Analysis of Annotation Strategies for Hypothetical Proteins: A Case Study of *Neisseria*

Neeraj Nagpal,
Neera Munjal,
Sayan Chatterjee

University School of Biotechnology, Guru Gobind Singh Indraprastha University, New Delhi

Address for correspondence:
E-mail: neerajnagpal49@gmail.com

Growing possibilities of biotechnology for genome sequencing lead to generation of sequences for millions of genes. However, function of majority of these genes is unknown, and can be determined experimentally only for a few of them. Therefore, a large part of proteomes is represented by hypothetical proteins (HP), i.e. proteins predicted from nucleic acid sequences only and protein sequences with unknown function. The usual scenario involving hypothetical protein is in gene identification during genome analysis. When the bioinformatics tool used for the gene identification finds a large open reading frame without an analog in the protein database, it returns "hypothetical protein" as an annotation remark.

Neisseria are a large family of commensal Gram Negative bacteria that colonize the mucosal surfaces of many animals. Of the 11 species that colonize humans, only two are pathogens- N. meningitidis, which causes meningococcal septicemia and *N. gonorrhoeae,* which causes gonorrhoea. This particular family of proteins was chosen because there is increasing prevalence of strains with resistance to antibiotics. So a way to development of drugs targeted specifically to these proteins remains as challenge to the researchers. The current study was functional annotation and localization of hypothetical proteins of *Neisseria gonorrhoea* FA 1090. The majorities of such 'hypothetical' genes have a wider phyletic distribution and therefore are usually referred to as 'conserved hypothetical proteins'. Therefore, the sequences of these proteins retrieved from Protein database, N.C.B.I. were subjected to comparative genomic approach by similarity searches using Protein Data Bank BLAST, COG database, Pfam database and PROSITE. Cluster of Orthologous Groups of Proteins (COGs) represent phylogenetic classification of proteins encoded in complete genomes. Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). PROSITE database describes protein domains, families and functional sites as well as associated patterns and profiles to identify them. In addition, Signal Peptide and Cleavage site was predicted using online tools-SignalP, SigPred, SigCleave, PrediSi and Phobius, and Transmembrane helices were detected in proteins using online tools-DAS, HMMTOP, TMHMM, TMPred, TopPred, Phobius and SOSUI. Since, membrane proteins play key roles in biological systems as pores, ion channels and receptors and are important in intracellular communication and coordination, they may serve as good drug targets - altering the function of signaling proteins may help correct defects in signaling that are the root of many diseases. Besides these, LipoP tool was also used which produces predictions of lipoproteins and discriminates between lipoprotein signal peptides, other signal peptides and n-terminal membrane helices. Further, prediction of localization of proteins was also performed with the help of online tools like CELLO which is a multi-class SVM classification system and PSORTdb based on information determined through laboratory experimentation (ePSORTdb dataset) and computational predictions (cPSORTdb dataset). This study was also successfully used to compare the different online tools available for similarity search, Signal peptide and cleavage site prediction, transmembrane topology and localization of proteins.