CHAPTER 1

GENERAL INTRODUCTION

1.1. DEFINITION OF THE PROBLEM

Proteins are important work force in all living being. Proteins need to be fold or unfold in a proper way to work efficiently. Fold or unfold in protein plays important role in contemporary cell and molecular biology, regulation of protein conformation and function, intracellular protein degradation, protein structure, stability and function (Frydman, 2001). Protein folding also plays a major role in mutation. During mutations in genes, encoding proteins reduce the levels of secretory proteins due to improper folding (Bernales et al., 2006). There is considerable evidence that hydrophobicity plays a major role in protein folding. The major factor in the folding of proteins is the burial of hydrophobic side chains (James et al., 1988). Identifying foldable regions in proteins sequence based on hydrophobic residues is tedious due to identical residues at identical positions, chemically similar residues at identical positions and residues tend to occur in cluster along the length of the chain (Pang et al., 2008). There are several methods available like Sweet and Eisenberg, Kyte-Doolittle, Bull and Breese, Nozaki and Tanford, Fauchere and Pliska, Hopp and Woods, Cornette et al, Rose et al. Manavalan, Roseman et al., Lesser and Rose, Abraham and Leo, Welling et al., Miyazawa and Jernigan, Rao and Argos are available to calculate hydrophobicity from protein sequence. Each methods uses different parameters like hydropathy index, melting point, octanol scale, numerical values, average area of buried amino acids, bulk hydrophobic character, polar and charged side chains to calculate hydropathy plot (Gasteiger et al., 2005).
1.2. REVIEW OF LITERATURE

The review of literature consist of proteins, the role of hydrophobicity in proteins, different types of hydrophobicity plots, the role of carbon atom in hydrophobicity, the best computer operating system and languages for bioinformatics tools and the sliding window method for sequence analysis are discussed as follows.

1.2.1. PROTEINS

Proteins are important biomolecule plays an important role in structure, organization and function in every living system. Depending on their function variability, they are classified as transport proteins, storage proteins, toxins, hormones, enzymes, etc. Generally protein polymers are composed of amino acids monomers in an organizational way to a specific biological role.

Figure 1.1. Amino acids.
There are 20 different L-α-amino acids all the amino acids possess common structural features, including an α-carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. 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 conformation (Nelson and Cox, 2005). The amino acids are classified into polar, non polar, hydrophobic, hydrophilic and charged as given in the figure 1.1. Amino acids are covalently bonded together in chains by peptide bonds. The short chain (example: less than 30 amino acids) is known as peptide and longer chains are called polypeptides or proteins. Peptide bonds are formed between the carboxyl group of one amino acid and the amino group of the next amino acid. The four distinct aspects of a protein's structure are primary structure, secondary structure, tertiary structure and quaternary structure are shown in figure 1.2.

**Figure 1.2.** Protein structures.
Primary structure is the linear arrangement of amino acids in a protein and the location of covalent linkages such as disulfide bonds between amino acids. Secondary structure corresponds to the areas of folding or coiling within a protein; examples include alpha helices and pleated sheets, which are stabilized by hydrogen bonding (Pragya Khann, 2008).

Tertiary structure is the final three-dimensional structure of a protein, which results from a large number of non-covalent interactions between amino acids. Quaternary structure is formed by non-covalent interactions that bind multiple polypeptides into a single, larger protein. Hemoglobin has quaternary structure due to association of two alpha globin and two beta globin polyproteins. The shape into which a protein naturally folds is known as its native conformation. Although many proteins can fold unassisted, simply through the chemical properties of their amino acids, others require the aid of molecular chaperones to fold into their native states (Pragya Khann, 2008).

In protein folding, the hydrophobic effect plays important role in defining structure of proteins. The hydrophobic amino acids, such as alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and methionine clustered together within the protein. The hydrophobic core in which side chains are buried from water will stabilizes the folded state and charged / polar side chains are situated on the solvent-exposed surface to interact with surrounding water molecules. Minimizing the number of hydrophobic side chains exposed to water is the principal driving force behind the folding process (Pace et al., 1996). The major free energy contribution in protein stability is provided by hydrophobic effect (Rose et al., 2006).

1.2.2. ROLE OF HYDROPHOBICITY IN PROTEINS

Proteins are large organic compounds made of amino acid arranged in a linear fashion. The side chains of these amino acids are chemically different from one another in some respect it can be classified broadly in two ways hydrophobic and hydrophilic (Schwartz et al., 2001). A study by James et al. (1988) reveals that the
major factor in protein folding was the burial of hydrophobic side chains. A specific example is the packing of alpha-helices on beta-sheets by interdigitations of nonpolar side chains. Privalov and Gill (1986) revealed that the hydrophobic interactions are the most important non-covalent forces that are responsible for different phenomena such as structure stabilization of proteins.

The presence of both hydrophobic and polar residues interspersed through polypeptides chains is one of the most general features of amino acid sequences encoded by genes (Chothia, 1984). On studying hydrophobic run-length distributions to test for randomness of protein sequence hydrophobicities, White (1994) concluded that most protein sequences are individually indistinguishable from random sequences using theoretic perspective.

Nozaki and Tanford (1971) and Jones (1975) explained that the amino acid residues in a protein molecule are represented by their carbon atoms and each residue is assigned with the hydrophobicity index obtained from thermodynamic transfer experiments.

Michael Gromiha and Selvaraj (2004) stated hydrophobic force increasingly drives the polypeptide chain to folded, while hydrogen bonds and vander Waals define the shape. Thus the hydrophobic factor is a dominant force, while hydrogen bonds contribute to the stability of the folded state and other factors contributing only marginally.

Funahashi et al. (2003) reported that the free energy component to explains the protein stability at atom or molecular level caused by hydrophobic mutations. He have also carried out all atom molecular dynamics simulation to address the molecular mechanism of protein stability using short-range and long range interactions in different mutations.

To describe the hydrophobic interactions, Gill and Wads (1976) modeled a tetrahedral localized water molecule with one corner defined by a carbon-hydrogen group. Reynolds et al. (1974) correlate the hydrophobic free energy to aqueous cavity surface area, while free energy of transfer from hydrocarbon solvent to aqueous medium was stated as measure of hydrophobic interaction.
Sun and Parthasarathy (1994) stated the primary sequence determines the three-dimensional folded structure of a protein, than the regular folding patterns, such as alpha-helix, beta-sheet, and other ordered patterns in the three-dimensional structure. This result corresponded to the periodic distribution of the physical properties of the amino acids along the primary sequence. An Auto Regressive Moving Average (ARMA) model of spectral analysis was applied to analyze protein sequences represented by the hydrophobicity of their amino acids.

Kovacs (1997) had studied the impact of an extensive, uniform and hydrophobic protein surface on the behavior of the surrounding solvent. In particular, focus was placed on the possible enhancement of the structure of water at the interface, and have model for the hydrophobic effect.

Feldman and Hogue (2002) reported that the protein structure prediction using sequence by brute force random methods is computationally an expensive problem. Estimation suggest that it could take all the computers in the world longer than the age of the universe to compute the structure of a single 200 residue protein.

Michael Gromiha (2005) was able to predict the protein folding rate statistically from amino acid sequence with structural class information. In relating the protein folding rates with energy and conformational properties of amino acid residues and formulated a simple linear regression model for predicting the protein folding rates. The classification is based on inter-residue interactions into short, medium and long range based on geometry. Based on this contact potentials, secondary structures prediction, solvent accessibility and fold recognition can be calculated.

1.2.3. DIFFERENT TYPES OF HYDROPHOBICITY PLOTS

The numerical values were assigned to all amino acids to predict hydrophobicity for two protein sequence. The correlation coefficient is computed for all pairs of residues in these two sequences. The results suggest that the two properly aligned proteins of similar three dimensional structure have a correlation coefficient
range from 0.3 to 0.7. This observed frequency of amino acid replacements among related structures, a set of optimal matching hydrophobicity was derived (Sweet and Eisenberg, 1983). Kyte-Doolittle (1982) scale has been widely used for detecting hydrophobic regions in proteins. This scale used for identifying surface-exposed and transmembrane regions. The window sizes of 5-7 were used for predicting putative surface-exposed regions. The large window sizes of 19-21 were used for finding transmembrane domains.

Bull and Breese (1973) identified 14 proteins melting point and compared with hydrophobicity scales of Nozaki and Tanford (1971) and Chon and Edsall (1965). The calculated correlation between both the parameters showed strong correlation. The study showed hydrophobic index strongly associated with melting transition of the proteins.

Fauchere and Pliska (1983) used octanol scale as the most commonly used for estimating the contribution of buried non-polar side chains to the conformational stability of proteins. Hopp and Woods (1981) have used negative values to polar residues to predict hydrophobicity. This was used for identification of potential antigenic sites in proteins based on window size as 7.

Cornette et al. (1987) have computed an optimal hydrophobicity scale based on 28 published scales. This optimized scale was found suitable for prediction of alpha-helices in proteins. Rose et al. (1985) used the average area of buried amino acids in globular proteins to predict hydrophobicity scale and surface accessibility of proteins.

Manavalan (1978) suggested the hydrophobic character as defined by Tanford and Jones (1971) method does not reflect hydrophobic environment within protein structures. He studied on secondary structure of globular proteins using parameter called bulk hydrophobic character from hydrophobic environment of amino acids residues in protein crystals. The compared results shows correlation coefficient r = 0.9, as better than Tanford and Jones (1971) method. This new method was used to predict and characterise tertiary structures.
Janin (1979) used experimental data to indicate surface area of monomeric globular proteins. The mean accessible surface area per residues decreases like molecular weight $-1/3$ with increasing M, from about $68 \text{ Å}^2$ for proteins of 6,000 molecular weight to $38 \text{ Å}^2$ for proteins of 35,000 molecular weight. Thus, the accessibility to solvent was not a characteristic of the amino acids. This method tells about the accessible and buried amino acid residues of globular proteins. The accessible polar residues are found in surface and non polar residues are buried inside the globular protein.

Eisenberg et al. (1984) developed an algorithm to identify $\alpha$-helices involved in the interactions of membrane proteins. The membrane associated helices were distinguished based on hydrophobic plot. The magnitude of hydrophobic moment measures amphiphilicity of helix. This scale also showed the similarity of hydrophobicity scales.

Roseman et al. (1988) have used a theoretically derived scale based on structure additivity, with an experimentally derived scale obtained with N-acetylamino acid amides to predict hydrophobicity. The results show that the flanking peptide bonds dramatically reduce the hydrophilicity of the polar side-chains with deviations up to several kilocalories for the charged side-chains at pH 7.0. This result was used to construct a hydropathy scale based upon the partitioning of solutes between water and non-polar solvents.

Black and Mould (1991) have determined hydrophobicity parameters for the side chains of 22 common post- or cotranslationally modified amino acyl residues and unmodified amino acids. The result showed good correlation and was used to analyse homology, membrane directedness and folding potential of proteins.

Lesser and Rose (1990) have calculated the mean area buried upon folding of every chemical group in each residue within a set of X-ray elucidated proteins together with a standard state cavity size for each group. This study suggested that on average, each type of group buries a constant fraction of its standard state area. The mean area buried by most, though not all, groups can be closely approximated by summing contributions from three characteristic parameters corresponding to three
atom types: (1) carbon or sulfur, which turn out to be 86% buried, on average; (2) neutral oxygen or nitrogen, which are 40% buried, on average; and (3) charged oxygen or nitrogen with 32% of buried on average.

Nozaki and Tanford (1971) calculated free energy transfer of diglycine and triglycine side chain backbone peptide units from water to ethanol and dioxane. The result showed was 100% identical in both ethanol and dioxane. These results were used to calculate hydrophobicity side chains.

Abraham and Leo (1987) study indicated that polar and charged side chains influence the hydrophobicity of atoms in the side chain in a predictable manner. Field effects, as evidenced through polar proximity factors and bond factors need to be considered for accurate estimation of transfer phenomena. Thus suggest that tightly bound water to polar moieties in amino acids and peptides may be transferred into the octanol phase during partitioning experiments. This methodology presented was used characterizes the thermodynamic partitioning of groups and individual atoms in amino acids and proteins.

Welling et al. (1985) used information on the relative occurrence of amino acids in antigentic regions to make a scale. This was useful for prediction of antigentic regions. This method was found better than the Hoppen Woods (1981) scale of hydrophobicity, which is also used to identify antigentic regions. Chothia (1976) calculated accessible surface areas for individual residues in 12 proteins and the extended chains, secondary and tertiary structure of six proteins. The surfaces buried between the secondary structures were very hydrophobic and found increased proportion in the non-polar surface.

Miyazawa and Jernigan (1985) used quasi-chemical approximation method to calculate inter residue contacts of proteins by residue-residue contact observed in crystal structure. Using this method, contact pair formation resembles a chemical reaction was applied to this system to obtain formulas that relate the statistical average of the numbers of contacts to the contacts energy. The result suggested, there was a linear relationship between the average contact energies for non polar residues and their hydrophobicity as reported by Nozaki and Tanford (1971).
Rao and Argos (1986) have identified a new conformational preference parameter to calculate membrane-buried in helices using rhodopseu-domonas sphaeroides with the three-dimensional X-ray crystal structure. Engelman et al. (1986) used polar and non-polar characteristics parameter to develop a suitable scale. This newly identified scale was able to locate transmembrane helical structures in integral transmembrane proteins.

1.2.4. ROLE OF CARBON IN HYDROPHOBICITY

The carbon atom based researches by Richard Heck et al., (2010) attracted the Nobel Prize for chemistry in 2010. They developed a means of knitting carbon atoms together to form a stable skeleton. The research has led to the development of new drugs and materials. Michael Gromiha et al. (1999) reported the properties reflecting hydrophobicity strongly correlated with stability of buried mutations and there was a direct relation between the property values and the number of carbon atoms.

Chen and Kurgan (2009) stated that vander Waals contacts are formed between carbon, oxygen and nitrogen atoms covers 94.8% in proteins and organic compounds. The most common van der Waals contacts are established between carbon atom of a residue and carbon atom of a compound; carbon atom of a residue and oxygen atom of a compound.


Patel (2002) argued that at the molecular level, the best component for encoding discretized structural information is carbon. Living organisms by evolution discovered this billion of years ago and used carbon as the back-bone for constructing proteins that function according to their structure. PBD News (2002) have reported
that cells cannot merely wait for proteins to fold properly. Misfolded proteins often have carbon-rich amino acids on the surfaces. The carbon-rich patches associate strongly with similar patches on other proteins, forming large aggregates. Random aggregates are leads death to cells. The diseases like sickle cell anemia, mad cow disease and Alzheimer's disease are caused by unnatural aggregation of proteins into cell-clogging fibrils.

Jason and Andreas (2007) shows that genes with high carbon and nitrogen content were less likely to have duplicates, indicating that atomic composition also plays a role in evolution by gene duplication. Taken together, the emerging view that protein atomic composition influences genome and transcriptome evolution. Hedges (2006) concerned of those engaged in using the fossil record to reconstruct ancient diet lead on to feeding experiments on laboratory rats and the isotopic analysis of the allocation of dietary carbon to protein.

Protein structure comparison programs typically use uniform parameters (e.g. alpha-carbon positions) along the entire peptide chain and apply them to all residues by Holm and Sander (1993) and Krissinel and Henrick (2004). Such approaches have proven to be reasonably successful in aligning proteins and generating R-R correspondences. Correct and accurate alignment of hydrophobic packed structure core regions or regions composed of secondary structure elements was the implied goal of these methods.

Kleywegt (1999) have released programs (SPASM, RIGOR) that look for small motifs (example: active sites) in a given protein using a brute-force approach. The method based on inter-point distances are invariant with respect to rotation and translation. This technique performs an exhaustive search of sets of matching inter-point distances in the two point-sets. This method took advantage of the labeling information available with motif and use it to prune the search space (e.g. a carbon atom should only match another carbon atom).
1.2.5. THE BEST COMPUTER OPERATING SYSTEM AND LANGUAGE FOR BIOINFORMATICS

To optimize the best computer language and operating system for bioinformatics application. Joel and Atul (2009) discussed in table1.1 used different languages that would be appropriate and effective to seek mastery for bioinformatics, modern interpreted scripting languages. Such as Perl, Python, Ruby and many free open source operating systems are available.

**Table 1.1** Comparative analysis of operating system by Joel and Atul (2009)

<table>
<thead>
<tr>
<th>Language</th>
<th>Macintosh</th>
<th>OS/2</th>
<th>VMS</th>
<th>Win – NT</th>
<th>Unix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple script</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Unix shell</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>C/C++</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Visual Basic</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Perl</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Java</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PHP</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Fourment and Gillings (2008) benchmark provided a comparison of seven commonly used programming languages under different operating systems. The overall comparison showed that a developer should chose an appropriate language carefully, taking into account the performance expected and the library availability for each language. The best choice of language for a task would be according to the original philosophy, keeping in mind that Java is portable web oriented language, Perl is a powerful script language, Python is an easily coded language and C and C++ are efficient languages used in operating systems and drivers.
1.2.6. SLIDING WINDOW METHOD FOR SEQUENCE ANALYSIS

The different methods which are available to calculate hydrophobicity using web applications are follows. Kyte and Doolite (1982) used computer program that progressively evaluates the hydrophilicity and hydrophobicity of a protein along its amino acid sequence had been devised. Jureti et al. (1999) developed SPLIT Web server was very fast because a) it uses very simple preference functions and hydrophobic moment functions in its digital predictor, b) it uses the graphics library created to enable a fast graphical presentation of results, the only required input is the protein sequence.

Yufeng and Milton (2001) designed a web-based program WHAT, which uses a sliding window to determine and plot the hydropathy, amphipathicity, secondary structure and transmembrane topology along the length of any protein sequence. This method is based on programs designed for hydropathy and amphipathicity but on JNET and MEMSAT for secondary structure and transmembrane topology predictions, respectively. It has a user-friendly interface and a convenient input format.
1.3. OBJECTIVE

The proteins fold or unfold is decided by hydrophobicity. The hydrophobicity strongly correlated with stability and direct relation in the number of carbon atom. The present work is focused to develop: A new method of predicting hydrophobicity based on carbon distribution of a protein sequence. This distribution can predict stretch of sequence which is hydrophobic or hydrophilic based on carbon content.

1.4. WORK PLAN

To achieve the objective, the following works were done and discussed in chapters.

Chapter 1: General introduction.
Chapter 2: CARBANA: Carbon content analysis program for proteins sequence.
Chapter 3: PROTOOL: Protein tool for extracting hydrophobic characters.
Chapter 4: Pattern recognition in proteins based on carbon content.
Chapter 5: Carbon distribution studies on alanine aminotransferase.
Chapter 6: Role of carbon in protein crystal structure.
Chapter 7: Summary and Conclusion.