1. Introduction
1. Introduction

Computational biology is an emerging and rapidly growing field of genomics and proteomics. As a consequence of the large amount of data produced in the field of molecular biology, most of the current projects relating to bioinformatics deal with structural and functional aspects of genes and proteins. The data produced by a number of research teams all over the world are collected and organized in databases specialized for relevant subjects. The existence of public databases with billions of data entries require a robust analytical approach to cataloging and representing this with respect to their biological significance. Therefore, computational tools are needed to analyze the collected data in the most efficient manner. For example, working on the prediction of the biological functions of genes and proteins based on their sequence data. The support vector machines (SVM) approach is a promising technique for machine learning and it has been applied to many biological problems. Here, we explore the development of Bioinformatics tools for predicting oxygen-binding proteins including bacterial hemoglobin proteins and Uncharacterized (Unannotated) protein sequences.

1.1 Overview of oxygen-binding proteins

Oxygen is an important part of the atmosphere and is necessary to sustain most terrestrial life of living organisms as it is used in respiration. Oxygen is required for the regulation of a variety of cellular functions. The appearance of oxygen on the earth’s atmosphere is nearly 1 to 2 billion years ago. Oxygen is vital for all aerobic life processes. With the arrival of free oxygen on the globe, aerobic life becomes possible, but initially earth’s milieu was so oxidizing that, free molecular oxygen would oxidize and thus destroy all organic molecules required for the origin of life. Thus appear that, it cannot sustain aerobic life directly. It is required for the regulation of a variety of cellular functions.

Early in the history of earth, there was essentially no free oxygen anywhere, although Oxygen has always been one of the most abundant elements on earth. When oxygen gets accumulated in earth’s atmosphere, the introduction of oxygen into an anaerobic world brought problems for the then existing organisms, for many of the by products of oxygen metabolism are toxic substances. On the other hand, certain organisms evolved aerobic metabolic pathways, much more efficient than the aerobic ones. These were the ancestors of all animals and higher plants, capable of both respiration and fermentation. Even, the primitive microbes managed life without
free oxygen. Examples of this less efficient anaerobic metabolism still persist, such as bacteria that live in oxygen poor environments.

The challenge for primitive obligate and facultative aerobes was to sense, capture and store the oxygen for energy production without suffering from its hazardous side effects. The only chemical arsenal available to organisms engaged in oxygen based metabolism was perhaps the porphyrin ring. A porphyrin molecule is a planar group of four connected rings, each of which contains a nitrogen atom that faces the center of the ring cluster. These four nitrogen’s provide an ideal environment for the insertion of a metal ion, such as iron or magnesium, which is extremely useful for a variety of oxygen related reactions. This basic chemical apparatus grew increasingly complex through time and evolution and later on, porphyrin ring became embedded in larger organic compounds called proteins. These organic compounds themselves became increasingly varied through time and evolution. The descendents of these compounds include the chlorophyll and heme. Each class of compound still contains a porphyrin ring at the center so that the basic interaction between metal atoms and oxygen has not changed. Some of Oxy-proteins catalyze a variety of oxidations.

In humans, the protein molecule most closely associated with the utilization of oxygen is hemoglobin (Hb), the best characterized oxygen-binding protein both at the functional and molecular level, known as the oxygen carrier of the blood. This was the first protein to have its amino acid sequence deciphered and later on its three dimensional structure resolved by X-ray crystallography (Fermi et al., 1984) and also this was the first proteins to have its amino acid sequence determined. The presence of this oxygen-binding protein had long been thought to be restricted to mammals, but subsequent findings indicate an almost ubiquitous existence of Hbs in mammals, non-vertebrates, plants and bacteria (Wakabayashi et al., 1986; Bogusz et al., 1990; Burmester and Hankeln 1999; Roesner et al., 2005). Scientists have so closely associated hemoglobin with oxygen transport that they have been surprised to discover hemoglobin in organisms which have no obvious need for oxygen transport.

Hemoglobins are found in virtually all kingdoms of life and are assigned to vertebrate and non-vertebrate Hb classes. The latter group includes the Hbs present in plants, fungi and protozoa (Iwaasa et al., 1989; Iwaasa et al., 1990; Zhu and Riggs 1992; Cramm et al., 1994). The detailed analysis of protein structures of globins reveals a typical tertiary structure, the classical globin fold. This highly conserved structure consists of six to eight a helical segments that are connected
by short intervening loops. The three-on-three helix structure forms a sandwich like assembly and
binds the heme moiety surrounded by hydrophobic residues.

However, it has become very clear that the Oxy-proteins are widespread in the biosphere
and are found in all groups of organisms including prokaryotes, eukaryotes as well as fungi, plants
and animals. The oxygen-binding proteins of various organism considerably differ from one
another and are classified as hemoglobin, hemocyanin, hemerythrin, myoglobin, leghemoglobin
and erythrocruorin based on their structure and physiochemical properties.

1.1.1 Oxygen-binding proteins in Bacteria (hemoglobin)

The oxygen-binding proteins such as hemoglobin was believed to be a protein that evolved
in eukaryotes but the discovery of hemoglobin like protein have been found in many bacteria, the
gram negative bacterium *Vitreoscilla* (Wakabayashi *et al.*, 1986) brought down the age old
paradigm that hemoglobin is exclusive to eukaryotes. It was a common assumption that single
celled organisms don’t need respiratory surfaces because they have a high surface, and are small
enough to allow oxygen to diffuse through them and thus, avoid burden of auxiliary mechanisms.
The hemoglobin discovery in bacteria is strongly suggested an early prokaryotic origin of the
globin gene. Since then, the presence of Hb-like proteins has been reported in various microbial
systems indicating their widespread occurrence. Amongst archaea, only *Halobacterium salinarium*
is reported to have a globin-like protein, which acts as an aerotactic transducer (Hou *et al.*, 2000).
Some of these hemoglobins reported from single-celled organisms are bifunctional proteins with
their N-terminus as the globin domain and C-terminus performing various enzymatic functions
*e.g.* NADH reductase activity in *Saccharomyces cerevisae* (Keilin and Ryley 1953; Keilin and
Tissieres 1953) and *E. coli* (Vasudevan *et al.*, 1991), kinase activity in *Rhizobium melliloti* (Gilles-
Gonzalez *et al.*, 1991; Gilles-Gonzalez *et al.*, 1994) and cytochrome reductase in *Ralstonia
eutrophus* formerly called *Alkaligenes eutrophus* (Probst *et al.*, 1979). Hemoglobin-like proteins
have also been reported from diverse bacterial species *e.g.* *Mycobacterium tuberculosis* (Couture
*et al.*, 1994; Couture *et al.*, 1999); *Bacillus* (LaCelle *et al.*, 1996), *Vibrio* (P40609), *Erwinia*
(Favey *et al.*, 1995); *Bradyrhizobium* (Gilles-Gonzalez *et al.*, 1994) and *Chromatium* (Gaul *et al.*, 
1988). The number of yeast, bacteria and fungi are being added to the list which point towards a
possibility of ubiquitous occurrence of hemoglobins. The diversity of hemoglobins found in
almost all living forms has prompted the workers to explore and compare the structure, 
physiochemical properties and evolution of the various hemoglobins. The small size hemoproteins
having 20-40 amino acid residues sequences shorter than (non) vertebrate hemoglobins have been identified in several pathogenic and non-pathogenic unicellular organisms, called as ‘truncated hemoglobins’ (trHbs) (Pesce et al., 2000). They have been proposed to be involved not only in oxygen transport but also in other biological functions such as protection against reactive nitrogen species, photosynthesis or to act as terminal oxidases. Thus, microbial and lower eukaryotic Hbs can be conveniently put into three categories as Single domain hemoglobins, Chimeric hemoglobins and Truncated hemoglobins

1.1.1.1 Single domain Hemoglobin

The homologous hemoglobin of bacterium was first discovered in Vitreoscilla Sp. has 3-over-3 globin fold single domain hemoglobin (Wakabayashi et al., 1986; Frey et al., 2011). The gene encoding VHb (vgb) was cloned and expressed at higher levels from its native oxygen responsive promoter in E. coli (Dikshit and Webster 1988). At present, VHb is the best characterized member of the group of bacterial Hb proteins. A FAD-containing NADH-met reductase domain that co-purifies with Hb has been partially characterized in Vitreoscilla. Hence, a non-covalent assembly of these domains could yield a flavoHb whose function, although still unclear, could be related to that of other flavoHbs (Tarricone et al., 1997; Kaur et al., 2002).

1.1.1.2 Two-domain Hemoglobin (Flavohemoglobin)

The ensuing years brought to light several new globins; one of which is chimeric globins, comprising N-terminal globin domain and a C-terminal variable domain performing diverse functions. These chimeric globins comprise the ‘flavohemoglobins’ with a C-terminal FAD-binding domain and the gene regulating ‘globin coupled sensors (GCSs)’ with variable C-terminal domains. In the last decade, members of a family of 43 kDa two-domain (‘chimeric”) flavohemoproteins also called flavoHbs have been discovered in several microorganisms. They are composed of a classical globin domain covalently coupled to a flavoprotein reductase or a kinase domain (Takagi et al., 1993). These proteins appear to be widely distributed in bacteria and fungi. A two-domain flavoheme protein in E. coli also combines reversibly with oxygen and has a heme domain that is 46% identical to Vitreoscilla hemoglobin; the second domain binds a flavin. The crystal structure shows the similarity of the N-terminal domain of the globin family, whereas the C-terminal domain is structurally similar to the ferredoxin -NADP+ reductase family (Ermler et al., 1995).
1.1.3 Truncated hemoglobins

Truncated hemoglobins (trHbs) are small heme proteins have been identified in bacteria, unicellular eukaryotes and higher plants, forming a distinct group within the hemoglobin superfamily (Pesce et al., 2000; Wittenberg et al., 2002). The trHbs have been characterized so far in the ciliated protozoa Paramecium caudatum and Tetrahymena pyriformis, in the unicellular alga Chlamydomonas eugametos and in the eubacteria Nostoc commune, Mycobacterium tuberculosis and Mycobacterium leprae etc. (Iwaasa et al., 1989; Iwaasa et al., 1990; Takagi et al., 1993; Couture et al., 1994; Thorsteinsson et al., 1999; Visca et al., 2002; Fabozzi et al., 2006; Ascenzi 2007). In this regard, an analysis of the currently available microbial genome sequences indicates that certain bacteria (Bacillus subtilis, Staphylococcus aureus, Campylobacter jejuni, Bordetella pertussis, Deinococcus radiodurans) contain both flavohemoglobin and trHb, which suggests specific functions for each of the two protein classes (Wu et al., 2003).

1.1.2 Oxygen-binding proteins in cellular organisms

Oxygen-binding proteins are widely present in eukaryotes ranging from non-vertebrates to humans (Zhang et al., 2007). Moreover, these proteins have also been reported to be present in many prokaryotes and protozoans. The occurrence of oxygen-binding proteins in all kingdoms of organisms, though not in all organisms, shows their biological importance. Extensive studies on oxygen-binding proteins have categorized them into six different broad types, including erythrocruorin, hemerythrin, hemocyanin, hemoglobin, leghemoglobin, and myoglobin, each has its own functional characteristics and structure with unique oxygen-binding capacity (Fig 1.1). These oxygen-binding proteins are crucial for the survival of any living organism. Many Oxy-proteins exhibit different functions such as oxygen binding, electron transfer, and metabolism. Erythrocruorin is one of the respiratory proteins found in haemolymph of many species from the Phyla Annelids, Mollusca and Arthropoda, particularly present in certain marine polychaetes. Hemerythrin is a non-heme iron protein used by marine invertebrates for oxygen-binding and transport. Hemocyanin is also the oxygen-binding proteins in arthropods and mollusks. Two copper atoms bind the oxygen. Leghemoglobin found in the nitrogen-fixing root nodules of leguminous plants, its response to the roots being infected by the nitrogen-fixing bacterium like Rhizobium, as part of the symbiotic interaction between plant and bacterium. Myoglobin found in vertebrate skeletal muscle cells, it has the ability to store oxygen by binding it to the iron atom.
(Morse et al., 1986; Decker and Terwilliger 2000; O'Brien and Sidell 2000; Svistunenko 2005; French et al., 2008). Though several methods have been developed to predict function or subcellular localization of proteins, and there is a need of a method for predicting oxygen-binding proteins and their classification with their protein profile. During the past decades, a number of methods have been developed for predicting protein secondary structure, protein function prediction, protein protein interaction, protein domain. Many studies have appeared to understand the Oxy-proteins and their functions.

1.1.2.1 Myoglobin

Myoglobin is one of the oxygen-binding protein found in the muscle tissue in vertebrates, in general almost all animals (Ball and Cooper 1947; Biorck 1948; Rossi-Fanelli 1948). Diving mammals such as whales and seals have muscles with particularly high myoglobin abundance (Keilin and Schmid 1948; Keilin and Ryley 1953; Keilin and Tissieres 1953). Myoglobin, an iron-containing protein in muscle, receives oxygen from the red blood cells and transports it to the mitochondria of muscle cells, where the oxygen is used in cellular respiration to produce energy. Each myoglobin molecule has one heme prosthetic group located in the hydrophobic cleft in the protein. The function of myoglobin is notable from Millikan's review (Millikan 1939) in which he put together an accomplished study to establish that myoglobin is formed adaptively in tissues in response to oxygen needs and that myoglobin contributes to the oxygen supply of these tissues. It is a single chain globular proteins of 153 amino acids, containing a heme (iron-containing porphyrin) prosthetic group in the center around which the remaining apoprotein folds. It has eight alpha helices and a hydrophobic core. It has a molecular weight of 16,700 daltons, and is the primary oxygen-carrying pigment of muscle tissues. Myoglobin was the first protein whose structure was determined. In 1958, Max Perutz and John Kendrew determined the 3D structure of myoglobin by X-ray crystallography. Four years later, they both received the Nobel Prize in chemistry for this innovation.
Oxygen binding Proteins

<table>
<thead>
<tr>
<th>Hemoglobin</th>
<th>Myoglobin</th>
<th>Hemocyanin</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Hemoglobin" /></td>
<td><img src="image2" alt="Myoglobin" /></td>
<td><img src="image3" alt="Hemocyanin" /></td>
</tr>
<tr>
<td>Hemerythrin</td>
<td>Erythrocrurorin</td>
<td>Leghemoglobin</td>
</tr>
<tr>
<td><img src="image4" alt="Hemerythrin" /></td>
<td><img src="image5" alt="Erythrocrurorin" /></td>
<td><img src="image6" alt="Leghemoglobin" /></td>
</tr>
</tbody>
</table>

Figure 1.1 : Three dimesnional structure of various Oxygen-binding proteins (the images of given protein structures a reproduced from PDB)
1.1.2.2 Hemoglobin

Hemoglobin is an iron-containing Oxygen-binding proteins present in the blood of all animals, and the main function is to carry oxygen from the respiratory organs lungs to the rest of the body (tissues). Hemoglobin is present in red blood cells of vertebrates and gives red in color. It contains four peptide units, each unit attached to heme that binds to oxygen (Bragg 1953). Hemoglobin and hemoglobin-like molecules are also found in many invertebrates, fungi, and plants (Wald and Riggs 1951). In these organisms, hemoglobins may carry oxygen, or they may act to transport and regulate other things such as carbon dioxide, nitric oxide, hydrogen sulfide and sulfide. A variant of the molecule, called leghemoglobin, is used to scavenge oxygen away from anaerobic systems, such as the nitrogen-fixing nodules of leguminous plants, before the oxygen can poison the system.

1.1.2.3 Leghemoglobin

Leghemoglobin is an iron-containing protein, similar in a number of properties to the hemoglobin found in the nitrogen-fixing root nodules of leguminous plants. It is produced by legumes in response to the roots being infected by nitrogen-fixing bacteria, called rhizobia, as part of the symbiotic interaction between plant and bacterial: roots uninfected with Rhizobium do not synthesize leghemoglobin (Ellfolk and Sievers 1967; Cutting and Schulman 1971; Ewing and Ionescu 1972; Noel et al., 1982; Powell and Gannon 1988). Leghemoglobin has close chemical and structural similarities to hemoglobin, and, like hemoglobin, is red in color. The protein was believed to be a product of both plant and the bacterium in which the apoprotein is produced by the plant and the heme (an iron atom bound in a porphyrin ring) is produced by the bacterium (Downie 2005). Newer findings however, indicate that the heme moiety is also produced by the plant.

1.1.2.4 Hemocyanin

Hemocyanins are copper-containing respiratory pigments found in many mollusks (some bivalves, many gastropods, and cephalopods) and arthropods (many crustaceans, some arachnids, and the horseshoe crab, Limulus) (Klotz and Klotz 1955; Thomson et al., 1959; Johnston et al., 1967). They are colorless when deoxygenated but turn blue on oxygenation. It contains two copper atoms that reversibly bind a single oxygen molecule (O₂). Oxygenation causes a
color change between the colorless Cu (I) deoxygenated form and the blue Cu (II) oxygenated form. Hemocyanins carry oxygen in the hemolymph of most molluscs, and some arthropods, including the horseshoe crab, *Limulus polyphemus*. They are second only to hemoglobin in biological popularity of use in oxygen transport.

1.1.2.5 Hemerythrin

Hemerythrin is a non-heme iron anoligomeric protein responsible for oxygen transfer and/or storage in the marine invertebrate phyla of sipunculids, priapulids, brachiopods, and in a single annelid worm, magelona (DePhillips 1971; Loehr *et al*., 1975; Wilkins and Harrington 1983; Toulmond 1985) (Negri *et al*., 1994; Meyer and Lieb 2010). It differs from the other oxygen-binding proteins (hemoglobin and hemocyanin) both in the polypeptide chain and in the metal complex used to reversibly bind dioxygen. Hemerythrin typically exists as a homo-octamer or heterooctamer composed of α- and β-type subunits of 13-14 kDa each, although some species have dimeric, trimeric and tetrameric hemerythrins. Each subunit has a four-α-helix fold binding a binuclear iron center. Hemerythrin is essentially colorless when deoxygenated, but turn a violet-pink in the oxygenated state.

1.1.2.6 Erythrocruorin

Erythrocruorin is a large oxygen-carrying protein found in many annelids (David and Daniel 1974; Ilan *et al*., 1982). The erythrocruorins have molecular weights of 3–4 ×10^6, contain 60–192 O_2-binding haem moieties per molecule and are much more complex than the tetrameric vertebrate haemoglobins. However, they perform the same function, carrying O_2 from the respiratory surfaces to the tissues, and exhibit similar coopera-tivity in O_2 binding and inhibitory heterotrophic interactions between O_2- and proton-binding sites (Bohr effects), although these functions show greater adaptive variation than in the vertebrate pigments. Whereas erythrocruorin–O_2 affinity is insensitive to the anionic organic phosphate cofactors like glycerate-2,3-bisphosphate and ATP, which depress the O_2 affinity of vertebrate hemoglobin inside the red blood cells, it is increased by inorganic salts. This effect is important physiologically because annelids lack significant capacity for osmotic regulation and experience large fluctuations in blood electrolyte levels (Steigemann and Weber 1979; Royer *et al*., 1987; Royer *et al*., 2000).
1.2 Overview of Uncharacterized (Unannotated) protein sequences

A large number of genes discovered in sequencing projects remain functionally Unannotated, motivating significant research in postgenomic biology. The computational function annotation, plays a crucial role not only in the annotation process of newly sequenced but also in the interpretation of high-throughput experimental data such as gene expression, patterns from microarray or protein–protein interaction data. In the analysis of these data, even if a detailed function cannot be predicted, prediction of a broader category of function or subcellular localization greatly helps to cluster genes and reduce unavoidable errors in the data. This computational prediction relies on all kinds of information available in a protein sequences, co-occurrence of genes across multiple genomes, protein co-expression patterns, protein interactions and protein structures to accurately confer function primarily to sequences that are otherwise Uncharacterized (Unannotated) and also to confer additional function to sequences of partially characterized proteins. This study will focus on the current state of the field of computational function prediction of protein toward future methods that unify current resources. The study is organized as follows: the availability of a number of Uncharacterized (Unannotated) protein sequences in the database, predicting the Uncharacterized (Unannotated) protein function towards developing methods.

1.2.1 Uncharacterized (Unannotated) proteins related to plasminogen activators

Plasminogen activators are serine proteases which convert plasminogen to plasmin, thus promoting fibrinolysis. The identification of Pg-activators is very important due to their function in blood clot formation. The fibrinolytic system more appropriately referred as plasminogen (Pg) activator system, controls not only the intravascular fibrin deposition but also participates in a wide variety of physiologic and pathologic processes. The Pg-activators cleave plasminogen to produce two chains of active plasmin by a single proteolytic cleavage of Arg560-Val562 peptide bond (Castellino and Ploplis 2005). Plasmin is responsible for the degradation of blood clots. Broadly, the Pg-activators are classified into two types based on their mechanism of function; direct and indirect. Pg-activators such as tissue plasminogen activator (tPA), widely found in tissues, and urokinase (UK), originally identified in urine, are classified into direct or multiple domain activators, since they have the functional capacity to convert plasminogen to active plasmin by cleaving the Arg-Val bond on the single-chain plasminogen to release two plasmin chains. UK is a 411 residue protein, consisting of three domains and is activated by proteolytic cleavage between L158 and I159 (Lijnen et al., 1993; Lijnen et al., 1996). In contrast, Pg-
activators of bacterial origins, such as Streptokinase (SK) and Staphylokinase (SAK), are often indirect activators (Belkin *et al.*, 1986; Ouriel *et al.*, 1995; Banerjee *et al.*, 2004). Usually, indirect Pg-activators serve as co–factors of plasminogen, forming an active 1:1 stoichiometric complex with plasminogen/plasmin that degrades fibrin clots (Baruah *et al.*, 2006). SK is a single chain polypeptide made up of 414 amino acid residue and has a molar mass of 47 kDa. SK by itself is not a plasminogen activator, but it binds with free circulating plasminogen to form a complex (1:1), which can convert additional plasminogen to plasmin. Similarly, bacterial SAK, consists of a single polypeptide chain of 136 amino acids. The SAK–plasminogen complex is however inactive, until it is converted to an active SAK-plasmin complex by other Pg-activators (Rajamohan and Dikshit 2000; Rajamohan *et al.*, 2002).

**1.2.2 Uncharacterized (Unannotated) proteins – blood related proteins**

Blood proteins are found in blood plasma, and also called as serum proteins. Major blood proteins are albumin, globulin, fibrinogen, and regulatory proteins (Anderson and Anderson 1977; Adkins *et al.*, 2002). Sixty percent of plasma proteins are made up of albumin protein (Zunszain *et al.*, 2003), which are major contributors to the osmotic pressure of plasma and which assists in the transport of lipids and steroid hormones. Globulins make up thirty five percent of plasma proteins and are used in the transport of ions, hormones, and lipids thus assisting in immune function (Roux 1999). Four percent is fibrinogen and it is essential in the clotting of blood and can be converted into insoluble fibrin (Laurens *et al.*, 2006). Regulatory proteins, which make up less than one percent of plasma proteins, are proteins such as enzymes, proenzymes and hormones. The main functions of the blood proteins are transporting lipids, hormones, vitamins and metal molecules. Thus, these proteins are playing an important role in the regulation of a cellular activity and many different functions in the immune system. Due to their great function of blood, some classification prediction system has been developed in order to facilitate better understanding of their roles. The superior facility of classification system using machine learning based approach rather than experimental techniques is apparent. Currently, there is no classification of blood proteins available based on amino (AC) and dipeptide composition. Support vector machine (SVM) is one of the promising kernel based machine learning for building effective model on the training and testing for predicting class labels of unknown protein data. Therefore, in this study, we have attempted to develop an integrative SVM based prediction on a two step approach to
predict the blood proteins and further classify them into different classes. In the developed method is highly specific and sensitive to predict the blood proteins.

1.3 Need of protein functional prediction

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 the number of genomes have been sequenced. Besides advancement in sequencing technologies, the metagenomic projects (Schloss and Handelsman 2005) are also contributing in populating the sequence database. This exponentially growing list of sequences poses a major challenge for 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 the past a number of methods have been developed to predict genes, both for prokaryotes (David and Daniel 1974; Lukashin and Borodovsky 1998; Delcher et al., 1999) and eukaryotes (Issac and Raghava 2004) and repeats (Sharma et al., 2004). Second level of challenge is to predict the 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 the complexity involved in function assignment. The gap between the 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 (Chang et al., 2008; Chang et al., 2013).

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 includes: free text descriptions such as those contained in SWISSPROT 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); GPCR protein classification prediction (Bhasin and Raghava 2004; Suwa 2014). Each level of experimental protein function determination is a laborious task that can take enormous resources. Hence, automatic elucidation of protein function has emerged as a major research area of Bioinformatics.

1.4 Functional Assignment Techniques

Large-scale genome sequencing projects are continually 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 protein. In this genomic era, the task of finding all levels 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 SwissProt, which is a manually annotated protein sequence database (Bairoch and Apweiler 2000; Bairoch et al., 2005; Boeckmann et al., 2005), 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.

1.4.1 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 a target protein sequence database using alignment techniques such as BLAST (Altschul et al., 1990; Altschul et al., 1997) and FASTA (Pearson 1990). If a query protein has sequence identity more than a particular threshold with any experimentally annotated protein, then the function of query and experimentally annotated protein may be similar. In other words, functional assignment is done on the basis of homology/similarity of query proteins to proteins with known function. But the conclusions drawn from BLAST and FASTA searches are applicable only when query and target proteins share at least a minimum level of sequence identity. The 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.

1.4.2 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 a 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.

1.4.3 Level of expression

Microarray experiments are typically the measurement of expression level of thousand of genes in one step. 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 types of information, like effect of a perturbation (chemical/environmental) on the cell can be obtained (Claverie 1999). It has been shown in the past that proteins can be classified using gene expression data. Recently, it has been shown that a combination of gene expression and amino acid composition can improve the functional classification of proteins (Raghava and Han 2005).

1.4.4 Protein structure and function prediction

It is a widely accepted fact that the function of a protein is decided by the three dimensional structure of proteins. Thus the prediction of a tertiary structure of protein is important to understand its function. 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). An attempt has also been made in the past to predict the function of proteins from the predicted structure of proteins.

1.4.5 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 secretary proteins) play an 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).

1.5 Objectives of thesis

The main objective of this work is to develop Bioinformatics tools for Oxygen-binding proteins, and Uncharacterized (Unannotated) proteins related to blood proteins and plasminogen activators, which can be used to predict the function of a protein. A protein may have different functions in different environments, tissue and context. Thus the prediction of the function of a protein is not only difficult but complex too. It is beyond the scope of this work 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. The first objective of the thesis is to predict function at the protein level, where an attempt has been made to predict/classify important class of protein. This study includes prediction servers, i.e. Oxypred: for predicting oxygen-binding proteins and their classes, Bac_Hbpred: predicting bacterial hemoglobins and sub classes, PgActPred: for plasminogen activator predictions with sub classes, blood_pred: for predicting major blood proteins and their major classes.

The following objectives have been proposed for developing methods in this study.

1) Computational analysis and prediction of oxygen-binding proteins.
2) Functional annotation of Uncharacterized (Unannotated) protein sequences.
3) Web interface: The emphasis will be laid on developing algorithms which can be plugged into freely accessible web-servers.