Abstract

INTRODUCTION:

Bioinformatics is an emerging and rapidly growing field of a cross-disciplinary science. As a consequence of the large amount of data produced in the field of molecular biology, most of the current bioinformatics projects deal with structural and functional aspects of genes and proteins. The data produced by thousands of research teams all over the world are collected and organized in databases specialized for particular subjects. The existence of public databases with billions of data entries requires a robust analytical approach to cataloguing and representing this with respect to its biological significance. Therefore, computational tools are needed to analyse the collected data in the most efficient manner.

Essential genes

The genome of an organism characterizes the complete set of genes that it is capable of encoding. However, not all of the 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 favourable 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 for practical purposes. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for such drugs (Sarangi A N, et al).

In the era of complete genomes, the total number of genes in a sequenced organism can now be predicted (Claverie), but the function and selective importance of a substantial fraction of genes remains unknown (Hollon). The conditional importance of genes in conferring robustness can be understood in the context of the functional attributes of these genes and their correlations to the defined environmental conditions. However, a priori prediction of such genes for a given condition is yet not possible.

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 (Gabriel del Rio et al). With the accumulation of data derived from experimental small-scale studies and high-throughput techniques, however, it is now possible to construct networks of gene and proteins interaction (De la Rivas J) and then investigate whether the topological properties of these networks would be useful for predicting essential genes.

Implementation of machine learning techniques to predict microbial essential genes

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 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* , *Mus musculus* (house mouse) etc . All these works have met varying degrees of success. In-silico techniques, 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 has made this possible (Gong et al). There have been attempts also to predict essential genes of *Saccharomyces*. In few cases this system has been used to predict disease causing genes of prokaryotes.

Application of neural network technique in identification of essential genes of S. cerevisiae

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).
In the database it is interesting to note that for the yeast (*Saccharomyces cerevisiae*) 1110 essential genes have been identified. It is very difficult and costly to knockout single genes from higher organisms and see their expression. Sometimes for ethical reasons it is inhibitory to conduct such experiments on primates and humans. Establishment of such a predictive model which may be later extended to higher organisms will translate into various benefits.

The machine learning techniques have been applied in different fields of bioinformatics (Brown et al, Furey et al, Hua et al) but little work has been done to identify yeast essential genes with a holistic approach. This study aims at optimizing the features to identify the essential genes of yeasts by machine learning technique. Availability of Saccharomyces genome database (YGD) (Engel S R et al.) will help to explore all the genes of yeast.

Feature selection

Identification of key parameters or features

A machine learning framework ideally depends on two key components training and testing. On the basis of training received the framework may predict outcomes of novel cases or inputs. Machine learning can be viewed as the acquisition of structural descriptions from examples. The kind of descriptions found can be used for prediction, explanation, and understanding.
Various kinds of features were considered to be tested under the machine learning framework. The features were selected based on their reference in published literature as important feature for essential genes or were proposed during the current research work.

i.Location of the gene in DNA strand

ii. ORF Length:

iii. Codon usage bias:

iv. Use of rare codons:

v. Expression Level:

vi. Disorderness of proteins

vi. Protein abundance:

vii. protein complex number :

viii. Protein-protein interaction (PPI) network

Machine Learning

Here in this work Rapidminer version 5.3.015 community edition was used.

Classifier selection

Neural network has been used as a classifier.

Machine learning framework

For machine learning framework, Rapidminer version 5.3.015 (community edition) which is 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.

Results and discussion:

The machine learning framework employed here could effectively segregate between essential and non-essential genes with 72.5% accuracy.