DEVELOPMENT OF NOVEL ALGORITHMS FOR FUNCTIONAL ANNOTATION OF PROTEIN SEQUENCES

Abstract

Protein molecules are the workhorses of cells. An in-depth knowledge of protein function is essential to understanding the working of a healthy organism and for mitigation of diseases. The most reliable form of function annotation is through biological experiments. However, the exponential rise in the number of protein sequences is making the job of experimental function detection a formidable challenge. With numerous methodological advancements in genome sequencing projects, number of new protein sequences reported has been steadily increasing. There are a total of 184,998,855 sequences in the protein sequence database. However, only ~69,306 entries are verified functionally using experimental methods. In this scenario, computational biologists must aid the experimental community in high-throughput function prediction. It would be immensely useful to gain insights into the activity of these biomolecules for guiding further experiments. Given that most of the drug targets are proteins, one of the pressing goals of protein function prediction is to aid in the development of novel drug targets and new therapeutics. While there are a total of 70,076 human proteins in UniProt and 25,995 experimentally determined structures, there are only ~3021 human protein targets known to bind with FDA approved drugs. Thus, predicting the function of all proteins is crucial to drug target identification and to initiate drug discovery. This thesis work focuses on the development of a few new computational protocols for protein function prediction, starting from sequence.

Chemical information residing in the protein sequences determines their structures and structures determine their functions. Hence, knowledge of both protein sequence and structure can be exploited for accurate function prediction. This thesis work focuses on evolving a methodology for function prediction using physicochemical features (Chapter-II). A comparison of diverse textbook classifications of amino acids in relation to function prediction is described and it is demonstrated that a new classification of amino acids (NCL) does better in function prediction (Chapter-III). Enzymes are vital protein molecules from both function and malfunction perspectives especially in the context of drug-discovery. An application of the NCL integrated with mask BLAST for enzyme function is demonstrated (Chapter-IV) and further extended to all proteins (69,306) with known function (Chapter-V). In function prediction, due to the rich diverse ideas and algorithms already generated, it pays to take a

consensus view. In this spirit, the NCL-mask BLAST method developed here is integrated with alternate approaches to create a metaserver, S2F, for function prediction which captures the merits of various methods (Chapter-VI). Perspectives emerging from the thesis work on accurate function prediction of millions of sequences in growing sequence databases are summarized (Chapter-VII).