PREFACE

The thesis entitled "Study of Structural and Conformational Aspects of Carbohydrates by Database Analysis and Theoretical Methods" embodies the work done by the author under the guidance of Dr. K. Veluraja, Department of Physics, Manonmaniam Sundaranar University, Tirunelveli.

Carbohydrates are the most abundant organic compounds found in nature. Carbohydrates in the form of sugar and starch represent a major part of the total caloric intake for human and most animal life. They are also important material in the biological recognition process. Glycosylation is a process in which oligosaccharides having three or more monosaccharide units are joined to non-sugar molecules (proteins or lipids) in hybrid structures. In N-linked glycosylation, the carbohydrate moiety is attached to the amide nitrogen of the side chain of asparagine and in O-linked glycosylation, the carbohydrate is attached to the hydroxyl oxygen of serine or threonine.

Vital information on the sequential, structural and conformational aspects of proteins, carbohydrates and nucleic acids are available in databases. The chemical sequences of a large number of complex carbohydrates are available in the Complex Carbohydrate Structure Database (CCSD). Protein sequence and structural databases such as SWISS-PROT, PIR and PDB contain information about the site of glycosylation. Exploiting
these databases for information, regarding the aspects of carbohydrate and protein structures, is an integral part of research in structural biology.

This study attempts to bring into the surface the various structural and conformational features of carbohydrates, both free and linked to amino acid residues in proteins, and that of the glycosylating amino acid sequences in proteins. Theoretical methods have been developed to find the preference of amino acids at positions around the site of glycosylation. Deviation Parameter (DP), a normalised factor of the difference between observed and expected counts, has been computed for all the amino acids at positions around the glycosylated Asn, Ser and Thr. Positive or negative value for DP for the amino acid X at position p indicates preferential or non-preferential occurrence of X at position p. Sigma (σ) value is a measure of the statistical significance of the computed DP values. Amino acids with statistically significant DP values are said to play a vital role at particular positions in the glycosylating sequences. The favourable interactions involved in stabilising the structure of glycoproteins and glycan-binding proteins are analysed. The thesis consists of five chapters and a brief summary of each chapter is given below.

Chapter 1 provides a review about carbohydrates, its classification into monosaccharide, oligosaccharide and polysaccharides together with their structural and conformational features. Carbohydrate-protein interactions are discussed in detail with special emphasis to
glycosylation reaction. A summary of the materials already in the literature related to N-linked glycosylation and O-linked glycosylation is given. A brief introduction has been given to various nucleic acid, protein and carbohydrate databases.

Chapter 2 deals with the analysis of the Complex Carbohydrate Structure Database (CCSD). The version CCSD29 of this database containing 21987 unique carbohydrate structures were used for the analysis. The monosaccharide distribution, disaccharide distribution and their linkage pattern together with the terminal occurrence of sugars are discussed. The $\alpha, \beta$ equilibrium ratio of six carbon sugars such as glucose, galactose and mannose corresponds significantly to the $\alpha, \beta$ equilibrium ratio in solution. The commonly occurring terminal sugars lack the hydrophobic plane formed by three axial hydrogens perpendicular to the sitting plane of chair configured structures. The homo- and hetero- linkage pattern of the sugars such as Glc, Gal, Man, GlcNAc, GalNAc, Fuc, Xyl and Rha are discussed.

Chapter 3 describes the analysis of the potential N-glycosylating consensus sequences from the Brookhaven Protein Data Bank (PDB). 488 non-homologous proteins bearing 696 Asn-X-Ser/Thr sequences, where X can be any amino acid except proline, were used for this analysis. These sequences were grouped into SET I and SET II sequences using the method of identification of spatial neighbours. SET I sequences are those with Asn
having 0 to 3 spatial neighbours and are at the surface of the protein, and in
SET II sequences Asn is having more than 3 spatial neighbours and are buried
inside the protein. Deviation Parameter is computed for all the amino acids at
position X and the results show preferential occurrence for Gly, Asn, Phe and
Trp in SET I sequences. These amino acids are less preferred in SET II
sequences indicating that surface occurrence favours glycosylation. In most of
the SET I sequences a direct or water mediated hydrogen bonding between
the side chain of Asn and the side chain/backbone of Ser/Thr is found out. The
($\phi$, $\psi$) values for Gly at X position fall as a clustering, in a region which is
disallowed for non-glycyl residues in the Ramachandran plot. The significance
of the results were compared with that of the glycosylated sequences.

An analysis of the O-glycosylated sequences from O-GLYCBASE
database has been carried out in Chapter 4. The 992 sequences from this
database are separated into single glycosylation sequences and multiple
glycosylation sequences. The DP values computed for the amino acids at
positions from -10 to +10 around the site of glycosylation revealed the
positional preference of proline at various positions around the glycosylated
Ser/Thr. Ser and Thr are favoured around multiple glycosylation sites due to
the clusters of closely spaced O-glycosylation sites. Ala, Gly, Asp, His and Val
are preferred at few positions around multiple glycosylation sites. The same
results hold good for the mucin type O-glycosylation, and in addition the acidic
amino acids are preferred at various positions around the glycosylated Ser/Thr. The preference of sequential motifs have been discussed. The conformational features of the O-glycosylated sequences have been analysed using the O-glycosylated proteins from Protein Data Bank.

In Chapter 5, a database analysis has been carried out to find out the interaction between the carbohydrate and the amino acids in glycoproteins and glycan-binding proteins. It is known that β-sugars of Glc, Gal, Man, GlcNAc, GalNAc and ManNAc possess three axial hydrogens, H1, H3, H5, pointing the same direction and can form a plane. Also the acetamido sugars possess a CH3 group. These can form stacking/ hydrophobic interaction with the aromatic and hydrophobic amino acids such as Phe, Trp, Tyr, Ala, Val, Leu and Ile. The hydrophobic interactions in glycoproteins and glycan-binding proteins are discussed separately. To find out the order of preference of various amino acids involved in the hydrophobic interaction with the sugars, a quantity termed as hydrophobic interaction index (HII) is defined. Aromatic amino acids possess high value of HII. Hydrophobic/stacking interaction is found in most of the glycoproteins and glycan-binding proteins and hence is one of the key interactions in stabilising the carbohydrates in glycoproteins and glycan-binding proteins.

A summary describing the salient features and the conclusions drawn from this study is presented at the end.