Review of literature

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 have made this possible (Gong et al). There have been attempts also to predict essential genes of *Saccharomyces* .In some cases this system has been used to predict disease causing genes of prokaryotes. There have been instances where the E.coli model and S. cerevisiae model have been tried to do cross-referenced prediction(Hwang et al).In all these cases several parameters like various topological properties of protein-protein interaction network (e.g clustering coefficient, betweennes centrality, common function degree), ORF length etc have been used. We feel incorporation of further parameters like codon usage analysis , use of rare codons, subcellular localization etc may enhance the prediction system further. The availability of experimentally established dataset of essential genes of various prokaryotes will help to assess the correctness of the prediction system.

The machine learning based predictions done so far have been concentrated on individual features like network topology or codon usage bias. This is the first attempt to have a holistic approach towards prediction of essentiality combining many heterogeneous parameters together. This may throw a better illumination towards the design of nature , how it views the essential functions of life and how it tries to preserve them.

Objectives:

- Elucidate the factors that may govern essentiality of the genes. *(selection of parameters)*
- Download data of genes with label “essential” or non- essential” along with parametric values
- Transform data to the format of a neural network package
- Tuning the learning rate and number of hidden layers
- Use cross-validation to statistically test the performance
- Use the best parameter to train the whole training set
- Test the data