1.1 Carbohydrates

Carbohydrates are widely diverse group of compounds that are ubiquitous in nature and are one of the four major class of biomolecules along with proteins, nucleic acids and lipids (Varki et al., 2002). Chemically, carbohydrates are mainly made up of carbon, oxygen and hydrogen atoms. The name carbohydrate was originally assigned to compounds believed to be hydrates of carbon, of the general formula $C_n(H_2O)_m$. In general carbohydrates are aldehyde or ketone compounds with multiple hydroxyl groups. Carbohydrates are the main source of energy for living organisms and the central pathway of energy supply in most of the cells. They are the major products through which the energy of the sun is harnessed and converted into a form that can be utilized by man and other animals, as well as by many other organisms. The amount of carbohydrates found in nature is larger than that of any other group of natural compounds (Sharon, 1975). Carbohydrates are classified as monosaccharides, disaccharides, oligosaccharides and polysaccharides. Monosaccharides are the basic building block for oligosaccharides and polysaccharides. Monosaccharides are single units of sugar which cannot be further hydrolyzed into smaller carbohydrates. Disaccharide yields two monosaccharide units during hydrolysis. Tri, tetra and petasaccharides gives 3, 4 and 5 monosaccharide units respectively. Sucrose, maltose, trehalose, lactose and melibiose are a few important disaccharides. Sucrose is commonly known as table sugar and is formed by the two monosaccharides glucose and fructose through an $\alpha(1\rightarrow2)$ linkage.
Maltose is glucose\(_{(1\rightarrow4)}\)glucose; trehalose is glucose\(_{(1\rightarrow1)}\)\(_{\alpha}\)glucose; lactose is galactose\(_{(1\rightarrow4)}\)glucose and melibiose is galactose\(_{(1\rightarrow6)}\)glucose. The term “oligosaccharide” is used to describe a molecule that contains two to ten monosaccharides and the term polysaccharide is used to describe a large carbohydrate molecule. The most abundant organic substance on earth is cellulose, a polymer of glucose, which is the structural material of plants (Klemm et al., 2005). Chitin is a polymer of \(N\)-acetylglucosamine and is the major organic component of the exoskeleton of arthropods such as insects, crabs and lobsters (Dahiya, 2006; Rinaudo, 2006).

The structure of a carbohydrate is complex when compared to the other biomacromolecules such as proteins or nucleic acids. This complexity originates from the following factors:

(i) Sugar molecules have more number of hydroxyl groups and so they can contribute to diversified linkages.

(ii) Complexity in linkage also increases due to the existence of \(\alpha\) and \(\beta\) anomers.

For example, glucose molecule when made into a disaccharide with another glucose molecule, can form eleven different disaccharides [Figure 1.1].

(iii) In addition to this, branching also create an added dimension to this complexity.

Carbohydrates exist in various forms such as polysaccharides (pure carbohydrates), glycoproteins (carbohydrates with proteins) and glycolipids (carbohydrates with lipids). Commonly glycoproteins and glycolipids are collectively
called as glycoconjugates. The glycans of glycoprotein can be divided into two groups viz N-glycans which are linked to proteins through amide group of asparagine and O-glycans which are linked to proteins through hydroxyl group of serine, threonine, tyrosine, 4-hydroxylysine or hydroxyproline.

Figure 1.1: Glycosidic linkage diversity of glucose molecule
1.1.1 Structural elucidation of carbohydrates

The three dimensional structures of carbohydrates play important roles in their functions. Single crystal X-ray crystallography, X-ray fiber diffraction and NMR spectroscopy are widely used experimental methods to deduce the conformation of carbohydrates (Atkins et al., 1974; Atkins et al., 1995; Rini et al., 1993; Bock and Duus, 1994; Poveda and Jimenez-Barbero, 1998; Leeflang et al., 2000; Loris et al., 2000; Bubb, 2003; Palmer and Niwa, 2003; O’Brien and Moyna, 2004; Perez and Mulloy, 2005; Yu and Prestegard, 2006). However they are inadequate to explain the multiple conformations of carbohydrates in solution. The heterogeneity of oligosaccharide structures make the isolation and crystallization of oligosaccharides as a challenging task. In solid state, the multiple conformations of carbohydrate cannot be explained. Nowadays, computational methods have a profound role in studying the conformational aspects of carbohydrates and protein-carbohydrate interactions. Theoretical methods such as Molecular Dynamics simulations (MD), Quantum Mechanical (QM) calculations are widely used to study about the conformations of carbohydrates (Qasba et al., 1994; Woods, 1998; Momany and Willet, 2000; Vasudevan and Balaji, 2001, 2002; Smith et al., 2003; Bosques et al., 2004, Veluraja and Margulis, 2005; Veluraja and Seethalakshmi, 2008; Fadda and Woods, 2010; Schnupf et al., 2010; Karakurt et al., 2011; Rogers et al., 2011; Patel and Balaji, 2011; Khebichat et al., 2012).
1.1.2 Conformational parameters of carbohydrates

Monosaccharides are the basic building blocks of higher carbohydrates. They are polyhydroxy aldehydes (aldoses) or polyhydroxy ketones (ketoses). Monosaccharides can exist in an open chain (acyclic) form or in closed chain (cyclic) form. In general, Fischer projection is used to represent the open chain form of monosaccharides, Haworth projection is used to represent cyclic form. The five-member closed chain form of a monosaccharide is known as furanose, while the six-member cyclic form of monosaccharide is called as pyranose. Often six-member monosaccharides can be represented in chair conformation. Due to the possible rotations of single bonds within the pyranose ring, the ring can have number of conformations such as chair, boat, skew boat and half chair conformations. Of the various conformations chair conformation is the most favoured conformations for numerous six membered monosaccharides. There are two chair conformations which are distinguished and designated as $^4C_1$ and $^1C_4$. It has been already reported that the pyranose ring of carbohydrates mostly exists in chair conformations through X-ray crystallography (Cox and Jeffrey, 1939; McDonald and Beevers, 1952; Jeffrey, 1990), Nuclear Magnetic Resonance spectroscopy (Lemieux et al., 1958; Lenz and Heeschen, 1961; Rao and Foster, 1965; Angyal, 1969; Durette and Horton, 1971; Rao, 1974). Theoretical studies also confirm that the ring exists in chair conformation (Rao et al., 1971; Vijayalakshmi and Rao, 1972, 1973a; Vijayalakshmi et al., 1973b; Joshi and Rao, 1979; Veluraja and Rao, 1980; Madsen et al., 1990; Rao et al., 1998; Bryce et al., 2001; Fadda and Woods, 2010). A typical monosaccharide unit in chair conformation is shown in Figure 1.2.
Conformation of a disaccharide

The conformation of a disaccharide can be described by the glycosidic torsional angles $\Phi$ and $\Psi$ which arise due to the free rotation around glycosidic bonds. Figure 1.3 shows the typical disaccharide along with the glycosidic torsional angles $\Phi$ and $\Psi$. 

Figure 1.2: Structure of $\alpha$-D-Glucose in chair conformation a) $^4C_1$ Chair b) $^1C_4$

Figure 1.3: Structure of a disaccharide Glc$\beta$(1-4)Glc along with glycosidic torsional angles
Φ represents the glycosidic torsional angle that arise due to the rotation about C1 – O bond and Ψ represents the glycosidic torsional angle arising due to the rotation about O – CX’ bond where X’ is either 1 or 2 or 3 or 4 or 6 that depends upon the type of the glycosidic linkage. The glycosidic torsional angle Φ is 0° when H1 – C1 bond eclipses the O – CX bond and Ψ is 0° when C1 – O bond eclipses the CX – HX bond. For both Φ and Ψ angles, a clockwise rotation is taken as positive. The range of values for Φ and Ψ varies from 0° to +180° and 0° to -180°. The allowed conformation for a disaccharide depends upon the individual monosaccharide conformation along with the type of linkage. A typical steric map for a disaccharide is shown in Figure 1.4.

![Figure 1.4: Typical steric map for Glcβ(1-4)Glc](image)
1.2 Sialic acid and its Structural Diversities

Sialic acids are nine carbon amino sugars and are discovered by Gunnar Blix and Ernst Klenk in 1957 (Blix et al., 1957). More than 50 diversified forms of sialic acids occur in nature (Varki et al., 2002). They are commonly found in animal tissues and in bacteria. Of the diversified forms of sialic acid, the predominant one is N-acetylneuraminic acid (Varki and Varki, 2007). The schematic representation of N-acetylneuraminic acid structure is shown in Figure 1.5.

![Figure 1.5: Structure of α-D-N-acetyleneuraminic acid (α-Neu5Ac)](image)

In human beings sialic acids are present in body fluids (blood plasma, breast milk, gallbladder, excretions and synovial fluid) and tissues (erythrocytes, leucocytes, platelets, salivary glands, throat, stomach, cervix, colon, cartilage) (Accili et al., 1994;
Sillanaukee et al., 1999; Wang et al., 2001; Schauer, 2004; Boehm and Stahl, 2007). The special structural features of sialic acids are amino group at C5 position, carboxylate group at C2 (anomeric) position which confers a negative charge on a molecule under physiological conditions and characterizes it as a strong organic acid (pKa value is 2.2) (Berg et al., 2002). It has a glycerol side chain group at C6 position. The two anomeric forms of sialic acids are α and β. α-anomer of N-acetylneuraminic acid occurs as a part of oligosaccharides and the β-anomer predominantly exists in solution. The pyranose ring of sialic acid exists in $^2C_5$ chair conformation (Flippen, 1973; Brown et al., 1975; Czarnecki and Thornton, 1977; Veluraja and Rao, 1980; Spiwok and Tvaroska, 2009).

The structural diversities of the sialic acids arise due to: (i) the functional group modification at C5 position. The amino group (NH$_2$) at C5 position leads to Neuraminic acid and is the parent molecule for all sialic acids. Most commonly they are called as Neuraminic acid or Neu. The substitution of N-acetyl group (NH-CO-CH$_3$) at C5 position leads to N-acetylneuraminic acid and is called as Neu5Ac, NANA, NeuNAc or Sia. N-glycolyl group (NH-CO-CH$_2$OH) at C5 leads to N-glycolylneuraminic acid and is denoted as Neu5Gc. Hydroxyl (OH) present at C5 position leads to deamino neuraminic acid (KDN) (Varki, 1992; Schauer, 2000; Angata and Varki, 2002). The unsubstituted form of neuraminic acid does not exist in nature. Among the C5 functional group modification, α-D-N-acetylneuraminic acid is the most widespread form. The other modifications are acetyl, lactyl, sulphate and methyl group substitution at 4th, 7th, 8th and 9th hydroxyl position of sialic acid. Of these different substitutions O-acetylation is the most important diversity.
The common form of linkages found in the sialic acid containing oligosaccharides are \(\alpha(2\rightarrow3)\), \(\alpha(2\rightarrow6)\) linkages with galactose (Gal) or N-acetylgalactosamine (GalNAc) and \(\alpha(2\rightarrow8)\), \(\alpha(2\rightarrow9)\) linkage with another sialic acid.

### 1.2.1 Biological Functions of Sialic acid

The structure of sialic acid with its negative charge and the terminal occurrence at the carbohydrate chain of cell membrane glycoconjugates lead to act as determinants of molecular and cellular interactions (Kelm and Schauer, 1997; Varki, 1997; Rinninger et al., 2006; Varki, 2008; Schauer, 2009; Chen and Varki, 2010a; Schauer et al., 2011; Li and Chen, 2012). The acetylated derivatives of sialic acid at different hydroxyl positions are involved in many biological functions. Murine hepatitis virus S and Influenza C viruses recognize 4-O-acetyl-N-acetyl-neuraminic acid and 9-O-acetyl-neuraminic acid respectively as receptors (Zimmer, 1992; Regl et al., 1999; Hellebo et al., 2004; Langereis et al., 2010). Due to the negative charge, sialic acids are involved in binding and transport of positively charged molecules as well as in attraction and repulsion phenomena between cells and molecules. Many pathogenic agents such as toxins (cholera toxin, tetanus toxin, botulinum neuro toxin), bacteria (Escherichia coli, Helicobacter pylori), viruses (influenza virus, sendai virus) and protozoa (Trypanosoma cruzi) bind with host cells through sialic acid containing receptors (Merritt et al., 1994; Singh et al., 2000; Majoul et al., 2002; Wurzer et al., 2002; Tsai et al., 2003; Zeng et al., 2008; Langereis et al., 2010). Sialic acid is commonly found at the terminal of complex carbohydrates with either \(\alpha(2\rightarrow3)\) or \(\alpha(2\rightarrow6)\) linkage to galactose moiety and this moiety is preferentially recognized by
viruses (Rogers and Paulson, 1983; Arnberg et al., 2000; Helander et al., 2003; Stevenson et al., 2004; Blackburn et al., 2005; Dugan et al., 2005). It has been reported that Influenza viruses recognize $\alpha(2\rightarrow3)$ linked sialic acid found on avian intestinal epithelium whereas human strains recognize $\alpha(2\rightarrow6)$ linked sialic acids found in human respiratory epithelial cells (Matrosovich et al., 2004). Equine viruses recognize $\alpha(2\rightarrow3)$ linked sialic acid and Feline calci virus recognizes $\alpha(2\rightarrow6)$ linkage (Stuart and Brown, 2007). Sialic acid containing oligosaccharides play an important role in the adhesion between cancer cells and endothelial cells and also demonstrated that metastatic potential of tumor cells is proportional to cell surface sialylation (Wang, 2005; Malati, 2007). It has been reported that sialic acid is used as tumor marker for lung cancer (Kakari et al., 1991; Patel et al., 1994; Patel et al., 1995; Krug et al., 2004; Gokmen et al., 2004), breast cancer (Marquina et al., 1996; Brooks and Carter, 2001a; Brooks et al., 2001b; Raval et al., 2004), prostate cancer (Suer et al., 1996). Sialic acid is also used as a marker for cardiovascular disease (Gopaul and Crook, 2006). The structure of sialic acid plays an important role in biological functions.

1.3. Protein

Proteins are linear polymers of amino acids formed by reaction between the carboxylic acid group of one amino acid and the amino group of another amino acid with elimination of water by means of linkage called peptide linkage. The ratio of the major elements in proteins are C=50-55%, H=6-8%, O=20-23%, N=15-18% and S=0-4%. The protein structures are described by four levels viz primary structure, secondary structure, tertiary structure and quaternary structure. Proteins have number
of functional groups which include alcohols, thiols, thioesters, carboxylic acids, carboxamides and a variety of basic groups. Proteins are classified as simple proteins, conjugated proteins and derived proteins based on their solubility and chemical properties (Zhao et al., 2008). Simple proteins contain only amino acids and they do not contain any other additional compounds. Albumins, Globulins, Glutelins, Prolamines and Histones are belongs to the category of simple proteins. The proteins with association of non aminoacids are called as conjugated proteins. The non amino acid components are called as prosthetic group. There are seven major categories of conjugated proteins and are classified as follows: Proteins contain nucleic acids (nucleoprotein), carbohydrates (glycoprotein), pigments (chromoproteins), lipids (lipoprotein), enzymes (metalloproteins), phosphoric acid (phophoproteins) and lectins (lecithoprotein). Primary proteins on hydrolysis with acids or alkalies yield derived proteins. Based on the biological functions of proteins they are classified as enzymoproteins, transport proteins, storage proteins, contractile protein, structural protein, defense proteins and regulatory proteins.

Proteins are very important macromolecules in living systems and play a wide variety of roles in all biological functions (Berg et al., 2002; Disney and Seeberger, 2004; Huang et al., 2010; Romani et al., 2010). Most of the vital activities of the cell are controlled by proteins and nucleic acids. They are involved in virtually all cell functions. Each protein within the body has a specific function. Some proteins are involved in structural support, while others are involved in bodily movement, or in defense against germs. Antibodies are specialized proteins involved in defending the body from antigens (foreign invaders). One way antibodies destroy antigens is by immobilizing them so that they can be destroyed by white blood cells. Contractile
proteins are responsible for movement. Actin and myosin are contractile proteins present in muscle cells in the form of filamentous protein and are functioning in muscle contractile systems (Webb, 2003; Varkuti et al., 2012). Enzymes facilitate biochemical reactions. They are often referred to as catalysts because they speed up chemical reactions. Lactase breaks down the sugar lactose found in milk. Pepsin is a digestive enzyme that works in the stomach to break down proteins in food. Hormonal proteins are messenger proteins which help to coordinate certain bodily activities. Examples include insulin, oxytocin, and somatotropin (Nannipieri et al., 2011). Hemoglobin is a globular protein present in Red blood Corpuscles of blood can bind with oxygen when blood passes through lungs and distributes oxygen throughout the body cells to affect cellular respiration. Blood plasma contains lipoprotein which carries lipids from the liver to other organs. Oxytocin stimulates contractions in females during childbirth. Somatotropin is a growth hormone that stimulates protein production in muscle cells. Structural Proteins are fibrous, stringy and provide support. Examples include keratin, collagen, and elastin. Keratins strengthen protective coverings such as hair, quills, feathers, horns, and beaks. Collagens and elastin provide support for connective tissues such as tendons and ligaments. Storage proteins are used to store amino acids and examples for such a protein are ovalbumin and casein. Ovalbumin is found in egg white and casein is a milk-based protein. Transport proteins act as carrier proteins which move molecules from one place to another around the body (Glavinas et al., 2004; Quick and Javitch, 2007; Stolarczyk et al., 2011). Regulatory proteins help to regulate cellular or physiological activity. Insulin is a regulatory protein formed in pancreatic tissue to help to regulate the blood sugar level. Growth hormones of pituitary and parathyroid harmones regulate
Ca$^{2+}$ and phosphate transport in body. Viral proteins play an important role either in regulating virus replication, such as entry, assembly and release or modulating the electrochemical balance in the subcellular compartments (Wang et al., 2011; Scott et al., 2006; Lin and Zhang, 2007; Giner et al., 2010; Rappoport and Linial, 2012).

1.3.1 Protein–Carbohydrate interactions

Many biological processes are mediated by recognition events involving proteins and carbohydrates, and the study of these interactions constitutes the rapid growing field of glycobiology (Lis and Sharon, 1993; Varki, 1993; Dwek, 1996; Dinglasan and Jacobs-Lorena, 2005; Bohua et al., 2009). The interaction between the carbohydrate and protein or any other macromolecule is the prior condition to initiate the biochemical reaction or biological processes (Iniguez-Palomares et al., 2011). To understand the interactions between the oligosaccharide and the specific binding site in proteins, knowledge of the three dimensional structure of the protein and its complex with the ligand is more important. The experimental method X-ray crystallography has provided a wealth of information on the three dimensional structures of protein–carbohydrate complexes (DePristo et al., 2004; Chen et al., 2012). In many cases, oligosaccharides, either in the free form in solution or when present as glycoconjugates or complexed to protein receptors such as antibodies, lectins, and enzymes have proved to be difficult to crystallize. NMR experiments have been frequently used both in the area of oligosaccharide conformational analysis and for studying oligosaccharide–protein interactions. However, and in general, NMR analysis of lectin–saccharide complexes has been hampered by rather the high molecular weight of sugar-binding proteins (lectins) (Siebert et al., 1997;
Lectins are generically defined as proteins which interact non-covalently with carbohydrate moieties and show low affinities for simple mono- and oligosaccharides (Imberty and Perez, 1994; Angata and Varki, 2000). In any case, NMR methods have provided insights into the structural characterization of small lectins and on the thermodynamic driving forces controlling their encounters with carbohydrates (Poveda et al., 1997). In general, however, experimental studies yield a representation consisting of one single static picture of the system. In order to understand the dynamics of recognition at atomic level, we have little choice but to turn to theoretical methods (Rao et al., 1990; Veluraja and Margulis, 2005).

Protein–carbohydrate interactions are the basis of numerous biological processes for both cellular and pathological processes (Vijayan, 2009). These interactions generally require high-affinity binding (Olausson et al., 2008; Ferrara et al., 2011). They include the enzymatic synthesis and degradation of oligo- and polysaccharides, intracellular sorting of glycoconjugates, transport of carbohydrates into living cells and of their derivatives into subcellular organelles, the immunological response to carbohydrate antigens, and migration of leukocytes to sites of inflammation. Protein–carbohydrate interactions also play key roles in a variety of cell adhesion phenomena such as the attachment of parasites, fungi, bacteria, and viruses to host cells (Goto, 2007; Cox, 2007; Bagchi, 2008; Fisk and Read, 2011; Doyle, 2011). Higher affinity interactions occur when lectins, which are oligomeric proteins, bind to the carbohydrate chains of cell surface glycolipids and glycoproteins, which possess multiple binding epitopes. They were originally discovered in plants and its important function is its ability to agglutinate blood cells. Lectins are also found in
animals, bacteria, fungi and viruses. They are involved in many biological functions which includes cell-cell communication, host pathogen interactions, cancer metastasis (Gabius, 1997; Ishikawa et al., 2005; Inata et al., 2007; Shao et al., 2009), embryogenesis (Brill et al., 2001; Ishizaki et al., 2002), and tissue development (Francois and Balzarini, 2012). Cholera toxin and heat labile toxin are AB$_5$ toxins and use GM1 as their molecular target to attack and penetrate the host cells. Bacterial adhesion is often prelude to infection. In mammalian innate immune system (Ni and Tizard, 1996), lectins play a crucial role by capturing pathogens through glycans on their surface and presenting the foreign epitopes to CD4$^+$ T lymphocytes. This concept has its origins in the lock-and-key hypothesis, introduced by Emil Fisher at the end of the 19$^{th}$ century to explain the specificity of interactions between enzymes and their substrates, i.e., between molecules in solution. It was subsequently extended to describe the interactions of cells with soluble molecules and with other cells. Understanding of protein-carbohydrate interaction will help to design an inhibitor for a pathogenic disease (Holgersson et al., 2005).

The major stabilizing interactions in protein-carbohydrate complexes are vander Wall’s interactions, hydrogen bonding, hydrophobic, hydrophilic and water mediated hydrogen bonding interactions (Stephanie et al., 2010; Nurisso et al., 2010).

### 1.3.2 Online resources for carbohydrates and protein structures

**Carbohydrate databases**

The 3D structures of carbohydrates are available in Cambridge Structure Database (http://www.ccdc.cam.ac.uk/products/csd/), which contains more than 4000 structures. GLYCO3D stores many sugar structures.
Several entries in the Protein Data Bank contain carbohydrates, either as ligands or as glycoproteins. It has been reported by Lutteke that the 3D structures of carbohydrates available in PDB contains at least one error (Lutteke and von der Lieth, 2004; Lutteke, 2009). Oligosaccharide conformations predicted by Francesco Strino 7 with SHAPE (Rosén et al., 2009) are stored in the 3D-BAO database (http://www.cermav.cnrs.fr/cgi-bin/bao/3D-BAO.cgi). Three-dimensional structures of sialic acid containing carbohydrate structures are available in 3DSDSCAR database (http://www.3dsdscar.org) (Veluraja et al., 2010).

**Swiss-Prot**

Swiss-Prot is a curated protein sequence database that strives to provide a high level of annotations which includes the description of the function of a protein, its domain structure, post-translational modifications and variations. This database can be accessed and searched through SRS system (http://us.expasy.org/srs5/) at ExPASY or one can download the entire database as one flat file. Swiss-Prot is maintained collaboratively by the Swiss Institute of Bioinformatics (SIB), Department of Medical Biochemistry of the University of Geneva and the EMBL outstation, the European Bioinformatics Institute (EBI) (Stoesser et al., 1999; Boeckmann et al., 2003). TrEMBL is a computer annotated supplement of Swiss-Prot that contains all the translations of nucleotide sequence entries not yet integrated in Swiss-Prot. As on 10th July 2012, Swiss-Prot contains 536789 entries.
Protein Data Bank

Protein Data Bank (PDB) is the three dimensional structural storage for biomolecules that mostly includes Proteins and nucleic acids. PDB was established by Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. By 1974 there were 12 protein structures which include myoglobin, hemoglobin, carboxypeptidase A and subtilisin. The crystal structures explored from X-ray and Nuclear Magnetic Resonance (NMR) spectroscopy techniques by the biophysicists, biologists and biochemists are deposited in PDB database. The archive is managed by the partners of the Worldwide Protein Data Bank (wwPDB, http://www.wwpdb.org) (Berman et al., 2003), RCSB PDB (Berman et al., 2000), PDBe (Europe; http://www.pdbe.org) (Velankar et al., 2011), PDBj (Japan; http://www.pdbj.org) and the Biological Magnetic Resonance Databank (BMRB, United States; http://www.bmrb.wisc.edu). wwPDB members host data deposition and annotation sites, distribute data, and collaborate on issues of policy, formats, standards and curation. Each members also develop different tools and resources to study and utilize the data. The RCSB PDB website provides resources and tools to mine and analyze sequences, structures, ligands and annotations. At present PDB contains 82992 crystal structures of macromolecules as on 10th July 2012.

1.4 *Vibrio Cholerae* Neuraminidase

Sialidases or neuraminidases make up an important class of glycoside hydrolases that catalyze the removal of the terminal sialic acid from various glycoconjugates. *Vibrio cholerae* is a gram-negative bacterium and is in rod shape.
They are facultative anaerobic, which means they can survive either with or without oxygen. *Vibrio cholera* produces cholera toxin and that can adhere to gangliosides GM1 present on the epithelial cells (Thompson and Schendurnd, 1998). The two different types of *Vibrio cholerae* are *Vibrio cholerae* Serogroup O1 and *Vibrio cholerae* Serogroup non-O1. *Vibrio cholerae* bacterium produces cholera toxin, whose action on the mucosal epithelium is responsible for the characteristic for cholera disease. *Vibrio cholera* of O1 and O139 serotypes are responsible for epidemic cholera in humans. All the other serotypes are grouped as “non-O1” strains. Neuraminidases are enzymes that catalyze the removal of terminal sialic acid from the carbohydrate moieties in the surfaces of infected cells. *Vibrio Cholerae* neuraminidase (VCN) plays a significant role in the pathogenesis of cholera by removing the sialic acid from higher order gangliosides to unmask GM1 as the receptor for cholera toxin (Hinou et al., 2005).

Structural studies on VCN by Taylor and Itzstein in 1994 revealed that this enzyme is composed of a central β-propeller catalytic domain flanked by two lectin-like domains (Moustafa et al., 2004; Taylor and Itzstein, 1994). The rigid active site of VCN appeared to be very similar to those observed in other sialidases such as *Salmonella* neuraminidase and influenza virus neuraminidase (Taylor and Itzstein, 1994; Crennell et al., 1994). The following features are common to these three members: (a) three arginine residues (arginine triad) which stabilize the carboxyl group of the sialic acid, (b) a glutamic acid which stabilizes one of these arginines, (c) a tyrosine residue whose hydroxyl group is close to the anomeric position of the sialic acid, (d) a glutamate with which the tyrosine has been implicated in the catalytic mechanism, (e) a hydrophobic cavity which accommodates the N-acetyl group of
sialic acid, and includes a conserved tryptophan, and (f) a solvent-exposed aspartic acid which could act as a proton donor to the glycosidic oxygen or could be involved in stabilizing a proton-donating water molecule (Crennell et al., 1994). This overall fold similarity defined by eight strictly conserved residues at the active site strongly suggests a similar mode of action for the entire superfamily of sialidases.

1.5 Influenza Virus

Influenza viruses are the member of orthomyxovirus family (Lamb and Krug, 2001). They are enveloped viruses with a genome of negative sense, single stranded, eight segments RNA. Influenza Virus particles are in spherical shape and approximately 100nm in diameter (Fujiyoshi et al., 1994). Eight segments encode of eleven proteins (Palese and Shaw, 2007); hemagglutinin (HA), neuraminidase (NA), matrix 1 (M1), matrix 2 (M2), nucleo protein (NP), non-structural protein 1 (NSP1), non-structural protein 2 (NSP2 also known as Nuclear Export Protein (NEP), polymerase acidic protein (PA), polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2) and polymerase basic protein 1-F2 (PB1-F2) (Betakova, 2007; Hale et al., 2008; Chen et al., 2004; Zhao et al., 2009). The viral envelope consists of transmembrane protein on the outside and the matrix protein on the inside. Hemagglutinin (HA), Neuraminidase (NA) and ion channel M2 protein are anchored in the lipid bilayer of the viral envelope. The structure of Influenza A virus is shown in Figure 1.6.
HA is type I transmembrane protein and is a homotrimer, each monomer is composed of two subunits HA1 and HA2 (Lazarowitz et al., 1973). The receptor binding site is located at membrane-distal tip of each monomer. HA is responsible for the viral binding with the receptors which contains sialic acid as a terminal residue on the host cell surface and mediates fusion of the viral envelope with the endosomal membrane after receptor-mediated endocytosis. Cleavage of HA is essential for fusion and viral infection. NA is the type II transmembrane protein and is a homotetramer. NA removes the cell surface receptor and is critical for the release of virus particles from the cell surface and spread of virus (Ghedin et al., 2005; Sylte and Suarez, 2009). M2 is the type III transmembrane protein with homotetramer. M2 act as an ion channel that permits ions to enter the virion during uncoating and equilibration of the pH
gradient between the acidic lumen of the trans-Golgi network and the neutral cytoplasm (Sugrue and Haye, 1991; Hay, 1992; Kido et al., 2009; Lazrak et al., 2009; Rossman et al., 2010; Cross et al., 2012) for dissociating the vRNP from M1 in the early phase of the infectious cycle (Schroeder et al., 2005; Nayak et al., 2009). Of these three transmembrane proteins, HA is the most abundant envelope protein at approximately 80% followed by NA which makes up around 7% of the viral envelope proteins. M2 is the very minor component of the envelope, with only 16 to 20 molecules per virion.

The viral core consists of helical ribonucleocapsids (i.e. vRNP) containing vRNA and NP along with little amount of the nuclear export protein (NEP) and three polymerase proteins PB1, PB2 and PA which form the viral RNA polymerase complex (Klumpp et al., 1997; Portela and Digard, 2002; Sugiyama et al., 2009). Based on the matrix protein (M) and NP Influenza viruses are classified into three distinct strain types viz A, B and C (Palese and Young, 1982; Tobita, 1997). Most of the Influenza outbreaks are associated with A and B type. Among the three types Influenza A infects human, swine, horse, seal, mink, whale and birds (Iqbal et al., 2009; Bui et al., 2012). Influenza B and C infects human and they do not infect birds. Influenza A viruses are further divided into distinct subtypes depending on the surface glycoprotein HA and NA. So far 16 HA subtypes and 9 NA subtypes has been identified (Karasin et al., 2000; Nielsen et al., 2001; Olsen et al., 2006; Boonsuk et al., 2008; Gagnon et al., 2009; Goyal et al., 2010; Kim et al., 2010; Shi et al., 2010). The combinations of 16 HA subtypes and 9 NA subtypes gives 144 possible combinations of HA and NA and hence 144 theoretically possible subtypes. H1N1, H5N1, H3N2 are some examples for the possible subtype combinations. Influenza is commonly
called as “flu”. H1N1 and H5N1 subtypes are commonly known as “Swine flu” and “Bird flu” respectively. Influenza A viruses infect humans, pigs, sea mammals, horses and birds. The phylogenetic studies of influenza A viruses revealed that aquatic birds are the source of all Influenza viruses.

### 1.5.1 Receptors for Influenza A Virus

Influenza viruses recognize the oligosaccharides which contain Neu5Ac\(\alpha(2\rightarrow3)\)Gal or Neu5Ac\(\alpha(2\rightarrow6)\)Gal moiety (Matchytka et al., 1993; Suzuki et al., 2001; Pekosz et al., 2009; Hidari and Suzuki, 2010; Nicholls et al., 2007; Shinya et al., 2006; Kimble et al., 2010; Sawada et al., 2009). The receptor binding site of the HA is composed of Helix190, Loop130 and Loop220 (Lin et al., 2009). It has been reported that HA distinguishes the differences in the linkage of Neu5Ac\(\alpha(2\rightarrow3)\)Gal or Neu5Ac\(\alpha(2\rightarrow6)\)Gal on the host cells. Human HA strains preferentially bind to the terminal Neu5Ac\(\alpha(2\rightarrow6)\)Gal structure that is found on the surface of some tracheal epithelial cells, whereas avian and equine strains bind to Neu5Ac\(\alpha(2\rightarrow3)\)Gal structure (Chua and Chai, 2012).

It is reported that a change in single amino acid can alter the receptor binding preferences from Neu5Ac\(\alpha(2\rightarrow3)\)Gal to Neu5Ac\(\alpha(2\rightarrow6)\)Gal. In the case of H1, H2 and H3 subtypes, it was shown that as few as two mutations at the receptor binding site are responsible for the adaptation to human host by genetic and mutational studies (Naeve et al., 1984; Rogers and D’Souza, 1989; Connor et al., 1994; Glaser et al., 2005). The substitutions Gln226Leu and Gly228Ser switch the binding specificity of H2 and H3 HAs from avian to human receptor, whereas the substitutions Glu190Asp and Gly225Asp change the receptor specificity of H1.
1.6 **Perspective of the Study**

Three dimensional structures of carbohydrates and proteins play an important role in many biological functions. Sialic acids, acetylated sialic acids, sialyldisaccharides and sialyloligosaccharides are important carbohydrate molecules that play vital role in molecular recognition phenomena. Hence this present work is aimed

(i) To study the conformational flexibility of $\alpha$-Neu5Ac in aqueous environment using MD simulation of 10ns duration

(ii) To deduce the conformational models for $\alpha$-Neu5Ac

(iii) To obtain the optimized geometry of $\alpha$-Neu5Ac using QM calculation

(iv) To find out the conformational flexibility and to deduce the conformational models for acetylated derivatives of $\alpha$-Neu5Ac ($\text{Neu4,5Ac}_2$, $\text{Neu5,7Ac}_2$, $\text{Neu5,8Ac}_2$, $\text{Neu5,9Ac}_2$) through MD simulation.

(v) To model the sialic acid and its acetylated derivatives at the binding site of VCN by rigid body dynamics.

(vi) To propose the binding specificity of VCN towards sialic acid and its acetylated based on atmistic level interactions between sugar molecules and the protein.

(vii) To investigate the binding specificity of influenza A hemagglutinins H1, H3, H5 and H9 towards the sialyl disaccharides Neu5Ac$\alpha$(2→3)Gal or Neu5Ac$\alpha$(2→6)Gal through molecular modeling and MD simulation.

(viii) To investigate the binding specificity of H1 towards the sialyldisaccharides Neu5$\alpha$(2→3)Gal, Neu5$\alpha$(2→6)Gal, NGc5$\alpha$(2→3)Gal, NGc$\alpha$(2→6)Gal, KDN$\alpha$(2→3)Gal and KDN$\alpha$(2→6)Gal by carrying out MD simulation on these complexes.