. Sequence complexity of disordered protein, Proteins. 2001 Jan 1;42(1):38-48.

Saha, S. and Heber, S., 2006. "In silico prediction of yeast deletion phenotypes.," Genet. Mol. Res. (electronic Resour.) GMR., Vol. 5, no. 1, pp. 224–232.

Sarangi, A.N., Aggarwal, R., Rahman, Q. and Trivedi, N. 2009. Subtractive Genomics Approach for *in Silico* Identification and Characterization of Novel Drug Targets in *Neisseria Meningitidis Serogroup B*. *J Comput Sci Syst Biol*., 2: 255-258.

Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M. and Gerstein, M. 2006. "Predicting essential genes in fungal genomes," pp. 1126–1135.

Smedley, D., Haider, S. ,Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. 2009. "BioMart--biological queries made easy.," *BMC Genomics*, vol. 10, p. 22, 2009.

Smith, L.K., Gomez, M.J., Shatalin, K.Y., Lee, H. and Neyfakh A.A. 2007. "Monitoring of gene knockouts: genome-wide profiling of conditionally essential genes," *Genome Biol.,* Vol. 8, no. 5, pp. 1–11.

Tompa, P. 2002. Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527–533

Uversky, V.N. *et al*. 2000. Why are 'natively unfolded' proteins unstructured under physiologic conditions? Proteins 41, 415–427

Wang, G., Dong, Y., Chen, H., Zhang, H., Song, Y., Zhang, H. and Chen, W. 2011. High Incidence of Disorder-Promoting Amino Acids in the Amino Terminus of Mature Proteins in Bacillus subtilis, *American Journal of Biochemistry and Biotechnology*, 7 (4): 172-178, 2011

Wright, P.E. and Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*., 293, 321–331

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N. and Obradovic, Z. 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, *J Proteome Res.*, May;6(5):1882-98.

Zhang, R. and Lin, Y. 2009. "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Res*., vol. 37, no. SUPPL. 1, pp. 455–458.

****