A central repository for functional and evolutionary information of Human Proteins

Inderjit Singh Yadav, Dhwani Raghav and Subhash Mohan Agarwal*

*Bioinformatics Division, Institute of Cytology and Preventive Oncology (*icmr*), I–7, Sector-39, Noida-201301,* India

Address for correspondence:
E-mail: inderjitbioinfo@gmail.com

Evolution is a natural phenomenon occurring in nature that creates pressure on the organism for its survival. Evolutionary novelties are brought about by the changes in gene interactions which may gives rise to new gene functions or genes with altered activities resulting in diseased states. Therefore, the evolution rate of the genes can be used to answer a number of biological questions related to population process, species diversification, conservation biology and diseases occurrences. If the genes are fast evolving this means that an equilibrium state in its function is not reached, so it keeps mutating to achieve a better or a modified function. This fast mutation rate can lead to the formation of highly species specific protein module or an intermediate highly disturbed

disease state. While there are a set of genes, in which the mutation rate is less. These can be stated as slow evolving genes. Therefore, we have developed a comprehensive database of whole human proteome so that one can identify the rapidly and slowly evolving human genes based on the synonymous and non-synonymous substitution changes occurring in them and their relationship with the diseased state like cancer.

We downloaded the human chromosome files from the NCBI and in-house PERL scripts were developed to parse the gene features like gene id, protein accession, nucleotide information etc. from the input files. The identified human proteins were used to search the homologs and then evolutionary rate was calculated. Thereafter, the extract entities were relationally linked to build the schema. The backend of the database was designed in MySQL. The web interface i.e. front end of the database was generated using PHP codes.

The study aimed to provide a first ever comprehensive central depository of evolutionary data of human genome. The database is augmented with the features like gene and protein sequence, its location, architecture (i.e. exon-intron structure), function, homology to other eukaryotic genomes, as well as protein multiple sequence alignment. Moreover, the database also exhibit the different types of information about genes, including chromosomal positions, accession numbers, gene and CDS sizes, orthologs, protein structures and external links to other databases. The database is build up with an interactive web interface, and a knowledge based schema which ensure the user to easily retrieve the information. The database is enriched with various browsing section as well as search facility for querying the database and retrieving desired information. We have developed a database that allows user to query and retrieve information regarding evolutionary rate or phylogenetically conservedness of the human proteins.