CHAPTER 1

PROTEIN FUNCTION PREDICTION

1.1 INTRODUCTION TO PROTEIN FUNCTION

The word “protein” comes from the Greek term prôtos which means “primary” or “first rank of importance.” They are the basis of life and they regulate all activities involving the cell including genetic code replication to transport of oxygen and regulate the machinery of the cell and consequently an organism’s phenotype. Their task is accomplished through three-dimensional tertiary as well as quaternary coordination among substrates that include DNA, RNA, and other proteins.

The Amino acids are the building blocks of proteins. Amino acids form peptide by the polymerization at the carboxylic group of one amino acid to the amino group of the next (Lau, 2005). A protein is a long polypeptide chain. The protein is given a distinct function based on the chemical properties of every amino acid and the distinct sequence of its peptide chain. Removal of an amino acid or converting it from the protein sequence can damage the structure as well as its biological meaning.

Depending on the number of the amino acids in a chain, they are named as di, tri, tetra while a small chain of peptide that has four to eight amino acids is called oligopeptide and those with nine to forty amino acids are known just peptides. If the chain has more than 40 amino acids, then it is called polypeptide.

The classification of proteins is a very difficult task due to the uniqueness of the sequence of amino acid, that is the order in which it is linked.
to the composition, i.e., the number and kind of the amino acid present in a protein, the three-dimensional folding pattern (conformation) (Khan & Saxena, 2006). Although there are some overlapping characters for all features, their functions and physiochemical properties are classified differently.

The functions of proteins are mostly dependent on its structures; certain molecules like DNA perform a fairly small set of function though highly important, have a fixed structure which does not depend on its sequence. Whereas, certain protein molecules perform specific functions such as digesting sugars or moving muscles have a different structure; thus, depending on the function performed, their structure also differs. Thus, the function of protein depends on its structure.

Enzymes are forms of protein molecules that are present in cells, which hasten the chemical reaction, but are not used up in the process, that is they work as catalysts. In a given three-dimensional space, enzymes should recognize and react with their substrate with positioning precisely the critical chemical groups. Many scaffold protein signal and precisely dock other proteins or components and position them in the correct pattern. Certain proteins such as Collagens are prone to mechanical stress and should build a proper matrix where cells adhere and proliferate. In a precise fashion, motor proteins should convert chemical energy into movement.

Enzymes are the most specialized proteins because of its catalyst activity. Almost all the biological reactions in the body have enzymes as catalyst. Biochemical reactions cannot be performed in structural proteins due to their inertness. The position and native form of the organs are maintained by structural proteins, for instance, collagen is the primary fibrous constituent of the cell wall with a high tensile strength, which is abundant. It is present in
connective tissues including cartilage, tendons, matrix of bones and cornea of eye.

The most pure form of collagen is leather. Elastin is a structural protein present in ligaments which can stretch in two dimensions, α keratin consists of almost the entire dry weight of hair, wool, feathers, nails, claws, quills, scales, horns, hooves, tortoise shell as well as most of the outer layer of skin. Fibroin is the chief component of silk fibres as well as spider web. Resilin is the protein that is present in the wing hinges of some insects which has perfect elastic properties.

The prediction of protein structural class from a pattern recognition perspective is a multi-class classification problem. In the past 30 years, much effort was taken to solve this problem. Generally, these approaches are made of two important steps: (1) protein sequence representation or feature extraction; (2) algorithm selection for classification.

The protein structural class is predicted through many classification techniques such as neural network, Support Vector Machine (SVM), fuzzy k-nearest neighbor, fuzzy clustering, Bayesian classification, Logistic regression, rough sets, and ensembles of classifiers (Liu et al. 2015).

SVM has reached the best prediction performance for this task among these techniques; simultaneously, more discriminatory information for protein structural class are revealed through a wide range of sequence features that include Amino Acid Composition (AAC), pseudo-AAC, Position-Specific Score Matrix (PSSM) profile as well as predicted secondary structure. Pseudo-AAC is extensively applied to the field of bioinformatics as a potential feature extraction tool to analyze DNA protein sequences.
Protein Classification

Sequence similarity is the most common approach in classification of protein, which is a well-researched topic with many relevant algorithms and software packages. These are the largely utilized software packages (BLAST) in biology and bioinformatics. Two sequences are taken as input in the case of sequence similarity algorithms and software and a measure of distance is provided or similarity among them. High similarity and small distances are obviously synonymous.

While calculating local or global similarity algorithms, equal cost to all possible steps is not assigned. Scoring matrices are utilized moreover to give varied weight to replace various pairs of characters. Scoring matrices that are commonly used are BLOSUM, as well as the older PAM. The most common protocol and software package utilized to calculate global as well as local similarity is BLAST with a variant which is known PSI-BLAST. Other beneficial algorithms are FASTA and the Smith-Waterman dynamic programming algorithm (Sasson 2005).

Local alignments are searched through Smith-Waterman method, instead of considering the entire length of every sequence, comparison takes between substrings of all possible lengths. Dynamic programming is the basis for Smith-Waterman algorithm. In this algorithmic method, a solution for a problem is obtained through caching solutions of sub problems and using it in later stages of computation.

Local alignment score is provided through the approximation in FASTA which is a heuristic technique. The following observation forms its basis – through the identities in sequences good local alignments are stemmed from. A lookup table is constructed through the FASTA algorithm in which all instances of k-tuples of amino acids that appear the sequence are stored. All
instances of pairs of proteins are stored by the lookup table where the typical value of k used is k=2.

An additional heuristic technique for local alignment is BLAST. It looks for similar k-tuples, instead of identical k-tuples in sequences. It looks for k-tuples (the typical value used for k in protein comparison is 3) in one sequence that score at least T when aligned with the other sequence, again using a scoring matrix. In both directions, such local similarities are extended so as to find locally optimal continuous alignments with a score of S. Such alignments are named as High-Scoring Pairs (HSPs) and can provide best local alignments of two strings. In the tunable parameters are the neighbourhood threshold T and score threshold S (Sasson & Linial, 2004).

Primary Structure

Linear sequence of amino acids residues in a polypeptide chain. The amino acids are connected through peptide bonds on each side of the Cα carbon atom (Kornelia Polok).

Secondary Structure

The arrangement of primary amino acid sequence into motifs including α helices, β-sheets and coils are referred to in secondary structure. This can be determined by the side chains in amino acids. 4-10 amino acids from adjacent on neighboring stretches form the α helices. The β-sheets are created from adjacent β strands composed of 5-10 residues. They are arranged in parallel or antiparallel orientation.

Tertiary structure (3D)

As the name indicates, it is a three-dimensional arrangement formed through the packing of secondary elements into globular domains. The
structure depends upon certain post-translational modifications including addition of sugars and disulfide bridges.

**Quaternary structure**

This involves the formation of several polypeptide chains using tertiary structures. Functionally vital areas such as ligand-binding sites or enzymatically active sites are formed through tertiary or quaternary structures.

Approaches to Structure Prediction

Protein structure prediction includes two major categories:

(i) Prediction of secondary structure: In this attempt are made to locate the polypeptide chain that adopts the α-helical or β-strand structure. Regions which lack these secondary structural elements adopt coil conformation.

Tertiary structure (3D) prediction: tries to predict the 3D structure or native structure of the protein are made (Selvaraj, 2014). Though this has been an elusive goal so far, various techniques have been put forward to attain the goal.

**Table 1.1 List of some secondary structure prediction servers**

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHD</td>
<td><a href="http://cubic.bioc.columbia.edu/predictprotein">http://cubic.bioc.columbia.edu/predictprotein</a></td>
</tr>
<tr>
<td>PSI-PRED</td>
<td><a href="http://insulin.brunel.ac.uk/psipred">http://insulin.brunel.ac.uk/psipred</a></td>
</tr>
<tr>
<td>JPRED</td>
<td><a href="http://jura.ebi.ac.uk:8888">http://jura.ebi.ac.uk:8888</a></td>
</tr>
<tr>
<td>NPSA</td>
<td><a href="http://npsa-pbil.ibcp.fr">http://npsa-pbil.ibcp.fr</a></td>
</tr>
</tbody>
</table>
**Protein function prediction**

**Step 1** – Family model libraries are generated.

In a superfamily, for every family identified, creation of an alignment utilizing MAFFT with high-quality settings are done (‘–amino –localpair –maxiterate 1000’) and on its basis, using HMMER3 hmm_build command a profile HMM is built with default settings.

**Step 2** - Associating families with functions

Protein functions are associated probabilistically in every family. For every particular GO term that is related to the family (through one or more of the parent proteins that are corresponding), the frequency with individual sequences are counted. The number that results ranges from 1 to the number of domains in the family N. Secondly, all the counts are up-propagated in the GO DAG where the coarser terms are favoured by more specific child terms. Ultimately, association of the family takes with all terms which received counts in the above procedure with a particular term-specific probability value p. This is simply derived as the normalised term occurrence count or annotation frequency: \( p = \frac{T}{N} \), ranging from 0 to 1.

**Step 3** - Family and function assignment

CATH domains are designated to the target proteins, with the help of a standard protocol which generates the Gene3D resource. Scanning against the superfamily’s family model library are identified for domain consequences in each superfamily employing the HMMER3 hmm_scan command with default options (Rentzsch & Orengo 2013).
1.2 GENE ONTOLOGY

Gene Ontology (GO) is a primary tool for representing and processing information about gene products as well as functions. A variety of biological databases are developed through GO in tandem within the umbrella project Open Biological Ontologies (OBO). A controlled description of cellular components, molecular functions as well as biological processes is provided (Smith et al. 2003).

Biological process is the biological objective to which the gene or gene products are contributed. One or more ordered assemblies are accomplished via a process. Molecular function is the biochemical activity which includes specific binding to ligands or structures of a gene product (Ashburner et al. 2000). This applies to the capability that a gene product or gene product complex carries as potential.

Cellular component in the cell is where a gene product is active based on the eukaryotic cell structure; though not all terms are applicable to the organisms, the set of terms are inclusive. Terms such as ‘ribosome or proteosome’ specify multiple gene products and there are terms such as ‘nuclear membranes’ or ‘golgi apparatus.’

A collaborative effort in constructing and using ontology in facilitating the biologically meaningful annotation of genes is the GO project and their products in a wide variety of organisms. The groups which participate in the project include major model organism bases and other resource center of bioinformatics. To mine functional as well as biological significance, GO annotations have proven to be largely beneficial from very large datasets which includes microarray results (Gene Ontology Consortium 2006). The organization of data is facilitated through GO from a novel and
fully annotated genomes and the comparison of biological information between clade members and across clades.

Two aspects of information integration are addressed in GO project: provision of consistent descriptors for gene products in case of distinct databases, and the standardization of classification for sequences as well as sequence features. There are three major goals in the GO project: (i) a set of structured, controlled vocabularies known as ontologies are developed so as to illustrate important domains of molecular biology, which includes gene product attributes as well as biological sequences; (ii) application of GO terms in the annotation of sequences, genes or products of genes in biological databases; and (iii) provision of centralized public resource allowing universal access to ontologies, annotation data sets and software tools developed for use with GO data (Gene Ontology Consortium 2004).

The simplest data structures sufficient for initial task is chosen by GO forming a directed acyclic graph; the structure of GO is detailed below. Three attributes are given importance in annotating GO: the particular molecular functions possessed by these products, participation in higher level biological processes, and components of cell in which they can be determined. Presently GO has been used to annotate more than one million gene products within the variety of participating databases (Bada et al. 2004).

**Characteristics of GO**

- Community Involvement
- Clear Goals
- Limited Scope
- Simple, Intuitive Structure
- Continuous Evolution
- Active Curation
Early Use

A gene is associated with a GO annotation with effect to ontologies and is created automatically or by a curator through methods of prediction. Genes associates themselves with many appropriate terms and with most specific terms that are available in order to reflect whatever is identified about the gene. When a gene gets annotated to a term, then the association among the gene and the parents of the terms are inferred inevitably. As GO inherits all the characteristics of the ancestors of those terms, each and every path from all terms back to its base be accurate biologically or the ontology revised.

There is continuous ontology and annotations updated so as to reflect current knowledge, to reduce errors as well as to enhance logical consistency. The GO ontology is updated daily while the annotation files are released weekly. GO also helps in inferring the function of the gene computationally (Rhee et al. 2008). The same theme is slightly varied in case of typical approaches: On the basis of certain criteria, genes are grouped together such as similar gene expression or a network with protein-protein interaction. Detection of enrichment of GO terms is done through ways that are aforementioned and presumption of characterized genes which could be involved in similar biological procedure as the genes with which they are grouped together. Thus, these genes that are uncharacterized can be associated presumptively with enriched GO terms.

Semantic similarities can be quantified through several similarities within terms or entities that are annotated entity in the study which is represented as DAG such as GO. The below are the two types of methods used to compare terms in graph-structured ontology:
- Edge-based – edges and their types are used as data source;
- Node-based – nodes and their properties are the main data sources

The basis for edge-based technique is primarily, counting the number of edges in a graph path between two terms (Ivanoska et al. 2014).

Distance is the most common method that chooses either the shortest path or the average of all paths, when there is one or more path in existence. A metric of distance between two terms is given through this technique which can get converted into a similarity metric. As this approach is considered intuitive, the specificity is not same for terms at the same depth and edges which are present at the same level need not denote the same semantic distance.

The basis for node-based technique is on the comparison of properties of the involved terms that can relate themselves to the terms, their ancestors, or their descendants. Information Content (IC) is the most extensively employed concept to these approaches and this brings forth a measure as to how particular and informative a term is.

1.3 CLUSTERING-BASED PROTEIN FUNCTION PREDICTION

The protein is grouped into sets in clustering (clusters) which demonstrate similarities between proteins in the same cluster than in varied clusters. As biological functions can take place through certain groups of genes as well as proteins, division of networks by naturally grouped parts (clusters or communities) is a crucial way in investigating certain relationships among function as well as topology of networks or to bring out the hidden knowledge behind them (Trivodaliev et al. 2011).
A similarity of computing is required in clustering sequences of protein. Based on the measure of similarity two approaches are required for measure of similarity. The first has its basis on the alignment of sequence. Alignment algorithms such as BLAST or FASTA help in measuring the similarity between two sequences of proteins.

While good solutions are achieved through alignment of sequence, it is comparatively difficult to assemble a higher number of sequences due its computational complexity. Also, if there is a disparity in the length of sequences, it is difficult to align satisfactorily which results in reduced clustering accuracy. Methods without alignments in the second approach to similarity measure (Abdel-Azim, 2016).

Proteins can be clustered through several approaches. One of the simplest clustering techniques is the Single Linkage (SL) method and is used widely for clustering sequence of protein. Though the technique is simple, it is fast and largely employed, the disadvantage being the difficulty in the detection of an appropriate cut-off score for similarity of sequence. There is variation in terms of length of signature and residue conservation while sharing a signature. Sometimes, they might not share a signature, but the proteins might turn out to be similar by chance.

A rigid or a set of cut off scores are used in the current protein clustering application with distinct levels. Proteins are classified at different levels and are selected automatically by a SL-based clustering technique by name SL-KL. A graphic illustration is used to represent a given set of proteins in which the vertices denote protein and edges denote pairs of proteins with sequence similarities above a particular cutoff score (Kawaji et al. 2004).
A variety of similarity measures are employed through classical graph-based agglomerative methods between nodes in partitioning PPI networks, but generally they result in a poor clustering arrangement that consists of only a few giant core clusters with multiple tiny ones. To enhance results of clustering, based on topological properties PPI networks were weighted like shortest path length, coefficients of clustering, degree of node or the degree of experimental validity. The novel clustering algorithm, the edge-betweenness is expressed as a global measure in separating PPI networks into sub-graph in a divisive manner.

The shortest paths that run through the edges are called as edge-betweenness. Significant modular structures are identified but lots of computational resources are required. An ensemble method is suggested to join multiple, independent clustering arrangement so as to coordinate typical clustering algorithms and to deduce consensus cluster structure.

Clustering of network $G(V, E)$ is decomposing a set of nodes $V$ into various highly interconnected subsets of nodes. A cost-based local search algorithm is RNSC algorithm which is based loosely on tabu search metaheuristic. In this algorithm’s context, a clustering of a network $G=(V, E)$ in partitioning of the node set $V$. The space of partitions of $V$ is searched efficiently by the search spaces, which is assigned a cost for a clustering with low cost.

A simple integer-valued cost function is searched by the algorithm (called the naive cost function) as a preprocessor, before a more expressive (but less efficient) real valued cost function (scaled cost function) is searched. The initial clustering is random or user-input (King et al. 2004). A low-cost clustering is searched by RNSC by composing an initial random clustering first, then repeatedly moving one node from cluster to another in a random
fashion, in improving the clustering cost. In a general move, the cost is reduced by a near optimal amount.

### 1.4 DOMAINS FOR PROTEIN FUNCTION PREDICTION

Proteins are made up of evolutionarily stable units called domains. Two types of domains are available:

- **Functional domains** – these are defined by Pfam and SMART;
- **Structural domains** – these are semi-autonomous, compact folding units with separate hydrophobic cores.

Both structural as well as functional domains may overlap and are valid though there are agreements rarely on perfect boundaries on the domain.

The basis for protein domains are as follows:

**Geometry:** group of residues with a high contact density, number of contacts within domains is higher when compared to the number of contacts among domains

**Kinetics:** domain as an independently folding unit

**Physics:** domain as a rigid body connected to other domains by flexible linkers

**Genetics:** minimal fragment of gene that can perform a particular function

Structural domains have separate hydrophobic cores and have the characteristic of more widespread contacts between residues within than between domains. A manually curated domain classification database is
Partitioning of domains takes place through automated as well as manual analysis of structures of proteins (Ezkurdia & Tress, 2011).

If there is similarity between protein chain and a structure which is previously parsed into domains, a new domain is similarly divided. Otherwise after using a range of protein, domain parsing is evaluated by eye through a domain parsing software. Sequence, structural and/or functional similarity are used in the classification of domain into a homology-based hierarchical system base.

There are multiple domains for most proteins with effect of the modular recombination of domains all through protein evolution (‘domain shuffling’). This implies that there are several other pair-wise homology relationships between individual domain sequence than between whole-protein sequences. Exploitation of this can increase the sensitivity of Protein Function Prediction (ProFP) methods has, as demonstrated for a simple protocol with basis on pairwise sequence comparison (Rentzsch & Orengo 2013).

With its extension, any ProFP method should be improved based on the sequence of protein by performing these comparisons on domain level instead and the results should be integrated subsequently through the results obtained for all domains. One of the primary challenges is the association of protein families and domain functions.

A term T is assigned to the domain composition of all proteins in a matrix, which is followed by an iterative protocol. In the first step, the domain D with highest occurrence matrix is related to T, all instances of D are labelled in protein sets as T-associated. Not T-associated are the domains that co-occurring. In the next step, all proteins that contain D are eliminated. The protocol goes on until there is no protein left. For every domain D, the
likelihood of ‘encoding’ function $T$ is evaluated. This process goes, and the functions can be associated with multiple domain types and vice versa.

Valuable insights can be given through the association of gene products with structure, since molecular details are provided through structural information of the function of a protein. The target is also prioritized through the assignment of structural domain for genomics consortium by the indication of gene products with no structural predictions (Pandit et al. 2004).

It is vital that two distinct predictions are made through the domain fusion analysis. First, protein pairs with related biological functions are predicted, i.e., participation of proteins in usual structural complex, metabolic pathway or biological process (Marcotte et al. 1999). Functional prediction is robust – for E. Coli, general functional similarity was observer in over half the predictions that are testable. Secondly the method predicts protein-protein interaction.

1.5 COMMON CHALLENGES IN PROTEIN FUNCTION PREDICTION

There are certain challenges irrespective of the kind of biological information or the technique involved and approaches to automated function prediction.

Incomplete Data

Only partial information is provided by the biological data because of the limitation is experimental techniques and resources. There might not be constant time intervals for gene expression profiles from microarray experiments (Chua & Wong, 2009).
Lack of a Common Protein Naming Convention

Overlapping or complementary information are contained many biological databases on the same proteins. In case of prediction of function, it is crucial in obtaining information from many sources. From various reasons, the adoption of different naming conventions may stem, including legacy or the nature of the data being referenced (for instance, sequences vs. Genes). Nevertheless, certain issues are posed in protein function prediction when data needs to be combined from various sources.

Noisy Data

Some biological data including high-throughput protein interaction assays and gene expression function can be noisy (i.e., may possess several false positives). For consistent prediction performance, approaches that utilize such biological data must consider noise. For achieving consistent prediction performance; noise should be taken into consideration for approaches which make use of such biological data. False positives can be reduced in experimental datasets in case of protein-protein interactions through many computational techniques.

In order to handle the challenges of protein prediction function and enzyme correction function, mis-annotations within databases, a challenge is taken upon by the community. These involve cooperation between various groups, employing theory, computation and experiment and a significant progress is made towards protein function confirmation, thus a substantial value is added to the information of structural genomics proteins currently available.

Ten years after the start of PSI began the Enzyme Function Initiative (EFI), funded by NIGMS. Substrate specificity of proteins of unknown
functions is determined through the combination of bioinformatics with experimental enzymology. Many techniques have been developed by research groups in order to offer beneficial information about enzymes of unknown function. However, the probability of mis-annotation is higher when only one type of analysis, sequence- or structure-based, is used when making predictions (Mills et al. 2015). The information got through genome projects can be more useful and complete if methods continue to be optimized and used in parallel with various other methods.

1.6 PROTEIN INTERACTIONS IN FUNCTIONAL PREDICTION

The cells are managed by interaction of proteins in metabolic as well as signalling pathway and in complex such as molecular machines which synthesize the use of Adenosine TriPhosphate (ATP) which replicate and translate genes or building up of cytoskeletal infrastructure.

Through various specific binding sites, proteins associate with each other. These Protein-Protein Interaction Sites (PPISs) are good contributors in recognizing binding residues under specified chemical and physical statuses. As the central position of interactions is marked through the PPISs, they are less effectively captured through experimental methods, and development of computational approaches to model the discrimination between interacting and non-interacting sites for prediction of PPIS (Chang et al. 2016). Structure and sequence-based evidence are done through PPI site prediction techniques and many studies have been proposed. Due to the rapid growth of structural information, PPI predictions using structural information have gained more attention.

Representation of protein-interaction network is done as an interaction graph with protein in vertices and interactions in edges. Multiple
topological features that include network diameter, distribution of vertex
degree, clustering coefficient etc., have been studied and these network
analyses show that protein interaction networks have features of a scale-free
network and ‘small-world effect.”

Proteins are grouped into sets for clustering in protein interaction
networks and these demonstrate greater similarity among proteins in the same
cluster than in different clusters. Two types of modules are represented in
protein interaction networks – protein complexes and functional modules
(Wang et al. 2010).

One of the reasons for missing interactions is that a number of
protein-protein interactions may have been done through other mechanisms,
including gradual accumulation of mutations evolving in a binding site.
Through domain fusion, false predictions of physical interactions maybe made
in cases where domains do not interact but fuse (Marcotte et al. 1999). This
can be because proteins are fused to regulate co-expression or protein
signaling.

One of the classification problems is the recognition of protein-
protein interaction sites where every amino acid residue is assigned to one of
the two classes – interacting or non-interacting residues. There might be a
solution to the problem when statistical and machine learning techniques such
as NN or SVM are used.

Two distinct methods are generally employed to define an
interaction site based on 3D structural data: (i) interatomic distance and
(ii) change in Accessible Surface Area (ASA) upon complex formation. In the
first method, interaction sites can be defined based on the distance between
non-hydrogen atoms of different protein chains. The second approach defines
an interaction interface by utilizing the idea of solvent accessibility or ASA (Cai & Hong 2012).

ASA or the solvent accessibility of an amino acid residue in an unbound protein chain is contrasted with the corresponding ASA value for the same residue in a complex. Residues with considerable difference in ASA among the isolated chain and complex structures are then classified as “interacting”. Both methods need high resolution structural data. However, the inter-atomic distance-based method is more likely to be more sensitive to problems with missing atoms or atoms with several occupancies.

Helpful information is present in protein interaction networks for understanding the role of proteins in cells and function prediction for annotated proteins.

Through computation, protein-protein interaction networks can be modelled as undirected graphs, where nodes are proteins and edges represent physical binding interactions. Many edges are missed in this graph and many incorrect edges or false positives are present. For complimenting and extending experimental techniques, a number of computational techniques have been applied successfully in predicting protein interactions. These methods can be categorized based on the type of data considered by them while making predictions, as follows (Qi & Noble 2011):

- To infer PPIs, over-represented domain pairs or motif pairs are observed to interact with protein pairs. Potential PPIs can be predicted through structural information and sequence evidence about PPI interfaces.

- Several genomic methods infer protein interactions based on the conservation of gene neighborhood, conservation of gene order,
gene fusion events, or the co-evolution of interacting protein pair sequences.

- Integrating various types of approach is an attractive alternative to several kinds of evidences from various sources in a statistical learning framework.
- In recent years, to model organism, high-throughput PPI experiments for elucidating protein-protein interactions have been applied.

1.7 PROBLEM STATEMENT

In data mining, prediction of protein functions is considered as a classification problem where the attributes of proteins are taken as a sample and its biological attributes as classes. The major issue in protein classification is the size of the feature space or curse of dimensionality. In Hierarchical Multi-Label Classification (HMC) problems, the X instances are processed to map every instances xi to a set of classes, where the constraints in the hierarchical structures and quality criterion are optimized. Non-deterministic Polynomial (NP)-hard problems computationally costly and time complexity. Feasible solutions are obtained through optimization techniques, though there is no certainty that an optimal solution can be obtained for such problems. As there are no efficient algorithms for NP problems, most of them are solved using trial and error method using different optimization methods. Also, for finding out whether these challenging optimization problems can be tackled, several innovative algorithms have been developed. Among these algorithms which are popular for their efficiency are Particle Swarm Optimization (PSO), Cuckoo Search (CS) And Firefly Algorithm (FA). In this work, hybrid optimized with Group Search Optimization (GSO) and Genetic algorithm (GA) are proposed.
1.8 OBJECTIVE OF THE RESEARCH

A protein is defined through the various forms of collected data in forms like its sequence of linear chain of amino-acids, three-dimensionally structure, various interactions that is protein-protein interactions and co-expression of the gene. It is computationally expensive to find the function of a protein using experimental approaches. The main objectives of the proposed work are listed below:

- To propose a multi-label protein function prediction based on hybridization of k-nearest neighbor-based EM algorithm (KNN+EM).
- To propose Group Search Optimizer (GSO) with Hierarchical Multi-label Classification withProbabilistic Clustering
- To propose Hybrid GSO with Hierarchical Multi-label Classification with Probabilistic Clustering to improve the performance of classification

1.9 ORGANIZATION OF THE THESIS

Chapter 1 discussed the basic information about the protein, function prediction, Gene Ontology and clustering in function prediction etc.

Chapter 2 reviews the related work to the protein function prediction and hierarchical classification etc.

Chapter 3 explains the multi-label protein function prediction based on hybridization of k-nearest neighbor based EM algorithm (KNN+EM).
Chapter 4 proposes the Group Search Optimizer (GSO) with Hierarchical Multi-label Classification with Probabilistic Clustering for protein function.

Chapter 5 uses the Hybrid GSO with Hierarchical Multi-label Classification with Probabilistic Clustering to improve the performance of classification of protein function.

Chapter 6 concludes the proposed work and gives the future scope for the protein function prediction.