Introduction
Chapter 1: Introduction

All activities of a living cell is governed by its genome which is defined as the 'sum total of nuclear and the mitochondrial genetic material'. Every organism (ranging from virus to human) possesses a genome that contains the biological information needed to construct and maintain a living organism. The complexity and architecture of genomes differ according to the phylogenetic groups. For example viruses contain naked genome protected by protein capsid. On the other hand prokaryotic genome lies freely in the cytoplasm. But eukaryotic genomes are confined inside a discreet membrane bound organelle called nucleus (Figure 1.1).

![Figure 1.1: Schematic structure of eukaryotic and prokaryotic cell.](image)

The biological information contained in a genome is encoded in the nucleotide sequences and is divided into discrete units called genes. Prokaryotic and eukaryotic genomes are also distinct from each other on the basis of gene organization. Prokaryotic genomes contain long open reading frame (stretch of genome sequence that has potential to code a protein) with very little intergenic space (Figure 1.2). In contrast, a typical eukaryotic cell has a very complex genome architecture consisting of non-coding and coding introns and exons respectively (Figure 1.3), used to generate splice variants of messenger RNA (mRNA) transcripts which in turn give rise to many different protein products from a given gene in eukaryotes. The information contained in a gene is read by a special class of proteins, which attach to
the genome at appropriate positions and initiate a series of biochemical reactions referred to as gene expression (Figure 1.4).

The expression pattern of genes eventually determines the phenotype of an organism. Gene expression is a highly regulated phenomenon involving two intricate steps of transcription and translation. During transcription the information contained in a gene is transferred to a class of RNA called as messenger RNA (mRNA) and the whole process is catalyzed and facilitated by RNA polymerase and transcription factors. In eukaryote cells the primary transcript (pre-mRNA) is often processed further via alternative splicing process where the blocks of mRNA are cut out and rearranged, to
produce different arrangements of the original sequence. Whereas, during process of translation the, mRNA is decoded to produce a specific polypeptide according to the rules specified by the triplet-nucleotide or codon by forming a sort of complex with ribosome. A different class of RNA called as transfer RNA (tRNA) reads the arrangement of codons and adds amino acids accordingly to form a protein. Therefore each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. In prokaryotic cells, which have no nuclear compartment, the process of transcription and translation may be linked together. But in eukaryotic cells, transcription occurs inside the cell nucleus, and the mRNA has to be transported out of the nucleus into the cytoplasm, where ribosomes can bind to it.

**Figure 1.4:** The process of transcription and translation in prokaryotic (a) and eukaryotic cell (b).

Proteins are large organic compounds made of linear chain of amino acids. Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential parts of organisms and participate in most of the process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle. Some proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. At residual level, two adjacent amino acids in a protein are joined together by peptide bonds between the carboxyl and amino groups. All amino acids possess common structural features that include a central α carbon to which an amino group, a carboxyl group, and a variable side chain are attached (Figure 1.5). Only proline differs from this basic structure as it contains an unusual ring to the N-end amine group, which forces the CO–NH amide moiety into a fixed

**Figure 1.5:** The general formula of an amino acid.
conformation (Nelson and Cox, 2000). The side chains of the amino acids have different chemical properties that produce three-dimensional protein structure and are therefore critical to protein function. For the sake of clarity protein structures are divided under four hierarchal levels.

- **Primary structure**: The linear arrangement of amino acid sequence is called as the primary structure of a protein.

- **Secondary structure**: Secondary structures are regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the alpha helix and beta sheet (Branden and Tooze, 1999). Because secondary structures are local, many regions of different secondary structure can be present in the same protein molecule.

- **Tertiary structure**: The spatial relationship of the different secondary structures which impart the overall shape of a protein molecule is called as tertiary structure. It is generally stabilized by non-local interactions, like hydrophobic core, salt bridges, hydrogen bonds, disulfide bonds, and sometimes by post-translational modifications. The term "tertiary structure" is often used as synonymous with the term *fold*.

- **Quaternary structure**: The shape or structure that results from the interaction of more than one protein molecule, usually called *protein subunits* in this context. Each subunit functions as part of the larger assembly or protein complex.

There has been a continuous advancement in sequencing technologies since the last decade, which has resulted in exponential growth in sequence databases. The first whole genome was sequenced in 1995 (Fleischmann et al., 1995), since then number of genomes has been sequenced. Besides advancement in sequencing technologies, the current progress in metagenomic projects (Schloss and Handelsman, 2005) is also contributing in populating the sequence database. This exponentially growing list of sequences poses a major challenge in front of bioinformaticians in order to extract some useful information from the four letters of nucleotides or 20 amino acids of protein chains, which makes no sense in raw form. Here, the first level of challenges is to predict/assign co-ordinate of protein coding regions (genes) and repeats in nucleotide sequences. In past a number of methods have been developed to predict
genes, both for prokaryotes (Delcher et al., 1999; Lukashin and Borodovsky, 1998) and eukaryotes (Issac and Raghava, 2004) and repeats (Sharma et al., 2004). Second level of challenge is to predict function of these gene products or proteins. However, between the two levels of challenges, first level still seems to be comparatively easy. The databases are flooded with tens of thousands of protein sequences of unknown function due to complexity involved in function assignment. The gap between protein of known sequence and known function is increasing with exponential rates over the years due to advancement in sequence technology and complexity involved in function assignment.

Protein function itself is a very complex phenomenon that is associated with many mutually overlapping levels: biochemical, cellular, organism-mediated, developmental and physiological, all intertwined intricately. There are several definitions of protein function ranging from a very general "a capability that a gene product carries as a potential" (Rison et al., 2000) to more complex and restrictive definitions based on knowledge representation (Karp, 2000). There are several methods to represent protein function which include: free text descriptions such as those contained in SWISS-PROT database; assignment of terms from a hierarchically arranged controlled vocabulary like Gene Ontology term (http://www.geneontology.org); description in terms of its interaction with, or relations to other molecules (protein-protein, protein-DNA, protein-RNA) (Riley, 1998). Each level of experimental protein function determination is a laborious task that can take enormous resources. Hence, automatic elucidation of protein function emerges as a major research area of Bioinformatics.

1.1. Functional Assignment Techniques

Large-scale genome sequencing projects are continuously discovering new genes and proteins. The ultimate goal of all sequencing projects is not only to find genes but also to describe the function of each resulting proteins. In this genomic era the task of finding all level of functions (molecular, biological and physiological) is too time consuming and costly. Hence the gap between proteins with known sequence and function and proteins with known sequence but unknown function is increasing continuously. Moreover, it has been observed that in Swiss-Prot, which is a manually annotated protein sequence database (Bairoch and Apweiler, 2000), experimental
functions have been determined for only a small fraction of proteins. In order to bridge the gap over the past years, many computational methods have been developed. Following is a brief description of bioinformatics approaches commonly used for function prediction.

1.1.1. Similarity search

One of the most powerful techniques, commonly used for predicting function of a newly found sequence is similarity search. In similarity search, query protein is compared with target protein sequences database using alignment techniques such as BLAST (Altschul et al., 1990) and FASTA (Pearson, 1990). If query protein has sequence identity more than a particular threshold with any experimentally annotated protein, then function of query and experimentally annotated protein may be similar. In other words functional assignment is done on basis of homology/similarity of query proteins to proteins with known function. But conclusions drawn from BLAST and FASTA searches are applicable only when query and target proteins share at least a minimum level of sequence identity. Remote homologous sequence can be searched using position specific iterated BLAST, popularly known as PSI-BLAST (Altschul et al., 1997). Development of PSI-BLAST has pushed down the level of sequence identity further down to infer the homology between two sequences.

1.1.2. Subcellular localization of proteins

The cellular localization of a protein is one of the most fundamental properties of any protein due to cellular division of labor. The correct prediction of subcellular location can be a major breakthrough for functional prediction, since to perform a function, proteins must be located in their native location, such as nucleus or mitochondria or outside the cell in case of secretory protein. So some insight can be gained about the basic function of a protein if its location can be predicted correctly.

1.1.3. Level of expression

Microarray experiments are typically the measurement of expression level of thousand of genes in one go. It can give information regarding which genes are up or down regulated at a particular condition. On the basis of increase or decrease in expression, genes are clustered into groups. The complete information is equivalent to a sort of signature or patterns or profiles. By observing these profiles several type of
information, like effect of a perturbation (chemical/environmental) on the cell, can be obtained (Claverie, 1999). It has been shown in past that proteins can be classified using gene expression data. Recently, it has been shown that combination of gene expression and amino acid composition can improve the function classification of proteins (Raghava and Han, 2005).

1.1.4. Protein structure prediction

It is a widely accepted fact that function of a protein is decided by the three dimensional structure of proteins. Thus prediction of tertiary structure of protein is important to understand the function of a protein. The experimental techniques (X-ray crystallography, NMR) of protein structure determination have their own limitations and unable to fill the gap between known sequences and known structure. In order to overcome this limitation attempts have been made to predict protein structure from its amino acid sequence using bioinformatics approach (Baker and Sali, 2001; Jones, 2001). Since protein structure is more conserved than sequence, some methods use structural information to predict function (Bartlett et al., 2003). Attempt has also been made in past to predict function of proteins from predicted structure of proteins.

1.1.5. Motifs, Patterns or Signals

It has been frequently observed that protein having similar function contain some conserved signal sequence or a well-defined pattern, which are responsible for their function. In some cases the sequence of an unknown protein is too diverged to detect its resemblance by overall sequence alignment, but the presence of conserved signal sequence or a well-defined particular cluster of residue types, known as a pattern, motif, signature, or fingerprint helps a lot in protein classification. These motifs arise because of particular requirements on the structure of specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity. These requirements impose very tight constraints on the evolution of those limited (in size) but important portion(s) of a protein sequence (Lesk, 1988). In other words, conserved pattern or signal sequence (e.g. N-terminal signals in secretory proteins) play important role in predicting protein function. A number of databases are available, which lists these motifs such as PROSITE (Hulo et al., 2004), PRINTS (Attwood et al., 2000), BLOCKS (Henikoff et al., 2000). Similarly, there exist intrinsic signal sequences, which guide the nascent protein to their location.
These sequences can be present either at N-termini or C-termini or at both ends. In most cases, the targeting sequence is removed during or after the transportation of protein to their destination by proteolysis (Arretz et al., 1991).

1.1.6. HMM profiles

This is an alternative approach to the motif-based characterization of protein family. The advantage of profile over motif is the fact that profile provides a more sensitive way of detecting distant sequence relationships in instances where very few residues are conserved. It relies on the fact that the variable regions between conserved motifs also contain valuable information. In this approach beside the complete conserved portion of the alignment, gaps are also used as discriminator, which is together termed as profile. In other words, profile defines which residues are allowed at a given position, which positions are highly conserved and which degenerate, and which positions or regions can tolerate insertions (Attwood, 2000). The popular method of homology search, hidden Markov Model (HMM) is also an extension of the concept of profiles (Hughey and Krogh, 1996).

1.2. Limitation of existing function assignment techniques

From the above points it is clear that there is no perfect bioinformatics tool which can be used to predict function of a protein. Though our understanding about protein function has improved over the years it is still far from satisfactory. Thus there is need to develop novel and better in silico tools for estimating function of a protein from its amino acid sequence.

1.3. Objectives of the thesis

The main objective of this thesis is to develop bioinformatics tools, which can be used to predict function of a protein. A protein may have different function in different environment, tissue and context. Thus prediction of function of a protein is not only difficult but complex too. It is beyond the scope of a thesis to work on all types of function of a protein. Thus attempts have been made to develop prediction methods only for important and major functions of a protein. First objective of thesis is to predict function at protein level, where attempt have been made to predict/classify important class of protein. This includes prediction of mitochondrial, nuclear and nucleotide (DNA & RNA) binding proteins. Second objective of thesis is to predict...
function at residue level, where attempt have been made to predict important binding residues in a protein. This includes prediction of DNA and RNA interacting residues in nucleotide binding protein. Following objectives has been decided for developing any method in this study

A. Benchmarking of existing methods;
B. Develop better method than existing methods;
C. Integration of two or more approaches;
D. Annotation of proteomes using newly developed method;
E. Launching web server.

1.4. Organization of the thesis

The central theme of this thesis is to develop prediction methods that can be used to predict function of newly discovered proteins with known amino acid sequence but unknown function. The whole thesis has been organized in four sections and eight chapters. Section I contains three chapters including introduction to thesis (this chapter). Chapter 2, "Review of Literature" describes existing bioinformatics tools used to address function annotation of proteomes. The complete methodology of these tools and their present status is briefly discussed in this chapter. In Chapter 3 (Materials and Methods) different computational techniques and performance measures used in present work are discussed. Section II describes methods developed for predicting function at protein level. It contains following three chapters i) chapter 4, describes composition based similarity search approach and its applications in biology; ii) chapter 5, describes methods developed for predicting major class of proteins (mitochondrial and nuclear proteins) and iii) in chapter 6, methods developed for predicting nucleotide binding proteins have been described. The methods developed for predicting function at residue level has been described in section III (Chapter 7 & 8). Methods have been developed for predicting RNA and DNA interacting residues in a protein (Chapter 7). In chapter 8, method has been developed for predicting beta-hairpins (or residues involved forming beta-hairpins). Last section of thesis deals with the summary and future prospects of the work.