Section 1: Introduction

1 Glycobiology, Glycomics and (Bio)Informatics

Claus-Wilhelm von der Lieth

Formerly at the Central Spectroscopic Unit, Deutsches Krebsforschungszentrum
(German Cancer Research Center), 69120 Heidelberg, Germany

1.1 The Role of Carbohydrates in Life Sciences Research

Despite their nearly complete neglect in databases and ‘traditional’ bioinformatics projects, carbohydrates are the most abundant and structurally diverse biopolymers formed in nature. Historically, the chemistry, biochemistry, and biology of carbohydrates were very prominent areas of research over a long period of time during the beginning and the middle of the last century. However, during the initial phase of the development of molecular biology, focusing on DNA, RNA, and proteins, studies of carbohydrates lagged far behind. Among the main reasons for this were the inherent structural complexity of carbohydrates, the difficulty in easily determining their structure, the fact that their biosynthesis cannot be directly predicted from the DNA template, and that no methods are available to amplify complex carbohydrate sequences. The more recent development of a variety of new and highly sensitive analytical tools for exploring the structures of oligosaccharides and for producing larger amounts of pure complex carbohydrates has opened up a new frontier in molecular biology. The term glycobiology, which was introduced in the late 1980s [1], reflects the coming together of the traditional disciplines of carbohydrate chemistry and biochemistry, with modern understanding of the cellular and molecular biology of complex carbohydrates, which are often named glycans in this context. The more recently introduced term “glycomics” [2] describes an integrated systems approach to study structure–function relationships of complex carbohydrates – the glycome – produced by an organism such as human or mouse. The glycome can be described as the glycan complement of the cell or tissue as expressed by a genome at a certain time and location. It includes all types of glycoconjugates: glycoproteins, proteoglycans, glycolipids, peptidoglycans, lipopolysaccharides, and so on. The aim of glycomics projects is to create a cell-by-cell catalog of glycosyltransferase (GT) expression and detected glycan structures using high-throughput techniques such as DNA glycogene chips, glycan microarray screening and mass spectrometric (MS) glycan profiling, combined with efficient bioinformatics tools.

Until recently, the role of complex carbohydrates to function as carriers and/or mediators of biological information was a widely neglected and unexplored area in science. However,

Bioinformatics for Glycobiology and Glycomics: an Introduction Edited by Claus-Wilhelm von der Lieth, Thomas Lütteke and Martin Frank

with the awareness that the human genome encodes for a significantly smaller number of genes than was estimated from genomes of lower organisms such as yeast [3], it became obvious that each gene can be used in a variety of different ways depending on how it is regulated. Consequently, the study of post-translational protein modifications, which can alter the functions of proteins, came increasingly into scientific focus. Since then, with glycosylation being the most complex and most frequently occurring co- and post-translational modification, glycobiology research has attracted increasing attention.

About 70% of all sequences deposited in the SWISS-PROT [4] protein sequence databank include the potential *N*-glycosylation consensus sequence Asn–X–Ser/Thr (where X can be any amino acid except proline) and thus may be glycoproteins. However, it is well known that not all potential sites are actually glycosylated. Based on an analysis of well-annotated and characterized glycoproteins in SWISS-PROT, it was concluded that more than half of all proteins are glycosylated [5, 6]. However, this number should be regarded as a very crude estimation since this study was hampered by the paucity of reliable, experimentally determined, and carefully assigned glycosylation sites.

The glycans are exposed on the surface of biomolecules and cells. They form flexible, branched structures that can extend 30 Å or further into the solvent. With a molecular weight of up to 3 kDa each, the oligosaccharide groups of mammalian glycoproteins frequently make up a sizable proportion of the mass of a glycoprotein and can cover a large fraction of its surface. The carbohydrate moiety of "proteins" may amount to a few percent of the molecular weight, but can be as much as 90% in some cases. *O*-Linked mucin-type glycoproteins are usually large (more than 200 kDa) with attached *O*-glycan chains at a high density. As many as one in three amino acids may be glycosylated and 50–80% of the total mass is due to carbohydrates [7]. An analysis of the available three-dimensional structures of glycoproteins contained in the PDB [8] revealed that the glycan and the protein parts of glycoproteins behave like semi-independent moieties. This behavior has several important biological consequences:

- *N*-Glycans can be modified without appreciable effects on the protein. Every *N*-linked glycan is subject to extensive modifications. This allows cells to fine-tune the biophysical and biological properties of glycoproteins and to generate the microheterogeneity [9] that is so characteristic of glycoproteins.
- The semi-independent nature of glycans also allows cell types and cells in different stages of differentiation and transformation to imprint on their glycoprotein pool their own specific biochemical characteristics, and thus give their exposed surface a "corporate identity."
- This "corporate identity" [10] exposed on their surface makes cells recognizable to other cells in a multicellular environment. It allows self-recognition and provides a central theme in development, differentiation, physiology, and disease.

1.2 Glycogenes, Glycoenzymes and Glycan Biosynthesis

The biosynthesis of carbohydrates attached to proteins or to lipids – called glycoconjugates – is fundamentally different to the expression of proteins. Whereas the enzymes required for the translation of the genetic information into a polypeptide chain in the ribosome are always the same for all proteins and amino acids, the subsequent glycosylation is a non-template-driven process where dozens of different enzymes are involved in the synthesis of the sugar chains attached to proteins or lipids. Depending on which of these enzymes are expressed in the cell that synthesizes a glycoprotein, various different glycan chains can be attached to the protein or lipid. Glycoproteins generally exist as populations of glycosylated variants – called glycoforms – of a single polypeptide [11, 12]. Although the same glycosylation machinery is available to all proteins in a given cell, most glycoproteins emerge with a characteristic glycosylation pattern and heterogeneous populations of glycans at each glycosylation site.

Glucose and fructose are the major carbon and energy sources for organisms as diverse as yeast and human beings (see, e.g., [7]: Monosaccharide Metabolism chapter). Organisms can derive the other monosaccharides needed for glycoconjugate synthesis from these major suppliers. It is important to appreciate that not all of the biosynthetic pathways are equally active in all types of cells.

The biosynthesis of oligosaccharides is primarily determined by sequentially acting enzymes, the glycosyltransferases (GTs), which assemble monosaccharides into linear and branched sugar chains. For this purpose, the monosaccharides must be either imported into the cell or derived from other sugars within the cell. However, a common factor is that all glycoconjugate syntheses require activated sugar nucleotide donors. It has long been known that a nucleotide triphosphate such as uridine triphosphate (UTP) reacts with a glycosyl-1-P to form a high-energy donor sugar nucleotide that can participate in glycoconjugate synthesis. Once the sugar nucleotides have been synthesized in the cytosol (or, in the case of CMP-Neu5Ac, in the cell nucleus), they are topologically translocated, since most glycosylation occurs in the endoplasmic reticulum (ER) and Golgi apparatus. As the negative charge of the sugar nucleotides prevents them from simply diffusing across membranes into these compartments, eukaryotic cells have devised no-energy-requiring sugar nucleotide transporters that deliver sugar nucleotides into the lumen of these organelles [7].

1.2.1 Biosynthetic Pathways

In eukaryotes, more than 10 biosynthetic pathways that link glycans to proteins and lipids [13, 14] are known. The KEGG PATHWAY resource [15, 16] – a collection of pathway maps representing current biochemical knowledge of the molecular interaction and reaction networks – has encoded 18 pathways for the biosynthesis of complex carbohydrates and their metabolism (see Figure 1.1), and 20 pathways for metabolism where carbohydrates are involved. More than 200 enzymes are involved in the biosynthesis of carbohydrate structures found on proteins and lipids. More than 30 different enzymes may participate directly in the synthesis of a single glycan. One of the best-characterized pathways is the biosynthesis of complex oligosaccharides that are subsequently attached to a protein through the side-chain nitrogen atom of the amino acid aspagarine (Asn) to give glycoproteins [10, 17, 18] (described in Section 8.1 in Chapter 8). Glycosylation of proteins occurs in all eukaryotes and in many archaea but only exceptionally in bacteria.

O-Linked glycosylation, where carbohydrates are attached to serine (Ser) and threonine (Thr), takes place post-translationally in the Golgi apparatus. The monosaccharides are added one by one in a stepwise series of reactions (Figure 1.2). This is in contrast to the *N*-linked glycosylation pathway where a preformed oligosaccharide is transferred *en bloc* to Asn. A second important difference is that there are no known consensus sequence

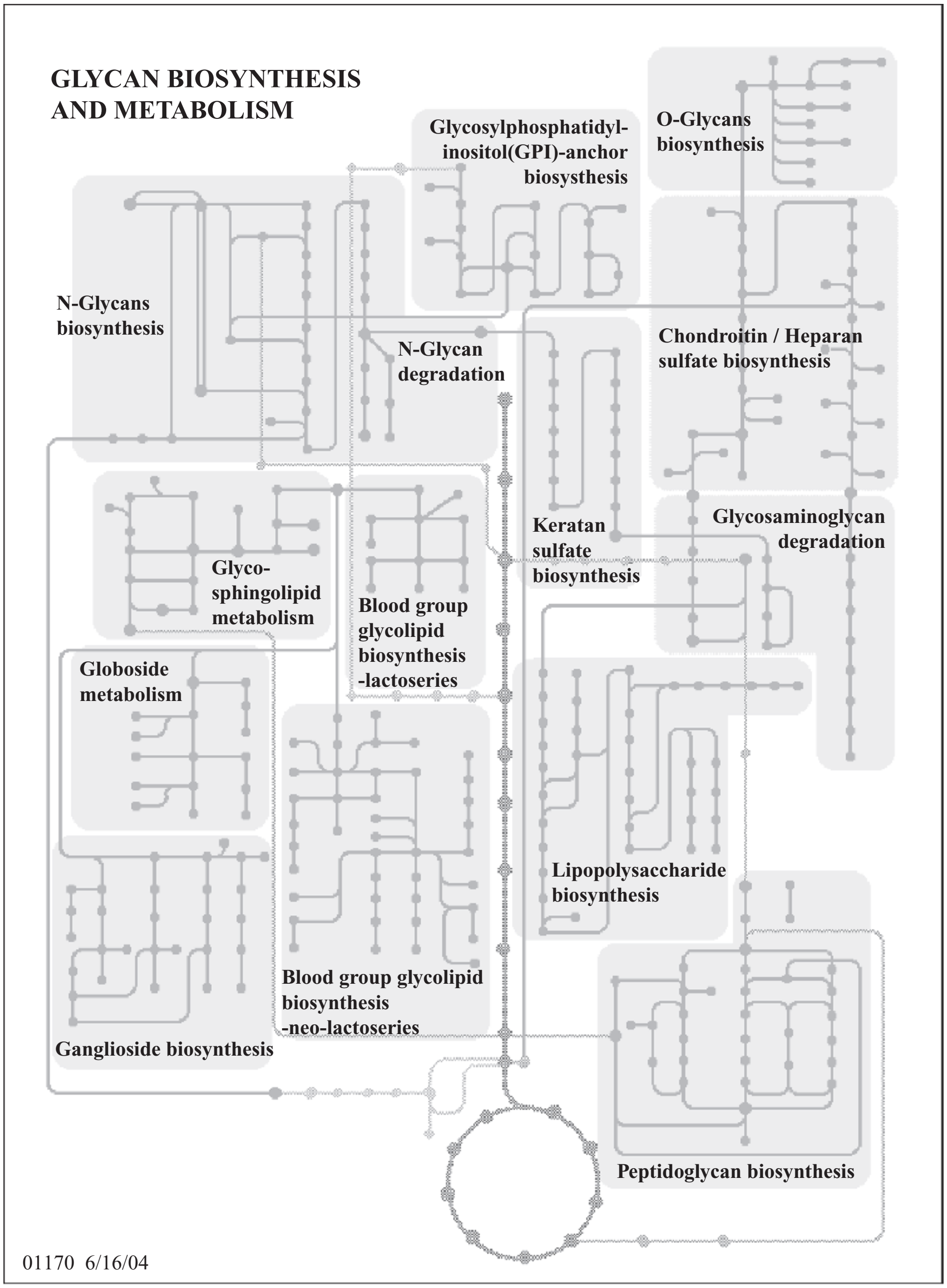
GLYCAN BIOSYNTHESIS AND METABOLISM
Glycosylphosphatidyl-inositol(GPI)-anchor biosysthesis
O-Glycans biosynthesis
N-Glycans biosynthesis
N-Glycan degradation
Chondroitin / Heparan sulfate biosynthesis
Keratan sulfate biosynthesis
Glycosaminoglycan degradation
Glyco-sphingolipid metabolism
Blood group glycolipid biosynthesis -lactoseries
Globoside metabolism
Lipopolysaccharide biosynthesis
Blood group glycolipid biosynthesis -neo-lactoseries
Ganglioside biosynthesis
Peptidoglycan biosynthesis
01170 6/16/04

Figure 1.1 Illustration of the pathways for the biosynthesis of complex carbohydrates and their metabolism encoded in KEGG PATHWAY [15, 16] available at: www.genome.jp/kegg/pathway/map/map01170.html.

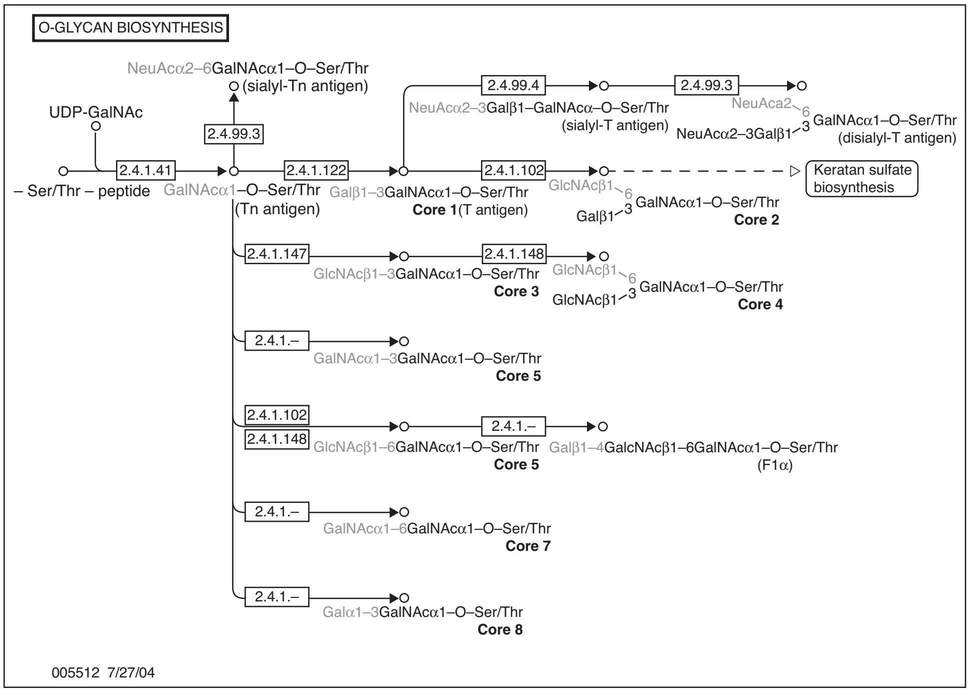

Figure 1.2 Known biosynthesis pathways for carbohydrates attached to the oxygen atom of the side chain of the amino acids serine or threonine as encoded in the KEGG PATHWAY resource [15, 16] (www.genome.jp/kegg/pathway/map/ map01170.html). An IUPAC like nomenclature (see Chapter 3) is used to characterize the monosaccharides and linkages. The enzymes are given in the square boxes by their corresponding Enzyme Commission (EC) numbers, which are based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB).

motifs that define an *O*-linked glycosylation site analogous to the Asn–X–Ser/Thr motif for *N*-linked glycosylation.

1.2.2 The Role of Bioinformatics in Identifying Glyco-related Genes

The enzymes required for the biosynthesis of complex carbohydrates can be classified into those needed for the conversion of monosaccharide building blocks to activated sugar nucleotides and their transport within the cell, and those which are used to build (glycosyltransferases) and remodel (glycosidases) glycoconjugates [19]. Many, but not all, of the latter enzymes are found within the ER–Golgi pathway for export of newly synthesized glycoconjugates.

The first mammalian GT gene was reported in 1986 [20]. The progress in identifying new GT genes at that time was slow because they had to be cloned by identifying the partial amino acid sequence of the purified enzyme, which was the limiting step. Thereafter, from the beginning of the 1990s when methods of expression cloning and PCR cloning with degenerated primers were employed, several novel GT genes were detected each year. It became obvious that GTs can be classified into several subfamilies which contain

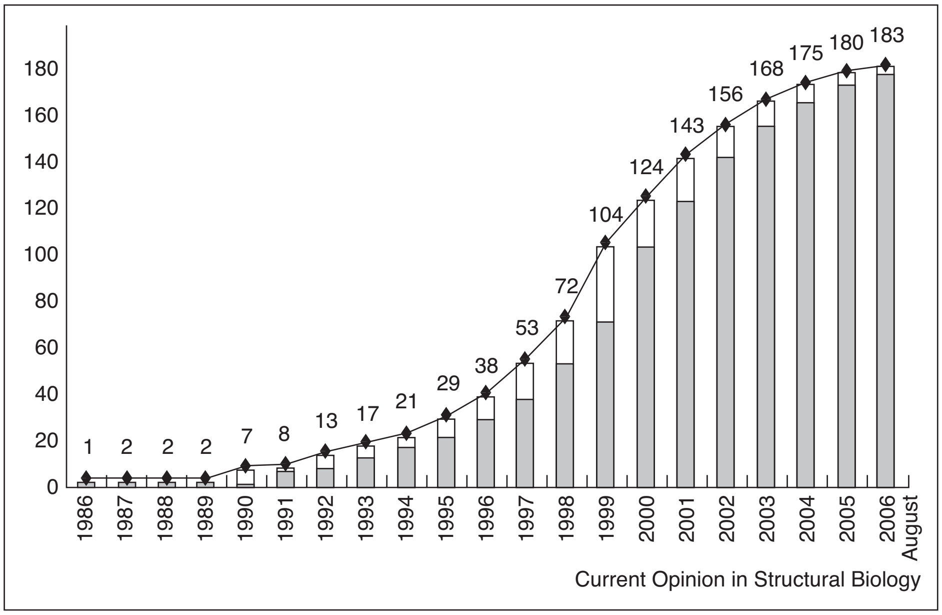

Figure 1.3 Progress in the cloning of glycogenes (including GTs, sulfotransferases, and sugar–nucleotide transporters). Filled columns indicate the cumulative number of glyco-related enzymes reported during the past two decades. Open columns indicate the number of novel enzymes reported in each year. Reprinted from [24] with permission from Elsevier.

well-conserved sequence motifs. Based on this knowledge and the increasing availability of gene sequences and the development of appropriate bioinformatics searching algorithms, the *in silico* identification of GT genes could be successfully applied [21, 22]. During the middle of the 1990s, the number of newly reported GT genes began to increase significantly, reaching a peak in 1999 (Figure 1.3). This was due to the substantial increase in sequenced genes and the ease of finding new GT genes by homology searching using well-known BLASTN searches. The number of newly identified GT genes began to decrease gradually after 1999 to only five by August 2006, suggesting that mammalian GT gene cloning seems to be approaching its completion. During the past two decades, more than 180 human glycogenes have been cloned and their substrate specificities analyzed using biochemical approaches [23, 24]. The current status of knowledge compiled for these human GT genes and their links with orthologous genes in other species is summarized in the GlycoGene database [22].

As demonstrated for the identification of GT genes, the application of classical bioinformatics tools and also the use of genomic databases had and will continue to have a significant impact on the rapid development of glycobiology research [25]. The same is true when searching for all lectins with similar binding affinity for a specific carbohydrate, which was also significantly accelerated through systematic analysis of gene sequences for the corresponding sequence motifs [26–29].

However, the use of (bio)informatics in glycobiology research has to be divided between those applications where an explicit description of the glycan structure is required, and those where the proteins to which carbohydrates are attached, the enzymes which build and modify carbohydrates, or the lectins which recognize a certain sugar epitope, are analyzed. The latter type of applications can be performed using well-known bioinformatics tools such as sequence alignment techniques and attempts to understand the evolutionary relationships through phylogenetic analysis. Where an encoding of the carbohydrate structure is required, however, for example when looking at carbohydrate specificity of a lectin or classification of the glycome of an organism, classical bioinformatics approaches cannot be directly applied.

1.3 Intrinsic Problems of Glycobiology Research

Glycobiologists have to deal with several intrinsic problems, making their research difficult and time consuming, as well as ambitious.

1.3.1 Carbohydrates Have to Be Analyzed at Physiological Concentrations

The first major challenge is to develop highly sensitive analytical methods. Since the biosynthesis of complex carbohydrates requires a variety of enzymes, which have to act in a defined and consecutive way, there are currently no methods available to amplify glycans readily in the sense that DNA is amplified using polymerase chain reaction (PCR) techniques. Consequently, highly sensitive analytical methods have to be applied, which are able to detect the small amounts of material found in cells. The chapters on experimental methods will discuss the central analytical methods – mass spectrometry, HPLC and NMR – which are used in different areas of glycobiology to identify glycan structures.

1.3.2 Complexity of Glycan Structures

The second major challenge lies in the complexity of glycan structures: each pair of monosaccharide residues can be linked in several ways, and one residue can be connected to three or four others (giving branched structures). The information content which can be potentially encoded by glycans in a given sequence is therefore high. The four nucleotides in DNA can be combined to give 256 four-unit structures, and the 20 amino acids in proteins yield 160 000 four-unit configurations. However, the number of naturally occurring residues is much larger for glycans which have the potential to assemble into more than 15 million four-unit arrangements.

Although oligosaccharides potentially carry this high capacity to store biological information, only a small part thereof is actually used in nature. A recent analysis of the KEGG glycan database [15] containing 4107 unique glycan entries [30], which consist of nine frequently occurring monosaccharides (glucose, galactose, mannose, *N*-acetylglucosamine, *N*-acetylgalactosamine, fucose, xylose, glucuronic acid, and sialic acid) showed, that only 302 (54%) of the 558 (nine monosaccharides, two anomers, 31 substitution possibilities) theoretically possible disaccharides appear in the database. Furthermore, while an enormous number of reaction pattern combinations are theoretically possible, only 2178 of these combinations actually appear in the database. These numbers suggest that the structural diversity of glycans is indeed large, but that the combination of reaction patterns which actually exist in a given cellular environment is limited by the availability of the glyco-related enzymes which build and modify the glycan structures.

1.3.3 Structural Heterogeneity

The third major challenge is the structural heterogeneity and "fuzziness" of glycans. Glycoproteins normally exhibit various glycoforms when isolated from cells and tissues [12, 31, 32]. Often several tens of different glycoforms for a given glycosylation site have been identified. Analytical techniques and also databases and bioinformatics applications have to cope with this phenomenon. Non-stoichiometric modifications to position and amount of chemical substitutions are another unique feature of complex carbohydrates, which requires the development of new concepts to analyze, encode, and handle, for example, the statistical occurrence of sulfate groups at specific positions in glycosaminoglycans such as heparin and heparan sulfate [33, 34].

1.3.4 Multivalent Interactions with Proteins

Glycan-binding proteins mediate diverse aspects of cell biology, including pathogen recognition of host cells, cell trafficking, endocytosis, and modulation of cell signaling [35]. However, the assignment of biological function to carbohydrates in recognition events is complex because individual glycan structures exhibit only very weak interactions with a protein surface. For example, the binding affinity of monovalent oligosaccharide ligands to their respective viral receptors is rather weak, with dissociation constants (K_d) of around 10^{-3}–10^{-4} M. This low affinity is in strong contrast to the high K_d values of 10^{-8}–10^{-12} M determined for the binding of complete virions to cell surfaces [36]. It is widely assumed that this high affinity is contributed to by multivalent binding of the repetitive virion surface carbohydrate-recognizing proteins/receptors to repetitive oligosaccharide structures on the cell surface. Unlike protein–protein interactions, which can be generally viewed as "digital" in regulating function, glycan–protein interactions impinge on biological functions in a more "analog" fashion that can in turn "fine-tune" a biological response. This fine-tuning by glycans is achieved through the graded affinity, avidity, and multivalency of their interactions.

1.3.5 New Insights Through Highly Sensitive Analytical Techniques

Much of the increase in a better understanding of the versatile regulatory role of glycans in life can be credited to improvements in existing, and the development of new, highly sensitive analytical techniques. The details of the current status of the analytical techniques will be discussed in detail in the chapters on experimental methods. Here, an especially impressive example will be briefly described, where the combination of modern biomolecular and analytical techniques was used to provide detailed insights into the molecular basis of the receptor specificity of the 1918 so-called Spanish flu.

It is well known that infection with viruses and bacteria often starts with specific interactions with glycans on the surfaces of host cells. For example, the host specificity of influenza A virus infection is mediated by the viral surface glycoprotein hemagglutinin (HA), which binds to host-cell receptors containing glycans with terminal sialic acids.

The impact of influenza infection is felt globally each year, as this disease develops in approximately 20% of the world's population. The 1918 "Spanish" influenza pandemic represents the largest recorded outbreak of any infectious disease, causing about 20 million deaths. At the end of the 1990s, an American research team was able to detect fragments of the viral genome in lung samples taken from the body of an Inuit woman victim of the pandemic buried in the Alaskan tundra and a number of preserved samples taken from American soldiers of the First World War. Using modern biomolecular amplification techniques, the entire coding sequence of 1701 nucleotides for the viral surface HA was amplified in 22 overlapping fragments such that the sequences for matching primers could be confirmed [37].

The HA of influenza virus mediates receptor binding and membrane fusion, the first stages of virus infection. The sequences found for the 1918 HA did not reveal any characteristics that were obviously responsible for the extreme virulence of the 1918 pandemic. Independently, two research groups succeeded in growing crystals of the 1918 HA and analyzed its binding properties [38, 39]. The carbohydrate recognition of influenza virus HA is highly specific: whereas human viruses infect epithelial cells in the lungs and upper respiratory tract which have α2,6-linked sialic acids on their surfaces, avian viruses preferentially bind to α2,3-linked sialic acids [40]. This slight structural difference in the recognized sugar epitope obviously prevents a spread of the influenza virus infection across species. Analysis of the binding specificity of the highly virulent 1918 influenza virus HA using the glycan array of the US Consortium for Functional Glycomics (CFG) revealed a clear preference for α2,6-linked sialylgalactose motifs [41]. Glycan microarrays are a relatively new and highly specific technology that allows rapid determination of glycan-binding protein interactions and specificities.

Subsequently it was shown [42] that a single amino acid substitution in the 1918 human influenza virus HA – Asp225 to Gly – changes receptor binding specificity from an HA which preferentially binds to the human α2,6-sialylgalactose motif to one which binds both the human α2,6- and the α2,3-sialylgalactose motif of the avian cellular receptors. Mutation of a further single amino acid back to the avian consensus – Asp190 to Glu – resulted in a preference for the avian receptor. Thus, the species barrier, as defined by the receptor specificity preferences, of 1918 human viruses compared with likely avian virus progenitors, can be circumvented by changes at only two positions in the HA receptor binding site.

A combination of highly sophisticated new techniques revealed that the HA of the 1918 influenza virus might be more like that found in avian influenza than was previously thought. Usually, avian influenza strains do not affect humans directly because bird-adapted HA proteins are not able to bind well to human receptors. Until very recently, it was thought that to make the leap to humans successfully a bird strain must pass through an intermediate animal that contains both bird and human receptors, such as a pig. The new findings suggest that minimal changes in the receptor binding domain of an avian HA may have been enough to broaden its binding targets to include the major sialic acid receptor expressed on human respiratory epithelium.

Modern molecular biology techniques and highly sensitive analytical tools have helped, 80 years after its outbreak, to give some new insights into why the 1918 influenza virus was so devastating. Additionally, glycan microarray technology has been proven to have the ability to detect rapidly strains which have the potential capability to cross species barriers, a major goal for worldwide influenza surveillance [43].

1.4 Carbohydrates as a New Frontier in Pharmaceutical Research

Except for sulfated glycan heparin [44], which belongs to the class of glycosaminoglycans (GAGs), synthetic carbohydrates have not been widely used as therapeutics. One obvious reason is that complex carbohydrates are difficult to synthesize. The recent development of a (semi-)automated oligosaccharide synthesizer greatly accelerates the assembly of complex, naturally occurring carbohydrates and also chemically modified oligosaccharide structures (mimetics) and promises to have major impact on the field of glycobiology [45, 46]. Synthetic carbohydrates and glycoconjugates will be more readily available for broad use, and will advance the study of their roles in biologically important processes such as inflammation, cell–cell recognition, immunological response, metastasis, and fertilization. Tools such as microarrays, surface plasmon resonance spectroscopy, and fluorescent carbohydrate conjugates to map interactions of carbohydrates in biological systems are available [47–49] and can be used to evaluate systematically the binding specificity and strength of naturally occurring carbohydrates and also mimetics thereof.

1.4.1 Carbohydrates in Drug and Vaccine Development

Bacteria, viruses, and parasites are the major agents leading to disease. All cells in nature are covered with a dense and complex coat of glycans. A wide variety of pathogens initiate infection by binding to the surface glycans of host cells. This is not surprising as cell-surface glycans are the first molecules encountered by pathogens when they contact potential host cells or their secretions. Outer, terminal glycan sequences such as those carrying sialic acid residues are even more likely to be preferred targets, as they are the first residues that pathogens encounter. Examples of disease in which cell-surface glycan recognition is involved include influenza virus infection of the lung and upper respiratory tract, erythrocyte invasion by the malaria parasite *Plasmodium falciparum*, *Helicobacter pylori* infection of the stomach, and intestinal diarrhea caused by the toxin of *Vibrio cholerae*.

In the case of influenza virus, as described above, infection is mediated by the viral surface glycoprotein hemagglutinin which binds to host-cell receptors containing glycans with terminal sialic acids. Surface binding is followed by penetration of the cellular membrane. Complex glycans are involved in cellular adhesion, internalization, and the release of newly formed virus particles, all of which are of high interest for preventive medicine and drug design. Highly potent inhibitors of the viral enzyme neuraminidase, which facilitates release of progeny influenza virus from infected host cells, have been designed with the help of computational chemistry methods using 3D structures of the enzyme. The neuraminidase inhibitors mimic the form of sialic acid seen in the transition state of the enzyme reaction, the cleavage of terminal sialic acid residues from glycans. Neuraminidase inhibitors have been shown to be effective against all neuraminidase subtypes and, therefore, against all strains of influenza, a key point in epidemic and pandemic preparedness. These new drugs have great potential for diminishing the effects of influenza infection [50, 51].

Glycoconjugate vaccines provide effective prophylaxis against bacterial infections. However, only a few vaccines have been developed by chemical synthesis of the key carbohydrate antigens. In Cuba, it was demonstrated that a conjugate vaccine composed of a synthetic capsular polysaccharide antigen of *Haemophilus influenzae* type b (Hib) elicited long-term protective antibody titers [52, 53]. This demonstrates that access to synthetic complex carbohydrate-based vaccines is feasible and provides a basis for further development of similar approaches for other human pathogens [54]. Hib was the leading cause of bacterial meningitis in many parts of the world before the introduction of conjugate vaccines. The use of vaccines against Hib in developing countries is expected to be an important tool for the reduction of vaccine-preventable morbidity and mortality among children less than 5 years old.

About 40% of the world's population live with the risk of contracting malaria. Although only about 1% of all malaria cases are lethal, malaria continues to claim the lives of over two million people annually. No viable vaccine candidate has been developed for malaria. Glycosylphosphatidylinositol (GPI) anchors are a class of naturally occurring glycolipids that link proteins and glycoproteins via their C-terminus to cell membranes. The malarial parasite *Plasmodium falciparum* expresses GPI in protein anchored and free form on the cell surface: the GPI constitutes a toxin which is implicated in the pathogenesis and fatalities of malaria in humans [55]. Recently, it could be demonstrated that mice vaccinated with a synthetic GPI glycan conjugated to a carrier protein produced anti-GPI antibodies and had a greatly improved chance of survival upon infection with *P. falciparum*. Between 60 and 75% of vaccinated mice survived, compared with 0–9% of sham-immunized mice. The parasite levels observed in the blood of the vaccine and control groups did not differ significantly, thus indicating that the synthetic GPI glycan conjugate serves as an anti-toxin vaccine [56].

1.4.2 Carbohydrates Play a Key Role in Many Diseases

Many diseases are caused by disruption of regulatory and control mechanisms within a particular organism. For example, a DNA point mutation may result into a single amino acid replacement in a protein, which may completely change or obliterate the function of the protein. Such mutations may occur in somatic cells of adult individuals, or they may be inherited, resulting in inborn defects, such as congenital disorders of glycosylation (CDGs) – defects in glycan biosynthesis, lysosomal storage diseases – defective glycan catabolism, and von Willebrand disease. Cancer and some autoimmune diseases, such as rheumatism, are other examples of diseases caused by failure of the organism's regulation and control system. Cancer is associated with changes in glycosylation of proteins exposed on the outer cell surface. Therefore, monitoring of temporal changes in glycosylation has potential as a diagnostic tool and as a prognostic indicator. Furthermore, cancer cell-specific complex glycans may also serve as targets for tissue- or cell-selective delivery of agents that can kill tumor cells.

In addition to the effects of altered glycan biosynthesis/catabolism in disease, complex carbohydrate epitopes play key roles in allergy and immune reactions against parasites. They are also of great significance in xeno-transplantation, where species-specific carbohydrate structures can be recognized as non-self and promote tissue rejection. On the other hand, synthetic manipulation of glycosylation patterns is being used to advantage in the biotechnological production of recombinant therapeutic glycoproteins; for example, an increase in the number of sialylated glycans on erythropoietin (EPO) increases its serum half-life [57].

An emerging area of research is so-called metabolic oligosaccharide engineering, the goal of which is a biosynthetically altered cell-surface repertoire through the introduction of unnatural sugar residues into cellular glycans. Such engineered cell surfaces are extremely useful systems for studying biochemistry and cell biology in a broad range of contexts, such as cell–cell interactions.

1.5 A Short History of Databases and Informatics for Glycobiology

It can be expected that the rapid evolution of glycomics, including glycan array technologies, will result in very large data collections that will have to be organized, analyzed, and compared, requiring standards for structural representation. The development and use of informatics tools and databases for glycobiology and glycomics research has increased considerably in recent years; however, it can still be considered as being in its infancy when compared with the genomics and proteomics areas. The intrinsic factors which make the development of informatics for glycobiology and glycomics a challenging task have been described above. However, there is a general consensus within the community of glycoscientists that the availability of comprehensive and up-to-date carbohydrate databases, and also efficient software to retrieve and handle the data, will be a prerequisite for successfully conducting large-scale glycomics projects aimed at deciphering new, so far unknown, biological functions of glycans. Here, a short overview of the history of databases and informatics for glycobiology will be given.

1.5.1 The Early Days: CarbBank

Before information technologies were available, it was a rather time-consuming task to cope with all structures of complex carbohydrates detected in nature, which were published in various journals using different ways to describe structural details. Normally, only a few specialists in the field could successfully access the available knowledge. When digital documentation systems and search engines were introduced into science during the 1980s, it was recognized that this new technology could also be very useful for encoding and retrieving all published glycan structures using a language which was well understood by glycoscientists. In light of this, the Complex Carbohydrate Structure Database (CCSD) [58, 59] – often referred to as *CarbBank* according to the retrieval software used to access the data – was established in the mid-1980s, the main purpose of which was to allow the user to find easily all publications in which specific carbohydrate structures were reported. The CCSD was developed and maintained by the Complex Carbohydrate Research Center of the University of Georgia (USA) and funded by the National Institutes of Health (NIH). The need to develop CarbBank as an international effort was clearly recognized and resulted in worldwide curation teams responsible for specific classes of glycans. During the 1990s, a Dutch group assigned NMR spectra to CCSD entries (*SugaBase*) [60, 61]. This was the first attempt to create a carbohydrate NMR database that complemented CCSD entries with proton and carbon chemical shift values.

For a variety of reasons, including disagreement on the best ways to integrate the CCSD into the growing bioinformatics environment and the need to provide more user-friendly tools compatible with new, Internet-based approaches, the funding for the CCSD stopped during the second half of the 1990s. In a letter sent to the provider of CarbBank in 1998, I wrote concerning the infrequent use of the database: "*One rather obvious reason for this situation is that carbohydrate data collections only rarely exhibit cross-referencing to other available data on the net. CarbBank uses efficient algorithms to provide rapid access to references following the input of a query expressed in terms of the carbohydrate nomenclature. Unfortunately, only pure bibliographic information such as authors, journal, and title are displayed, but not abstracts. Using modern Web techniques, it should be rather straightforward to send a request to the public WEB-Medline (PubMed) and provide elegant access to abstracts. One of the big disadvantages of essentially all carbohydrate Web applications is that there are no annotated and/or cross-referenced implementations, which allow glycoscientists to find important data for the compound of interest in a compact and well-structured representation. Most carbohydrate applications on the Web are designed to answer just one special question*."

Unfortunately, CarbBank was not developed further and, beyond 1996, the CCSD was no longer updated. An attempt to transfer responsibility for updating the CCSD to a volunteer team of glycoscientists around the world obviously failed. Nevertheless, with 49 897 entries, which correspond to 23 118 distinct glycan structures, the CCSD is still the largest publicly available repository of glycan-related data. All subsequent open access projects initiated at the beginning of the new century made use of the CCSD data.

1.5.2 Beyond CarbBank

The collapse of CarbBank was extremely frustrating, especially for those who were involved in this international venture. There was very little support for renewal of the project, as the bioinformatics field – concentrating at that time on the sequencing of the human genome – completely ignored the potential of carbohydrates as a repository of biological information.

A small informatics oriented group of scientists at the DKFZ (German Cancer Research Center) in Heidelberg, initially interested in elucidating the conformational space of complex carbohydrates, first put forward the imperative to develop informatics for glycobiology as an independent sub-branch of bioinformatics. This group also realized the need to make the CCSD entries publicly available using modern Internet-based tools and to cross-reference the glyco-related data with proteomics and glycomics information. These ideas led to the development of the *GLYCOSCIENCES.de* [62, 63] portal and the *EUROCarbDB* (www.eurocarbdb.org) project.

At the beginning of the new century, when the gap between encoded and published glycan structures became obvious, several companies started to provide commercial access to glyco-related data, which they extracted from the literature. However, due to limited commercial success, most of these services stopped. The Australian GlycoSuite [64], the only one of these services that survives today, is willing to provide academic users with free access to the data they have extracted from literature.

1.5.3 Glycomics Initiatives – a New Stimulus for Glycoinformatics Development

An important stimulation for glycoinformatics development was the establishment of the Consortium for Functional Glycomics (www.functionalglycomics.org) in 2001. This was the first large-scale project that clearly emphasized the need for informatics to manage and annotate automatically the vast amount of experimental data generated by glycomics research. The development of algorithms for the automatic interpretation of mass spectra – a severe bottleneck that hampers the rapid and reliable interpretation of MS data in high-throughput glycomics projects – is critical for all glycomics projects [65]. This is still the most active area of software development, where various primarily experimentally oriented groups have been developing software solutions and algorithms to solve their specific scientific questions.

Another important step was the integration of glyco-related biological pathways into the schemata of the first 'classical' bioinformatics initiative – the Kyoto Encyclopedia of Genes and Genomes (KEGG). Subsequent development of associated databases for glycan structures led to the KEGG GLYCAN [16, 66] approach, which elegantly established the connection between glycan structures and the knowledge of enzymatic reactions to build the glycan structures. Additionally, the KEGG group made significant progress in applying bioinformatics algorithms to the tree-like structures of glycans for comparison and alignment, to develop similarity scores, and to establish a global view of all glycans belonging to related pathways (see also Chapter 7).

As a consequence of the increasing interest in glycomics research, various new databases were started in recent years (for examples, see the link list at www.eurocarbdb.org/links/). Among these, the EUROCarbDB project (a distributed bottom to top initiative for primary experimental data), the Russian Bacterial Carbohydrate Structure Database (aiming to cover all known structures produced in bacteria), and the *Bioinformatics for Glycan Expression* initiative (development of glyco-related ontologies) of the Complex Carbohydrate Research Center are the largest ones. In general, the development of glyco-related tools and databases can be described as a small but fairly active field of research.

1.5.4 The current situation

The current situation in glycoinformatics is characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources, and specific applications, managed by various consortia, institutions, or local groups. These resources rarely provide communication mechanisms that would permit the widest advantage to be taken of these data by allowing their combination and comparison. However, approaches to link the distributed data have been conceptually worked out and examples are already being implemented. The collaborative spirit recently exhibited by all of the major glycomics initiatives will significantly help to overcome the current unfavorable situation. This positive spirit has recently led to an important milestone, the agreement of an XML standard format for the exchange of glycan structures (GLYDE-II) [67].

None of the existing initiatives has the capacity to fulfill completely the goal of CarbBank at the beginning of the 1990s, that is, to provide comprehensive access to all published carbohydrate structures. In particular, the existing initiatives do not have the worldwide resources to fill the gap of published glycan structures that were not included in CarbBank after its termination in the mid-1990s.

It is likely that the tendency to set up local databases designed to support specific areas of research in glycobiology will continue in the near future. The existence of a centralized glycan structure database would substantially increase the ability to annotate and cross-reference local data with other bioinformatics resources. Offering clear guidelines describing the minimal requirements of data exchange formats, which are required for databases to communicate with each other, will hopefully lead to strong interconnections and compatibility among glycobiology and glycomics databases.

1.5.5 The Future

It is clear that there is an urgent need to develop databases and informatics for glycobiology and glycomics. The developments in glycomics will produce an enormous amount of data and there is a need to cut across multiple datasets to understand fully the structure–function relationships of complex carbohydrates. A critical component that will facilitate this process is a bioinformatics platform to store, integrate, and process the recorded data, to condense them to information and knowledge. Several statements of international scientific institutions underpin this direction:

- The European Science Foundation has published a statement *Structural Medicine: The Importance of Glycomics for Health and Disease* (see www.eurocarbdb.org), which emphasizes the need to develop glyco-related databases further.
- In September 2006, the NIH organized a workshop, *Frontiers in Glycomics*. The workshop was the largest meeting focused on the development of databases and informatics for glycomics and glycobiology. A white paper was compiled which set priorities for the most important steps to develop the field [67].
- The outcome of this meeting was, on the one hand, an agreement to accept a standard exchange format for glycan structures called GLYDE-II, and on the other, a list of the most urgent needs – top priority is a centralized, comprehensive, and highly curated carbohydrate structure database.
- The European Strategy Forum for Research Infrastructures (http://cordis.europa.eu/esfri/) published a roadmap emphasizing that "*modern science is inconceivable without recourse to well structured, continuously upgraded (. . .) and freely accessible databases (. . .). The bioinformatics infrastructure (. . .) will continue to expand, requiring successive investments for major upgrades, and will remain the depository of biological information for as long as we now can foresee.*"

References

1. Rademacher TW, Parekh RB, Dwek RA: Glycobiology. *Annu Rev Biochem* 1988, **57**:785–838.
2. Hirabayashi J, Arata Y, Kasai K: Glycome project: concept, strategy and preliminary application to *Caenorhabditis elegans*. *Proteomics* 2001, **1**:295–303.
3. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004, **431**:931–945.
4. Apweiler R, Bairoch A, Wu CH: Protein sequence databases. *Curr Opin Chem Biol* 2004, **8**:76–80.
5. Apweiler R, Hermjakob H, Sharon N: On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1999, **1473**:4–8.
6. Ben-Dor S, Esterman N, Rubin E, Sharon N: Biases and complex patterns in the residues flanking protein *N*-glycosylation sites. *Glycobiology* 2004, **14**:95–101.
7. Varki A, Cummings R, Esko J, Freeze H, Hart G, Marth J: *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
8. Petrescu A-J, Milac A-L, Petrescu SM, Dwek RA, Wormald MR: Statistical analysis of the protein environment of *N*-glycosylation sites: implications for occupancy, structure and folding. *Glycobiology* 2004, **14**:103–114.
9. Rudd PM, Wormald MR, Stanfield RL, Huang M, Mattsson N, Speir JA, DiGennaro JA, Fetrow JS, Dwek RA, Wilson IA: Roles for glycosylation of cell surface receptors involved in cellular immune recognition. *J Mol Biol* 1999, **293**:351–366.
10. Helenius A, Aebi M: Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev Biochem* 2004, **73**:1019–1049.
11. Haslam SM, North SJ, Dell A: Mass spectrometric analysis of *N*- and *O*-glycosylation of tissues and cells. *Curr Opin Struct Biol* 2006, **16**:584–591.
12. Rudd PM, Dwek RA: Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol* 1997, **32**:1–100.
13. Spiro RG: Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology* 2002, **12**:43R–56R.
14. Freeze HH: Genetic defects in the human glycome. *Nat Rev Genet* 2006, **7**:537–551.

15. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, **32**:D277–D280.
16. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, **34**:D354–D357.
17. Kornfeld R, Kornfeld S: Assembly of asparagine-linked oligosaccharides. *Annu Rev Biochem* 1985, **54**:631–664.
18. Helenius A, Aebi M: Intracellular functions of N-linked glycans. *Science* 2001, **291**:2364–2369.
19. Taniguchi N, Miyoshi E, Jianguo G, Honke K, Matsumoto A: Decoding sugar functions by identifying target glycoproteins. *Curr Opin Struct Biol* 2006, **16**:561–566.
20. Narimatsu H, Sinha S, Brew K, Okayama H, Qasba P: Cloning and sequencing of cDNA of bovine *N*-acetylglucosamine (beta 1–4)galactosyltransferase. *Proc Natl Acad Sci USA* 1986, **83**:4720–4724.
21. Kikuchi N, Kwon YD, Gotoh M, Narimatsu H: Comparison of glycosyltransferase families using the profile hidden Markov model. *Biochem Biophys Res Commun* 2003, **310**:574–579.
22. Kikuchi N, Narimatsu H: Bioinformatics for comprehensive finding and analysis of glycosyltransferases. *Biochim Biophys Acta* 2006, **1760**:578–583.
23. Narimatsu H: Construction of a human glycogene library and comprehensive functional analysis. *Glycoconj J* 2004, **21**:17–24.
24. Narimatsu H: Human glycogene cloning: focus on beta 3-glycosyltransferase and beta 4-glycosyltransferase families. *Curr Opin Struct Biol* 2006, **16**:567–575.
25. Schachter H: Protein glycosylation lessons from *Caenorhabditis elegans*. *Curr Opin Struct Biol* 2004, **14**:607–616.
26. Drickamer K, Taylor ME: Identification of lectins from genomic sequence data. *Methods Enzymol* 2003, **362**:560–567.
27. Drickamer K, Fadden AJ: Genomic analysis of C-type lectins. *Biochem Soc Symp* 2002, **69**:59–72.
28. Houzelstein D, Goncalves IR, Fadden AJ, Sidhu SS, Cooper DN, Drickamer K, Leffler H, Poirier F: Phylogenetic analysis of the vertebrate galectin family. *Mol Biol Evol* 2004, **21**:1177–1187.
29. Amado M, Almeida R, Schwientek T, Clausen H: Identification and characterization of large galactosyltransferase gene families: galactosyltransferases for all functions. *Biochim Biophys Acta* 1999, **1473**:35–53.
30. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M: Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 2005, **21**:3976–3982.
31. Chalabi S, Panico M, Sutton-Smith M, Haslam SM, Patankar MS, Lattanzio FA, Morris HR, Clark GF, Dell A: Differential *O*-glycosylation of a conserved domain expressed in murine and human ZP3. *Biochemistry* 2006, **45**:637–647.
32. Arnold JN, Wormald MR, Sim RB, Rudd PM, Dwek RA: The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol* 2007, **25**:21–50.
33. Coombe DR, Kett WC: Heparan sulfate–protein interactions: therapeutic potential through structure-function insights. *Cell Mol Life Sci* 2005, **62**:410–424.
34. Capila I, Linhardt RJ: Heparin–protein interactions. *Angew Chem Int Ed* 2002, **41**:390–412.
35. Collins BE, Paulson JC: Cell surface biology mediated by low affinity multivalent protein–glycan interactions. *Curr Opin Chem Biol* 2004, **8**:617–625.
36. Herrmann M, von der Lieth CW, Stehling P, Reutter W, Pawlita M: Consequences of a subtle sialic acid modification on the murine polyomavirus receptor. *J Virol* 1997, **71**:5922–5931.
37. Reid AH, Fanning TG, Hultin JV, Taubenberger JK: Origin and evolution of the 1918 “Spanish” influenza virus hemagglutinin gene. *Proc Natl Acad Sci USA* 1999, **96**:1651–1656.
38. Stevens J, Corper AL, Basler CF, Taubenberger JK, Palese P, Wilson IA: Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* 2004, **303**:1866–1870.

39. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, Vasisht N, Steinhauer DA, Daniels RS, Elliot A, Wiley DC, Skehel JJ: The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science* 2004, **303**:1838–1842.
40. Connor RJ, Kawaoka Y, Webster RG, Paulson JC: Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. *Virology* 1994, **205**:17–23.
41. Stevens J, Blixt O, Glaser L, Taubenberger JK, Palese P, Paulson JC, Wilson IA: Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *J Mol Biol* 2006, **335**:1143–1155.
42. Glaser L, Stevens J, Zamarin D, Wilson IA, García-Sastre A, Tumpey TM, Basler CF, Taubenberger JK, Palese P: A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *J Virol* 2005, **79**:11533–11536.
43. Stevens J, Blixt O, Paulson JC, Wilson IA: Glycan microarray technologies: tools to survey host specificity of influenza viruses. *Nat Rev Microbiol* 2006, **4**:857–864.
44. Volpi N: Therapeutic applications of glycosaminoglycans. *Curr Med Chem* 2006, **13**:1799–1810.
45. Sears P, Wong CH: Toward automated synthesis of oligosaccharides and glycoproteins. *Science* 2001, **291**:2344–2350.
46. Seeberger PH, Werz DB: Automated synthesis of oligosaccharides as a basis for drug discovery. *Nat Rev Drug Discov* 2005, **4**:751–763.
47. Wang D: Carbohydrate microarrays. *Proteomics* 2003, **3**:2167–2175.
48. de Paz JL, Horlacher T, Seeberger PH: Oligosaccharide microarrays to map interactions of carbohydrates in biological systems. *Methods Enzymol* 2006, **415**:269–292.
49. Feizi T, Chai W: Oligosaccharide microarrays to decipher the glyco code. *Nat Rev Mol Cell Biol* 2004, **5**:582–588.
50. Wilson JC, von Itzstein M: Recent strategies in the search for new anti-influenza therapies. *Curr Drug Targets* 2003, **4**:389–408.
51. Stiver G: The treatment of influenza with antiviral drugs. *CMAJ* 2003, **168**:49–56.
52. Torano G, Toledo ME, Baly A, Fernandez-Santana V, Rodriguez F, Alvarez Y, Serrano T, Musachio A, Hernandez I, Hardy E, Rodriguez A, Hernandez H, Aguila rA, Sanchez R, Diaz M, Muzio V, Dfana J, Rodriguez MC, Heynngnezz L, Verez-Bencomo V: Phase I clinical evaluation of a synthetic oligosaccharide-protein conjugate vaccine against *Haemophilus influenzae* type b in human adult volunteers. *Clin Vaccine Immunol* 2006, **13**:1052–1056.
53. Fernandez Santana V, Pena Icart L, Beurret M, Costa L, Verez Bencomo V: Glycoconjugate vaccines against *Haemophilus influenzae* type b. *Methods Enzymol* 2006, **415**:153–163.
54. Werz DB, Seeberger PH: Carbohydrates as the next frontier in pharmaceutical research. *Chem Eur J* 2005, **11**:3194–3206.
55. Kwon Y-U, Soucy RL, Snyder DA, Seeberger PH: Assembly of a series of malarial glycosylphosphatidylinositol anchor oligosaccharides. *Chem Eur J* 2005, **11**:2493–2504.
56. Schofield L, Hewitt MC, Evans K, Siomos M-A, Seeberger PH: Synthetic GPI as a candidate anti-toxic vaccine in a model of malaria. *Nature* 2002, **418**:785–789.
57. Vansteenkiste J, Rossi G, Foote M: Darbepoetin alfa: a new approach to the treatment of chemotherapy-induced anaemia. *Expert Opin Biol Ther* 2003, **3**:501–508.
58. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem Sci* 1989, **14**:475–477.
59. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
60. van Kuik JA, Vliegenthart JF: Databases of complex carbohydrates. *Trends Biotechnol* 1992, **10**:182–185.
61. van Kuik JA, Hård K, Vliegenthart JFG: A 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates. *Carbohydr Res* 1992, **235**:53–68.
62. Loss A, Bunsmann P, Bohne A, Loss A, Schwarzer E, Lang E, von der Lieth CW: SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 2002, **30**:405–408.

63. Lütteke T, Bohne-Lang A, Loss A, Götz T, Frank M, von der Lieth CW: GLYCOCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* 2006, **16**:71R–81R.
64. Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH: GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* 2003, **31**:511–513.
65. Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, Sasisekharan R: Advancing glycomics: implementation strategies at the Consortium for Functional Glycomics. *Glycobiology* 2006, **16**:82R–90R.
66. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, Kawasaki T, Kanehisa M: KEGG as a glycome informatics resource. *Glycobiology* 2006, **16**:63R–70R.
67. Packer NH, von der Lieth CW, Aoki-Kinoshita KF, Lebrilla CB, Paulson JC, Raman R, Rudd P, Sasisekharan R, Taniguchi N, York WS: Frontiers in glycomics: Bioinformatics and biomarkers in disease. An NIH White Paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006). *Proteomics* 2008, **8**:8–20.