Preface

Why Is It Timely to Publish a Book on (Bio)informatics for Glycobiology and Glycomics?

The four essential molecular building blocks of cells are nucleic acids, proteins, lipids, and carbohydrates, often also referred to as glycans. Nucleotide and protein sequences are the heart of nearly all bioinformatics applications and research, whereas glycan and lipid structures have been widely neglected. Glycans are the most abundant and diverse of Nature's biopolymers. Complex carbohydrates are often covalently attached to proteins and lipids and thus constitute a significant amount of the mass and structural variation in biological systems. The field of glycobiology is focused upon understanding the structure, chemistry, biosynthesis, and biological function of glycans and their derivatives.

It has long been known that carbohydrates encode biological information. For example, it was shown already in 1952 that variation of blood group determinants is a consequence of glycosylation, that is, the addition of complex carbohydrates to proteins and lipids. Thus, the chemistry, biochemistry, and biology of carbohydrates were prominent areas of research during the beginning and the middle of the last century. Nevertheless, determining the biological consequences of glycosylation has been extremely difficult. In an editorial article in the March 2001 edition of *Science* devoted to glycobiology, it was described as a "Cinderella field" of research. This means that it is "an area (of research) that involves much work but, alas, does not get to show off at the ball with her cousins, the genomes and proteins."

With the awareness that the human genome encodes for a significantly smaller number of genes than estimated from genomes of lower organisms such as yeast [1], it became obvious that each gene can be used in different ways depending on how it is regulated. Consequently, the study of post-translational protein modifications, which can alter the functions of proteins, has entered an era of renaissance and come increasingly into the scientific focus. Glycobiology research has attracted increasing attention because glycosylation is the most complex and most frequently occurring post-translational modification. Similar to the developments in genomics and proteomics, high-throughput glycomics projects to decipher the role of carbohydrates in health and disease are emerging [2–9]. With the increasing amount of experimental data, the need to develop appropriate glycan related databases and bioinformatics tools is obvious. However, until recently, informatics have been only poorly represented in glycobiology [10–12].

Unfortunately, most of the tools and applications developed in bioinformatics for the description and analysis of DNA (RNA) and protein sequences cannot be directly applied to carbohydrates. This is mainly due to the fact that oligosaccharides can exhibit various ways to link together their building blocks, the monosaccharides. In nature, monosaccharides are

found which are connected to up to four others. This has the consequence that branched structures can be formed. Such structures can no longer be described as linear sequences, but rather need to be described as topologies of specifically connected building blocks. In this respect, the encoding of complex carbohydrates based on building blocks is more similar to the digital description of organic molecules, which are described in chemoinformatics through the topology of atoms.

As with proteomics, mass spectrometry (MS) has become a key technology for the identification of glycans. The experimental procedures in MS-based glycomics analysis are similar to those applied in proteomics. However, until recently, no efficient software tools were available to interpret the spectra. Therefore, the annotation of spectra and assignment of glycan structures to the mass peaks had to be done manually by an expert.

The lack of efficient automatic assignment procedures is still the major bottleneck for an automatic high-throughput analysis of glycans in glycomics projects. Consequently, several computational attempts to overcome this unfavorable situation have been published in recent years. The developed algorithms were mainly implemented by experimentalists focused on solving the specific needs of their experimental setup and scientific questions. However, it is obvious that glycomics calls for more general solutions, highly sophisticated algorithmic approaches, and standardization.

Since the beginning of this century, a small but rather active community of researchers emerged with the aim of working out the foundations of the informatics for glycobiology and glycomics. The development and use of informatics tools and databases for glycobiology and glycomics research have thus increased considerably in recent years. However, this field must still be considered as being in its infancy compared with genomics and proteomics.

It is the aim of this book to give an introduction to this emerging field of science for the experimentalist working in glycobiology and glycomics, and also for the computer scientists looking for new challenges in the development of highly sophisticated algorithmic approaches.

Glycomics: an Exotic and Somewhat Forgotten Area of Bioinformatics

The European Bioinformatics Institute (EBI) (www.ebi.ac.uk/) [13, 14] is one of the largest centers world-wide for research and services in bioinformatics, managing a broad range of freely available databases of biological sequences, information, and knowledge. However, in 2007, the EBI did not provide access to any collection containing glycan structures. The US National Center for Biotechnology Information (NCBI) [15] (www.ncbi.nlm.nih.gov/) provides the PubChem service (pubchem.ncbi.nlm.nih.gov/), a fairly new service containing information on the biological activities of small molecules, which also includes carbohydrate structures. However, the main focus of PubChem is to provide access to glycans that can be used as chemical probes or ligands to study the functions of genes, cells, and biochemical pathways. The Japanese Kyoto Encyclopedia of Genes and Genomes (KEGG) (www.genome.ad.jp/) [16] is a suite of databases and associated software that aims to integrate the current knowledge on molecular interaction networks in biological processes with the information about the universe of genes and proteins, in addition to information about the universe of chemical compounds and reactions. KEGG contains a GLYCAN database with about 11 000 structures [17]. Most of the data were taken from CarbBank [18, 19], the largest attempt to build up a comprehensive collection of all known carbohydrate structures

during the 1980s and 1990s (see also the paragraph on the history of glyco-related databases in Chapter 1). KEGG is currently the most developed project which links glycan structures with available proteomics and glycomics data though biosynthetic pathways.

The Bioinformatics Links Directory (www.bioinformatics.ca/links_directory/) [20, 21], an online community resource that contains a directory of freely available tools, databases, and resources for bioinformatics and molecular biology research – which is compiled by collecting applications which have been published in the annual *Nucleic Acids Research Web* issues – lists only six tools dealing with the bioinformatics of carbohydrate structures. This is in contrast to the 376 useful resources for DNA sequence analyses reported, and the 651 links to useful resources for protein sequence and structure analyses, which include also tools for phylogenetic analyses, prediction of protein structures and functions, and analyses of protein–protein interactions. On the other hand, collections of links compiled with special emphasis on glyco-related web applications (see, e.g., www.eurocarbdb.org) already show more than 60 dedicated websites. This discrepancy reflects the current points of view from both sides: while the bioinformatics community widely ignores the existence of macromolecules other than DNAs, RNAs, and proteins, scientists developing software applications for glycobiology do not regard themselves as part of the bioinformatics community.

The Role of (Bio)informatics in Glycobiology Research

The involvement of (bio)informatics in glycobiology research can be divided into:

1. The application of classical bioinformatics tools to analyze the DNA and protein sequences, which have a relation to carbohydrates. Such sequences may be the proteins to which glycans are attached, the enzymes which build or modify oligosaccharides, or the lectins which recognize a certain sugar epitope.
2. Applications and databases where an explicit description of the glycan structure is required. All analytical tools to determine glycan structures and structure–function relations depend heavily on appropriate encoding of glycan structures.
3. Atomic descriptions of carbohydrate-protein recognition processes in which the spatial structures of complex carbohydrates, their conformational preferences, and their energetics are analyzed.

In this book, we try to provide a comprehensive overview of all three areas of active research. The chapters are written by active researchers who have made essential contributions to the development of the field of glycobioinformatics in recent years.

Use of Classical Bioinformatics Databases and Tools

Chapter 1, *Glycobiology, Glycomics, and (Bio)Informatics*, briefly describes the biological role of carbohydrates, their biosynthesis, and the enzymes which are responsible for the stepwise synthesis of the branched oligosaccharides – the glycosyltransferases. Special emphasis is placed on the role of bioinformatics in accelerating the identification of human carbohydrate active enzymes with the help of bioinformatics databases and appropriate alignment tools.

Chapter 5, *Carbohydrate-active Enzymes Database: Principles and Classification of Glycosyltransferases*, gives a comprehensive overview of the enzymes responsible for the biosynthesis of the glycosidic bonds in living organisms. The authors develop and maintain the CAZy database, the world's most complete resource describing the families of structurally related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds. The subsequent chapter gives a short overview of the focus of other existing data resources which provide access to glyco-related enzymes.

Chapter 7, *Bioinformatics Analysis of Glycan Structures from a Genomic Perspective*, describes how informatics can help to elucidate the biological function of glycans, which are intertwined with the rest of the biological system such as interacting proteins and chemical compounds. In this chapter, an approach is presented based on the data of the Kyoto Encyclopedia of Genes and Genomes (KEGG) including a comprehensive glycan data resource called KEGG GLYCAN. It encompasses all aspects of the biological system, incorporating genomic information integrated with pathways and reactions and also chemical compounds.

Chapter 8, *Glycosylation of Proteins*, gives a short introduction to this phenomenon, and Section 8.2 therein, *GlySeq: an Analysis of Experimentally Determined Occupied Glycosylation Sites*, describes services which provide access to the respective glyco-related experimental data contained in the Protein Data Bank (PDB) [22] and SWISS-Prot. The *GlySeq* service – although focusing on analyzing the data from the viewpoint of carbohydrates – provides access to both moieties: the carbohydrates and the proteins. Glycosylation is known to affect protein folding, localization, and trafficking, protein solubility, antigenicity, biological activity and half-life. Chapter 9, *Prediction of Glycosylation Sites in Proteins*, summarizes the ideas of pattern recognition for protein glycosylation site prediction from peptide sequence alone. It provides a general introduction to data-driven prediction methods to solving this problem, including a discussion on artificial neural networks.

Informatics Applications Where a Special Encoding of Glycan Structures is Required

Due to the structural complexity of complex carbohydrates which can form highly branched structures, most of the tools and applications developed in classical bioinformatics for the description and analysis of DNA (RNA) and protein sequences cannot be directly applied to carbohydrates.

Chapter 2, *Introduction to Carbohydrate Structure and Diversity*, provides a short description of the major types of carbohydrate structural motifs found in nature. The structural diversity of carbohydrates that is currently available in databases has been analyzed and is also presented. As excellent reference books are available summarizing the cellular location and biological function of the various types of glycan, the Editors decided not to repeat this information here and recommend that readers consult the existing compendia for further reading.

Special encoding schemes are required which are able to describe all structural features of complex carbohydrates found in nature. Unfortunately, no standard description existed until recently that was capable of coping with all structural features of carbohydrates as needed, for example, for the emerging glycomics projects. Chapter 3, *Digital Representations of*

Oligo- and Polysaccharides, gives a comprehensive overview of all structural features which have to be encoded to cope with all carbohydrate-specific structural elements. It discusses various often used digital formats, which have so far been suggested and are used for specific applications or to encode carbohydrate structures in databases. The chapter gives an outline of future directions of developments.

The introductory part of the book is rounded off by a chapter on evolutionary considerations in studying the structural diversity of the most important class of terminal monosaccharides – the sialic acids. The contribution suggests that the structural diversity of sialic acids reflects the often conflicting pressures of evading pathogens, while simultaneously maintaining endogenous functions. Since most pathogens replicate much faster than their hosts, they can rapidly evolve different ways to target or mimic structures that are critical for host processes, which may be especially relevant to glycans.

An in-depth understanding of the biological functions of complex carbohydrates requires a detailed knowledge of all structural features of their primary sequence and also the conformational space that they can access. Analysis of carbohydrates has proved to be difficult in the past. Fortunately, modern analytical methods have the ability to elucidate most structural details at the concentration levels required for glycomics projects. However, at present, informatics tools give only limited support to enable an automatic, reliable interpretation of the vast amount of data recorded by the analytical methods. This deficiency currently represents a severe bottleneck for the practical implementation of high-throughput glycomics projects. Therefore, it is not surprising that the development of algorithms and tools to interpret analytical data constitutes the most active field of software design in the glycomics field.

The chapters included in Section IV, *Experimental Methods – Bioinformatics Requirements*, summarize the status of analytical procedures used in glycomics and the algorithms, software tools, and services which are available to support the interpretation of data. Similarly to proteomics, MS, HPLC and NMR are the main experimental techniques in glycomics research. However, the concrete experimental procedures applied by various researchers vary considerably, so that it is necessary to outline the analytical procedures since otherwise the informatics requirements would be difficult to explain.

Spatial Structures of Carbohydrates and Atomic Descriptions of Carbohydrate–Protein Recognition Processes

The elucidation of conformational preferences of complex carbohydrate structures has a long history, which dates back to the early 1980s. Chapter 18, *Conformational Analysis of Carbohydrates – A Historical Overview*, reviews these developments, and is followed by Chapter 19, *Predicting Carbohydrate 3D Structures Using Theoretical Methods*, where the various commonly used theoretical approaches and simulation methods are introduced and their applications to predict 3D structures of carbohydrates are discussed. Section 19.4, *Generation of 3D Structures of Glycoproteins*, describes a service available through the GLYCOSCIENCES.de portal [23], which generates 3D structures of glycoproteins. Chapter 20, *Synergy of Computational and Experimental Methods in Carbohydrate 3D Structure Determination and Validation*, describes methods to find and analyze experimental 3D structures of carbohydrates in the Protein Data Bank.

Lectins are carbohydrate-binding proteins or glycoproteins which often recognize a specific sugar epitope and are thus important for a broad variety of specific recognition processes and signaling events. Crystal structures of members of the different animal and plant lectin families have revealed a wide variety of lectin folds and carbohydrate binding site architectures. Despite this large variability, a number of interesting cases of both convergent and divergent evolution among plant, animal, and bacterial lectins are noted. These similarities are reviewed in Chapter 21, *Structural Features of Lectins and Their Binding Sites*.

Chapter 22, *Statistical Analysis of Protein–Carbohydrate Complexes Contained in the PDB*, provides a detailed overview of the interactions, which specific carbohydrates or classes of glycans exhibit with proteins in the available experimentally determined protein–carbohydrate complexes.

Current Status of Informatics for Glycosciences

Recent years have seen a variety of new databases and software tools emerging in the field of glycomics. However, the current situation in glycobioinformatics 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. For example, no comprehensive carbohydrate data collections similar to those currently available for genomic and proteomic data have been compiled so far. There is currently no location where information about all carbohydrates reported in peer-reviewed scientific papers is systematically stored. Procedures (similar to those for protein sequences) have not yet been established for scientists to report the observation of specific glycan structures in specific environments and to store these observations in a generally accepted database.

These are the reasons why we have chosen not to describe in detail the currently available glyco-related databases. It can be anticipated that the development of databases will be subject to rather rapid changes within the next few years, so that the descriptions provided could quickly become obsolete. As compensation, we provide a comprehensive list of web links to glyco-related databases, services, consortia, and communities. As many of the listed URLs are maintained by larger consortia and longer lasting bioinformatics projects, it is the hope that these will provide a more sustainable solution for this book.

Acknowledgments

The authors thank Dr Robin Thomson, Institute for Glycomics, Griffith University, for carefully reading the manuscript and many useful suggestions to improve its readability.

Heidelberg, June 2007

Claus-Wilhelm von der Lieth
Martin Frank
Thomas Lütteke

References

1. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004, **431**:931–945.

2. 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.
3. Mechref Y, Novotny MV: Miniaturized separation techniques in glycomic investigations. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* 2006, **841**:65–78.
4. Alvarez RA, Blixt O: Identification of ligand specificities for glycan-binding proteins using glycan arrays. *Methods Enzymol.* 2006, **415**:292–310.
5. Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J, *et al.*: Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc. Natl. Acad. Sci. USA* 2004, **101**:17033–17038.
6. Feizi T, Chai W: Oligosaccharide microarrays to decipher the glyco code. *Nat. Rev. Mol. Cell Biol.* 2004, **5**:582–588.
7. 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.
8. Prescher JA, Bertozzi CR: Chemical technologies for probing glycans. *Cell* 2006, **126**:1–4.
9. Pratt MR, Bertozzi CR: Synthetic glycopeptides and glycoproteins as tools for biology. *Chem. Soc. Rev.* 2005, **34**:58–68.
10. von der Lieth CW, Bohne-Lang A, Lohmann K, Frank M: Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform.* 2004, **5**:164–178.
11. von der Lieth CW, Lütteke T, Frank M: The role of informatics in glycobiology research with special emphasis on automatic interpretation of MS spectra. *Biochim. Biophys. Acta* 2006, **1760**:568–577.
12. von der Lieth CW: An endorsement to create open databases for analytical data of complex carbohydrates. *J. Carbohydr. Chem.* 2004, **23**:277–297.
13. Brooksbank C, Camon E, Harris MA, Magrane M, Martin MJ, Mulder N, O'Donovan C, Parkinson H, Tuli MA, Apweiler R, *et al.*: The European Bioinformatics Institute's data resources.. *Nucleic Acids Res.* 2003, **31**:43–50.
14. Brooksbank C, Cameron G, Thornton J: The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* 2005, **33**:D46–D53.
15. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, *et al.*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2007, **35**:D3–D4.
16. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004, **32**:D277–280.
17. 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.
18. Doubet S, Bock K, Smith D, Darvill A, Albersheim P: The Complex Carbohydrate Structure Database. *Trends Biochem. Sci.* 1989, **14**:475–477.
19. Doubet S, Albersheim P: CarbBank. *Glycobiology* 1992, **2**:505.
20. Fox JA, Butland SL, McMillan S, Campbell G, Ouellette BF: The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Res.* 2005, **33**:W3–W24.
21. Fox JA, McMillan S, Ouellette BF: A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res.* 2006, **34**:W3–W5.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res.* 2000, **28**:235–242.
23. 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.