Bioinformatics for glycomics: Status, methods, requirements and perspectives

Claus-Wilhelm von der Lieth, Andreas Bohne-Lang, Klaus Karl Lohmann and Martin Frank

Date received (in revised form): 10th March 2004

Claus-Wilhelm von der Lieth
is the leader of the molecular modelling group of the Central Spectroscopic Department of the German Cancer Research Center, Heidelberg, Germany. His main research interests are in molecular modelling and computational structural biology, with the main focus on glycobiology.

Andreas Bohne-Lang
is a computer scientist who has developed a variety of well-known algorithms and web applications to encode, generate and represent structures of complex carbohydrates.

Klaus Karl Lohmann
is a pharmacist who has developed various algorithms and web applications assisting the automatic interpretation of mass spectra of complex carbohydrates.

Martin Frank
is a chemist. His main research interests are in molecular modelling and the application of advanced simulation techniques for structural biology. He has developed various computational approaches to efficiently explore the conformational space of glycans.

Keywords: *glycobiology, glycomics, glycans, structural description, databases, automatic MS spectrum interpretation*

Claus-Wilhelm von der Lieth,
German Cancer Research Center,
Central Spectroscopic Department
B090,
Im Neuenheimer Feld 280,
D-69120 Heidelberg, Germany

E-mail: w.vonderlieth@dkfz.de

Abstract

The term ‘glycomics’ describes the scientific attempt to identify and study all the glycan molecules – the glycome – synthesised by an organism. The aim is to create a cell-by-cell catalogue of glycosyltransferase expression and detected glycan structures. The current status of databases and bioinformatics tools, which are still in their infancy, is reviewed. The structures of glycans as secondary gene products cannot be easily predicted from the DNA sequence. Glycan sequences cannot be described by a simple linear one-letter code as each pair of monosaccharides can be linked in several ways and branched structures can be formed. Few of the bioinformatics algorithms developed for genomics/proteomics can be directly adapted for glycomics. The development of algorithms, which allow a rapid, automatic interpretation of mass spectra to identify glycan structures is currently the most active field of research. The lack of generally accepted ways to normalise glycan structures and exchange glycan formats hampers an efficient cross-linking and the automatic exchange of distributed data. The upcoming glycomics should accept that unrestricted dissemination of scientific data accelerates scientific findings and initiates a number of new initiatives to explore the data.

INTRODUCTION

Glycan structures are encoded indirectly in the genome.[1–3] Compared with the biosynthesis of proteins, there is an additional step in the decoding process (see Figure 1). A variety of carbohydrate-active enzymes (glycosyltransferase, glycosidases, carbohydrate-modifying enzymes) that create, degrade or modify glycosidic bonds by which monosaccharide units are connected (see Figure 2) determine the structure of glycans, which are attached to a protein in a specific cellular environment. More than 5,000 genes have been assigned as putative carbohydrate-synthesising enzymes in the Carbohydrate-Active enZYmes (*CAZy*) database[4,5] but the function of less than 10 to 20 per cent of these genes is known to date. The enzyme-one-linkage rule suggests that it will eventually be possible to describe the full repertoire of glycan structures that can be made in a particular cell by determining which enzymes are expressed in the cell. However, different copies of a glycoprotein of the same regulatory pathway can be modified with different glycans (glycoforms) and the extent of heterogeneity varies from protein to protein.

GLYCOMICS

The term ‘glycomics’ describes the scientific attempt[6] to identify and study all the carbohydrate molecules – the glycome – produced by an organism such as human or mouse. Rapid and sensitive high-throughput analytical methods employing mass spectrometry (MS) and high-performance liquid chromatography (HPLC) techniques are currently applied to provide information on the glycan repertoire of cells, tissues and organs.[7,8] One of the aims of the emerging glycomics projects is to create a cell-by-

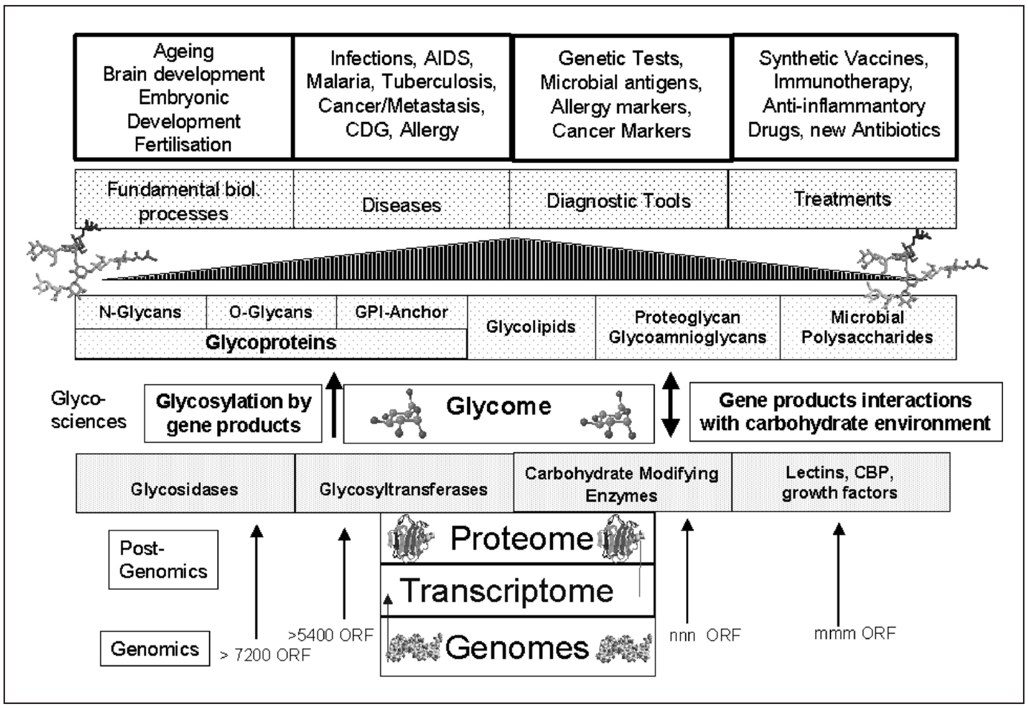


Figure 1: Biologically active glycan structures are encoded indirectly in the genome. A variety of glycosidases, glycosyltransferase and carbohydrate-modifying enzymes create, degrade or modify glycosidic bonds by which monosaccharide units are connected and attach the glycan structures to other macromolecules like proteins and lipids. Carbohydrates are involved in many regulatory processes and can therefore act as well as diagnostic markers or can be used as the target structure to inhibit a specific regulatory pathway. CBP, carbohydrate binding proteins; ORF, open reading frame; GPI, glycosylphosphatidylinositol; CDG, congenital disorder of glycosylation

Glycan profiles can distinguish between disease and normal cell state

cell catalogue of glycosyltransferase expression and detected glycan structures. Glycan profiling of normal and diseased forms of a glycoprotein has provided new insights for future research in rheumatoid arthritis, prion disease and congenital disorders of glycosylation.[9–13] In all of these diseases, differences in glycosylation indicate that there are cellular or genetic changes that affect the activity of specific glycosyltransferases.

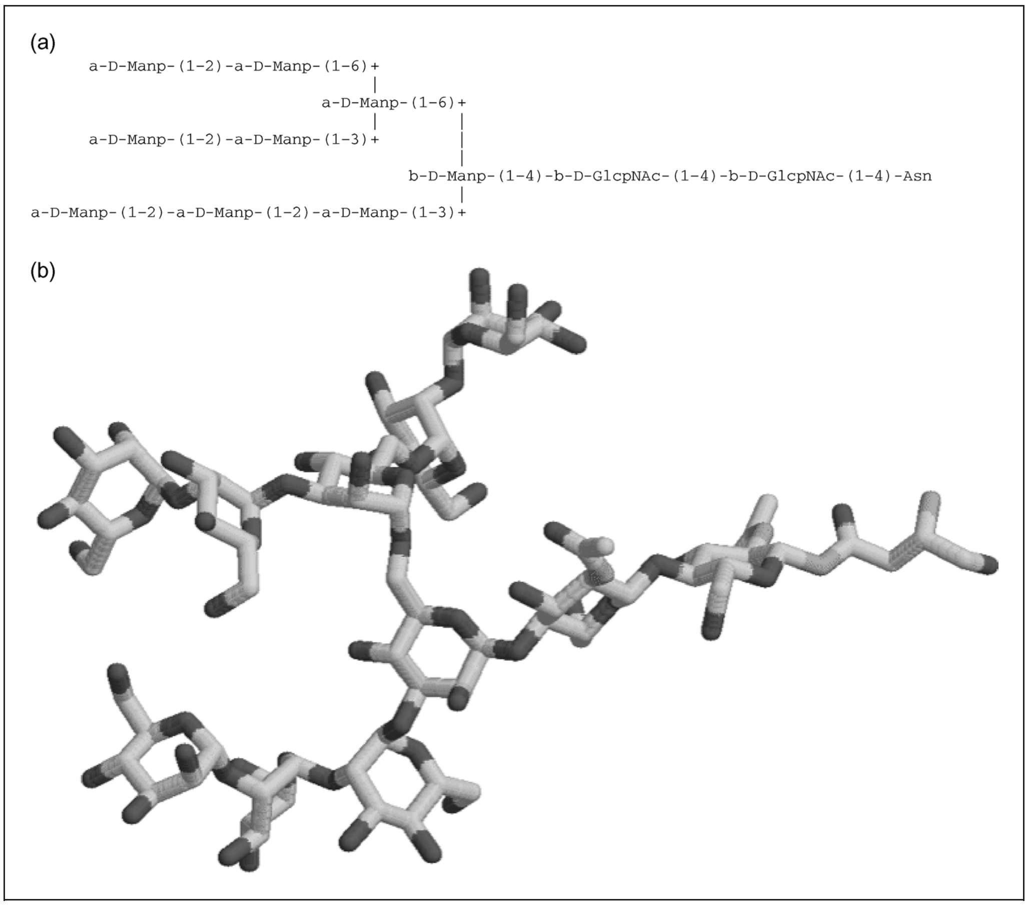
Knockout mice help us to understand the metabolic pathways of glycans

Genetically modified mice[14–16] provide a means of investigating the importance of specific glycan epitopes via ablation of certain glycosyltransferases. The detailed characterisation of any change in glycosylation occurring as a result of these experiments is essential to gain new insights about the role of glycosyltransferases in a specific cell, organ or tissue. Structural elucidation of the glycan composition of normal murine tissues and organs is therefore an essential prerequisite to investigate changes occurring in knockout mice.

Although carbohydrate microarrays[17–20] for the study of carbohydrate–protein interactions are largely at proof-of-concept stage, they have shown promise for establishing whether proteins bind carbohydrates and for assigning their ligands. After initial investment in the technical advances necessary to prepare suitable arrays, the advantages in speed and simplicity of identification of carbohydrate ligands are expected to be considerable. The use of carbohydrate microarrays promises to be a particularly valuable tool to understand how glycosylation of a cell changes[21] during differentiation and/or activation.

GLYCOPROTEOMICS

The terms 'glycomics' and 'glycoproteomics' are often used simultaneously within the literature to



(a)

```
      a-D-Manp-(1-2)-a-D-Manp-(1-6)+
                                   |
                              a-D-Manp-(1-6)+
                                   |        |
      a-D-Manp-(1-2)-a-D-Manp-(1-3)+        |
                                            |
                                       b-D-Manp-(1-4)-b-D-GlcpNAc-(1-4)-b-D-GlcpNAc-(1-4)-Asn
                                            |
a-D-Manp-(1-2)-a-D-Manp-(1-2)-a-D-Manp-(1-3)+
```

(b)

Figure 2: (a) Extended alphanumeric description of a glycan. A complete structural characterisation has to include the sugar sequence, the monosaccharide stereochemistry, the anomeric configuration and the linkage information. (b) 3D model of the same glycan generated with SWEET-II

Glycosylation affects many activities and functions of proteins

generally describe methods pertaining to the analysis of glycosylation. However, the term glycomic refers to the global category encompassing all glycoconjugates (eg glycolipids, glycoproteins, lipopolysaccharides, peptidoglycans, proteoglycans) while glycoproteomics refers only to the characterisation of the glycosylation of proteins.

PROTEIN GLYCOSYLATION

The human genome seems to encode for not more than 30,000 to 40,000 proteins.[22] This relatively small number of human genes compared with the genome of other species has been one of the big surprises coming out of the analysis of the human genome project. A major challenge is to understand how post-translational events – among these, glycosylation is by far the most abundant – affect the activities and functions of these proteins in health and disease. More than half the proteins in the human body have carbohydrate molecules attached.[23,24] Glycosylated proteins are ubiquitous components of extracellular matrices and cellular surfaces where their oligosaccharide moieties are implicated in a wide range of cell–cell and cell–matrix recognition events.[9,25] As an attempt to predict protein function solely from protein chain global properties (molecular weight, length etc.) and potential post-translational modifications, glycosylation turned out to be one of the most important determinants for functional classification.[26,27]

PREDICTION OF *N*-GLYCOSYLATION SITES

Potential *N*-glycosylation sites can be identified by the presence of the Asn-X-Ser/Thr (where X cannot be proline) sequon in peptide sequence databases. For reasons that are not understood, not all such sequons are glycosylated. Unfortunately, the unambiguous determination of occupied *N*-glycosylation sites is experimentally demanding and may vary between different cellular locations. Therefore fewer than 1,000 annotated, experimentally verified, glycosylated sequons are listed in Swiss-Prot,[24] a curated protein sequence database which strives to provide high level of annotation for structural and functional properties of proteins including post-translational modifications.

SWISS-PROT contains nearly 1000 annotated and verified glycosylated sequons

Another independent resource for the unambiguous evidence for the occupancy of a glycosylation site is the Protein Data Base (PDB), the single worldwide repository for 3D biological macromolecular structure data. The attached *N*- and *O*-glycans have been detected by X-ray crystallography. A recent study (September 2003) found 770 entries with 2851 attached N-glycan chains.[28] The *GlySeq* web-interface enables a detailed statistical analysis of unambiguously assigned glycosylation sites derived from both sources. Although some general trends such as increased occurrence of aromatic residues before the ASN at position +2 have been described, no additional general rules could be detected so far.[29] Probably the conformation of the protein backbone in the vicinity of potential *N*-glycosylation sites may play an important role.

SWISS-PROT and the Protein Data Base (PDB) contain information about possible glycosylation

NetNGlyc, based on a model of trained neural networks,[30] predicts *N*-glycosylation sites. Since *N*-glycosylation occurs in the secretory machinery, only the proteins that contain a secretion signal peptide can be *N*-glycosylated. To check that the submitted sequence is a secretory protein, *SignalP* is run in parallel. *SignalP*[31] predicts with a high degree of confidence whether a protein contains a signal peptide and predicts the position of the cleavage site. If no signal peptide is found, the user is warned that the protein is unlikely to be *N*-glycosylated.

***NetOGlyc* allows the prediction of *O*-glycosylation sites in mammalian proteins**

PREDICTION OF *O*-GLYCOSYLATION SITES

O-GLYCBASE[30] (Version 6.00 has 242 glycoprotein entries) is a database of glycoproteins with *O*-linked glycosylation sites. Entries with at least one experimentally verified *O*-glycosylation site have been compiled from protein sequence databases and literature. Each entry contains information about the glycan involved, the species, sequence, a literature reference and cross-references to other databases.

NetOGlyc[32] predicts *O*-glycosylation (mucin-type glycosylation) sites in mammalian protein sequences. The neural network behind the program was trained to recognise the sequence context and surface accessibility of 299 known and verified mammalian mucin-type *O*-glycosylation sites from the O-*GlycBase* database. The algorithm correctly predicted 83 per cent of the glycosylated serine and threonine residues in an independent test set and rejected 90 per cent of the non-glycosylated residues. Moreover, it can correctly predict 51 per cent of *O*-mannosylated sites in fungal proteins and as much as 85 per cent of non-glycosylated sites.

Mass spectrometry is a common method to determine glycosylation sites in proteins

ANALYTICAL TECHNIQUES

In recent years, MS[8,33–36] has become the method of choice for highly sensitive protein identification and characterisation. Fundamentals that enable a rapid identification of peptides are the availability of large protein sequence databases and the development of efficient algorithms for tandem mass spectrometry (MS/MS) fragment identification techniques. There are numerous reports describing the characterisation of glycans released from proteins and detected by MS. Additionally, some more recent papers outline the oligosaccharide profiling directly from 1- and 2-dimensional gel-separated proteins.

NMR (nuclear magnetic resonance) techniques can lead to a full structural characterisation of oligosaccharides including the monosaccharide stereochemistry, the anomeric configuration, the linkage type and the complete sugar sequence.[37] However, NMR spectroscopy is relatively insensitive with respect to the amount of samples needed in order to obtain good-quality structural data.[38] A complete structure determination by ^{1}H-NMR requires a glycan's availability in virtually pure state and amounts of material at the microgram level. The additional steps to scale up the amount of oligosaccharides produced and their purification normally exclude NMR techniques that are applied in high-throughput sequencing projects. Since there have been two recent reviews on NMR application for structural elucidation,[39,40] this topic will not be discussed here in detail.

Automated carbohydrate synthesis can be used to produce sufficient amounts of glycans to allow NMR-experiments

On the other hand automated carbohydrate synthesis technologies are now available that can rapidly produce sufficient amounts of pure oligosaccharides as required for drug discovery projects.[41–43] Such investigations benefit pivotally from the detailed structural information provided by NMR spectroscopy.

Until recently there were few bioinformatics databases and web applications dealing with glycobiology questions. The latest annually published list of molecular biology databases[44] showed only 3 among about 300 databases dealing with glycobiology-related aspects. However, mainly driven by the need to create algorithms for automatic interpretation of MS spectra produced by the upcoming glycomics project, more than ten projects where presented during an informatics workshop at the annual meeting of the US Society for Glycobiology in December 2003.

Only a few databases are dealing with glycobiology-related content

HOW TO DEAL WITH THE STRUCTURAL COMPLEXITY OF GLYCAN STRUCTURES?

Sequences for complex carbohydrates (see Figure 2) differ significantly from the simple linear one-letter code that describes genes and proteins: the number of naturally occurring residues is much larger for glycans, each pair of monosaccharide residues can be linked in several ways, and one residue can be connected to three or four others (branching). The information content that can be encoded by glycans in a given sequence is high.[45] 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. But glycans have the potential to assemble into more than 15 million four-unit arrangements. As a consequence, glycans bearing biological information consist of a few up to about 20 residues. Unfortunately, unlike the one-letter code of amino acid sequence for proteins, there is no representation of carbohydrates that is similarly accepted by scientists from all disciplines. In the following section, frequently used structural descriptions will be discussed. It is demonstrated that for bioinformatics applications only one internal representation is necessary, from which all other illustrations can be generated.

FREQUENTLY USED STRUCTURAL REPRESENTATIONS

The most general characterisation of a carbohydrate is its total mass or, less frequently used, the gross molecular formula. MS, the experimental method most often used to identify glycans, cannot distinguish between isomeric monosaccharides, which have the same mass. Therefore, glycans are often described by their composition. Here, the number of hexoses such as glucose, galactose or mannose as well as the number of all modified residues are reported. The description $\mathrm{Hex_5HexNAc_2dHex_1}$ indicates a glycan consisting of five hexoses, two hexose substituted with a *N*-acetyl group and one hexose where one hydroxyl group is substituted by a hydrogen.

For specific scientific questions only certain classes of glycans showing a limited number of residues and linkages

The web-tool *LiGraph* allows the translation of an alphanumeric glycan representation into graphical pictograms

have to be examined. Since the exact linkage information is often not available and can be estimated only through knowledge of the secretory pathway or additional experiments, pictograms indicating monosaccharides by circles, squares, stars and rhombuses connected by various types of lines are the preferred representation in most biomedically oriented publications dealing with glycans attached to proteins (see Figure 3). Unfortunately, at present, no generally agreed scheme for representing monosaccharides by symbols has been established. The US Consortium for Functional Glycomics (CFG), a large research initiative to understand the role of carbohydrate–protein interactions at the cell surface in cell–cell communication,[46] has recently announced a scheme that mainly aims to enable a convenient annotation of mass spectra for *N*- and *O*-glycans. However, it can be anticipated that the CFG scheme will be widely used. The *LiGraph* web tool enables the extended alphanumeric nomenclature (see below) to be translated into several graphical pictograms as often used in glycomics projects.

FULL STRUCTURAL CHARACTERISATION OF GLYCANS

A full structural characterisation has to include the complete sugar sequence, the monosaccharide stereochemistry, the anomeric configuration and the linkage information. The IUPAC-IUBMB (International Union of Pure and Applied Chemistry – International Union of Biochemistry and Molecular Biology) 'Nomenclature of Carbohydrates'[47] specifies how to uniquely describe complex oligosaccharides based on a three-letter code to characterise monosaccharide units (gal = galactose,

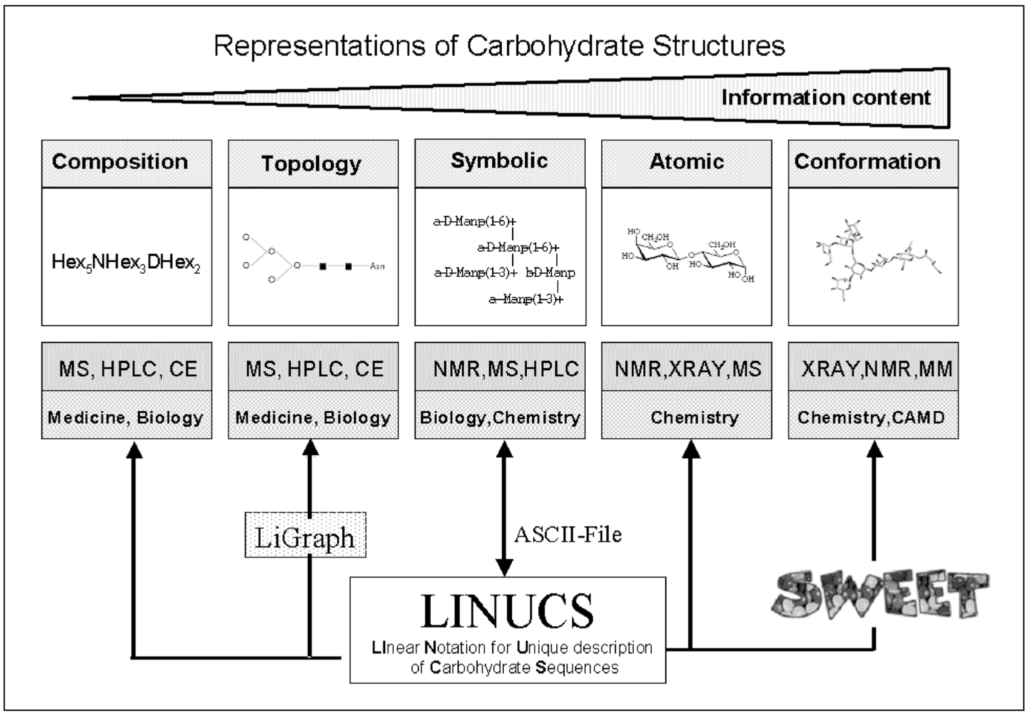


Figure 3: Various structural representations of glycan are used by different scientific communities and for the illustration of experimental results. However, only one computer representation of glycan structures is required, which has to include a complete structural description. Based on this encoding, the various representations can be generated. CAMD, computer-aided molecular design; CE, capillary electrophoresis; HPLC, high-performance liquid chromatography; MM, molecular modelling; MS, mass spectrometry; NMR, nuclear magnetic resonance

man = mannose, etc.). Each symbol for a monosaccharide unit is preceded by the anomeric descriptor and the configuration symbol. The ring size is indicated by an italic *f* for furanose or *p* for pyranose. The number of the ring C-atoms that link two monosaccharide units are given in parentheses between the symbols, an arrow indicates a linkage between two anomeric positions. The branches are displayed in separate lines (see Figure 2) and are connected by the linkage information. In such a way, long carbohydrate sequences can thus be adequately described in abbreviated form. This extended alphanumeric nomenclature is well suited for computer processing.

PSEUDO-3D AND 3D MODELS

LINUCS allow a linear unique representation of a carbohydrate

Chemists prefer the commonly used schematic all-atom and bond drawings of molecules where the stereochemistry is indicated graphically with wedges at each individual stereo centre. These pseudo-3D pictures are just another representation of the structural information contained in extended alphanumeric IUPAC-IUBMB description. However, no software is currently available that is capable of converting one representation into the other.

In drug discovery projects, virtual screening approaches[48] are increasingly applied to find molecules, which promise to exhibit a high affinity for a specific target protein. Spatial structures of the tested ligands as well as the protein are required to perform these studies. The freely available *SWEET-II*[49,50] server provides, for most types of carbohydrates, an efficient conversion of the extended alphanumeric IUPAC nomenclature into a reliable spatial structure.

SWEET-II allows the generation of 3D-structures

INTERNAL REPRESENTATION OF GLYCAN STRUCTURES

For the internal computer encoding and to enable an efficient link to glyco-related data from various data collections, it is obvious that a canonical description is required that includes a full structural description of a carbohydrate chain. However, the IUPAC rules are not sufficiently comprehensive to cover all ambiguities especially regarding the ordering of branches. Therefore, two linear codes have been described. The 'LINUCS' notation,[51] LInear Notation for Unique description of Carbohydrate Sequences, uses the extended alphanumeric IUPAC description and takes the glycosidic linking information to build up a hierarchy of the various branches starting from the reducing end of the oligosaccharide chain. The company Glycominds has built their database of glycan structures on the Linear Code$^{\mathrm{TM}}$,[52] a simple one to two-letter representation of saccharide units and linkages. The ordering of glycan branches is established using a special look-up table where the hierarchy of monosaccharide structures is defined.

The linear codes contain all required information to generate a variety of structural representations used by different communities (see Figure 3). The generation of masses and oligosaccharide compositions is a straightforward procedure. The LiGraph web tool translates the extended alphanumeric nomenclature into several graphical pictograms as often used in glycomics projects. The *SWEET-II*[49,50] service is able to generate reasonable 3D structures, which can be viewed by many molecular display programs and can be used for subsequent drug design studies. Currently, a tool to convert the alphanumeric description into a pseudo-3D model is missing.

DATABASES OF GLYCAN STRUCTURES

All existing digital data collections of glycan structures[52–57] (see Table 1, Berteau and Stenutz[39] and Marchal *et al.*[58] for recent reviews of the topic) use an encoding of the chemical structure as the primary key to access related bibliographic, biological, chemical or physical data. Unfortunately, currently there is no generally accepted way to normalise glycan structure and to exchange related data. This unfavourable situation currently hampers efficient cross-linking and the automatic exchange of data.

The CFG has started to develop a carbohydrate database that will make available various data sets pertaining to glycan structures. It can be expected that the specification for glycan structures and exchange formats made by CFG will have a large impact on future developments of standards in data formats for glycomics projects.

CARBOHYDRATE-ACTIVE ENZYMES

CAZy, a web-based database, contains information about carbohydrate active enzymes

It is estimated that about 1 per cent of the open reading frames (ORFs) of each genome is dedicated to the task of glycosidic bond synthesis. Furthermore, protein glycosylation, a glycosyltransferase-catalysed process, massively expands the functional proteome of higher organisms. To utilise the genomic resource to the full, it is essential to understand the sequences of the enzymes themselves, and how these sequences relate to enzyme structure, mechanism and specificity.

The *CAZy* server describes[4,5] families of structurally related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify or create glycosidic bonds. Within the server, the enzymes are grouped either by class (glycosidases and transglycosidases, glycosyltransferases, polysaccharide lyases, carbohydrate esterases, carbohydrate-binding modules) or by organism (access by completely sequenced organism). *CAZy* aims to list all known enzymes of sugar metabolism, and to provide information on structure but also on the mechanisms of activity. Over 120 organisms are represented, among which a vast majority are prokaryotes.

METABOLISM OF GLYCANS

KEGG contains information related to carbohydrate metabolic pathways

KEGG (Kyoto Encyclopedia of Genes and Genomes)[59,60] is a bioinformatics resource which is intended for linking genomes to cellular pathways, containing a complete set of building blocks (genes and molecules) and wiring diagrams (interaction networks) for cellular functions. Among others, *KEGG* includes manually drawn pathway diagrams for metabolism of complex carbohydrates and metabolism of complex lipids. These pathway diagrams are linked to individual entries of carbohydrate structures in the *GLYCAN* database, which has recently been added. It contains currently a total of 10,445 entries, among which only a few hundred were manually entered and linked to *KEGG* pathways. The rest represent unique structures derived from Complex Carbohydrate Structure Database, also known as *CarbBank*[53,61] which was maintained until 1999 by the Complex Carbohydrate Research Center, Athens (USA). The reactions catalysed by glycosyltransferases and other sugar-related enzymes are stored in the *REACTION* database where they are represented in a simple alphanumeric form of carbohydrate structures.

GLYCAN PROFILING

One strategy in functional glycomics is to compare the glycan repertoire found in normal and diseased or treated tissue. The rapid determination of the glycan composition for each tissue is an essential prerequisite to find new insights about the role of an enzyme in a specific cell, organ or tissue. To analyse changes in glycosylation occurring as a result of mouse strains with deleted or altered genes coding for glycosyltransferases, several experimental techniques are in use:

- Digestion with exoglycosidases, labelling of the glycan with a fluorescent tag and detection with HPLC techniques.
- Complete methylation of the glycan

Table 1: Glyco-related web tools and data collections

Name	Description/organisation	URL
Related carbohydrate information in protein databases		
CAZy	Carbohydrate-active enzymes	afmb.cnrs-mrs.fr/CAZY/
KEGG Pathway	Carbohydrate metabolism	www.genome.ad.jp/kegg/pathway.html
KEGG	Carbohydrate-active enzymes	www.genome.ad.jp/kegg/ligand.html
Lectines	3D structure of lectins	www.cermav.cnrs.fr/lectines/
CTDL	Animal lectins	ctld.glycob.ox.ac.uk/
PDB2LINUCS	Glycoproteins in PDB	www.dkfz.de/spec/pdb2linucs/
GlySeq	Analysis of glycoprotein sequences	www.dkfz.de/spec/glyseq/
Prediction of glycosylation positions in proteins		
NetNGlyc	*N*-glycosylation	www.cbs.dtu.dk/services/NetNGlyc/
NetOGlyc	*O*-glycosylation	www.cbs.dtu.dk/services/NetOGlyc/
YinOYang	Glyco-phosphorylation	www.cbs.dtu.dk/services/YinOYang/
big-PIPredictor	GPI-anchor prediction	mendel.imp.univie.ac.at/sat/gpi/gpi_server.html
DGPI	GPI-anchor prediction	129.194.185.165/dgpi/index_en.html
Tools for glycan structure analysis		
Glycofragment	Masses from glycan fragments	www.dkfz.de/spec/projekte/fragments/
GlycoSearchMS	MS-spectrum comparison	www.dkfz.de/sweetdb/
GlycoMod	Glycan structure from MolPeak	www.expasy.org/tools/glycomod/
GlycoMass	Masses from compositions	www.expasy.org/tools/glycomod/glycanmass.html
GlyPeps	Glycoprotein detection	www.dkfz.de/glypeps/
CASPER	$^{1}H,^{13}C$-NMR estimation	www.casper.organ.su.se/casper/
SugarBase	$^{1}H,^{13}C$-NMR search	boc.chem.uu.nl/sugabase/sugabase.html
NMR-Search	$^{1}H,^{13}C$-NMR search	www.dkfz.de/sweetdb/
Graphical representations and nomenclature		
LINUNCS	Linear encoding of sugars	www.dkfz.de/spec/linucs/
LiGraph	Graphical representation	www.dkfz.de/spec/ligraph/
IUPAC	Nomenclature	www.chem.qmw.ac.uk/iupac/2carb/
3D structures		
SWEET-II	Generation of 3D structure.	www.dkfz.de/spec/sweet2/
Disaccharides	Conformation maps	www.cermav.cnrs.fr/cgi-bin/di/di.cgi
Glydicts	Ensemble of glycan conformations	www.dkfz.de/spec/glydict/
GlycoMaps DB	Conformation maps	www.dkfz.de/spec/glycomaps/
DynamicMolecules	Molecular dynamics of glycans	www.md-simulations.de
Carbohydrate databases		
CarbBank	Complex Carbohydrate Research Center, Athens, USA	www.boc.chem.uu.nl/sugabase/carbbank.html
Glycociences_DB	DKFZ-Heidelberg	www.glycosciences.de
Glycan	KEGG Kyoto Encyclopedia of Genes and Genomes	www.genome.ad.jp/ligand/ glycan.genome.ad.jp
Carbohydrate DB	Consortium Functional Glycomics	web.mit.edu/glycomics/carb/carbdb.shtml
GlycoSuite	Proteome Systems Ltd	www.glycosuite.com/
Glycomic DB	GlycoMinds	www.glycominds.com/GlycoInfo.asp
General information, meetings, discussions, journals		
Glycoforum	Carbohydrates Coming of Age	www.glycoforum.gr.jp/
GlycoWord	Introduction to glycosciences	www.glycoforum.gr.jp/science/word/wordE.html
Glycobiology Res.	Introduction to glycanstructures	glycores.ncifcrf.gov
Glycobiology	US Society for Glycobiology	www.glycobiology.org
CFG	Consortium Functional Glycomics	web.mit.edu/glycomics/consortium/main.shtml
Glycobiology	Journal; Society for Glycobiology	glycob.oupjournals.org/
Carbohydrate Research	Journal	www.sciencedirect.com/science/journal/00086215
Carbohydrate Chemistry	Journal	www.dekker.com/servlet/product/productid/CAR

to produce volatile compounds and detection with gas-liquid chromatography coupled to a mass spectrometer.

- Complete derivatisation or tagging of the reducing end and detection with fast-atom bombardment mass spectrometry.

- Direct characterisation/profiling of oligosaccharides from 1- and 2-dimensional gel separated proteins, characterisation of oligosaccharides released from proteins and detection by matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) and/or tandem electrospray mass spectrometry.

Different glycan profiles describe different metabolic states

The CFG for example uses the following protocol: after the extraction of proteins and their tryptic digestion, the attached sugar moieties are cleaved off by an enzymatic digestion with peptide *N*-glycosidase F (PNGaseF). The permethylated *N*-glycans are subsequently analysed with MALDI-TOF MS and the assignments of peaks are based on compositions taking biosynthetic considerations into account. Other sequences are possible. In such cases, electrospray tandem mass spectrometry will differentiate alternative possibilities.

Currently, the assignment of MS-peaks is mainly done manually and only a few applications are available to support this process. However, high-throughput techniques will soon overwhelm the current capacity of methods if no automation is incorporated into glycomics.

ONLINE TOOLS ASSISTING MS SPECTRA INTERPRETATION

GlycanMass, GlycoMod and GlycoFragment can be used to assign peaks in glycan MS-spectra

'GlycanMass'[62] is a simple web tool that enables the mass of an oligosaccharide composition to be calculated. '*GlycoMod*'[63] is designed to find all possible compositions of a glycan structure from its experimentally determined mol peak. The program searches for all combinations of composition agreeing with the mass read in. It can be used to predict the composition of any glycoprotein-derived oligosaccharide comprising either underivatised, methylated or acetylated monosaccharides, or with a derivatised reducing terminus.

The web tool GlycoSearchMS can be used to identify glycans by comparing a measured MS-spectrum with calculated US-spectra

The '*GlycoFragment*'[64] tool allows the generation of all theoretically possible A-, B-, C-, X-, Y- and Z-fragments of oligosaccharides according the definitions of Domon and Costello.[65] The extended IUPAC nomenclature[47] is used to input structures. Several forms of derivatisation and substitution of the reducing end are implemented.

AUTOMATIC MS SPECTRUM INTERPRETATION

The appearance of several papers in recent years[64,66–70] describing algorithms for automatic MS spectrum interpretation reflects the fact that there is an urgent need for such tools. The solutions reported either use databases of known glycan structures or are purely theoretical *de novo* approaches, seeking to determine oligosaccharide structures incrementally, monosaccharide by monosaccharide, from the fragments observed. While the success of database approaches depends heavily on the completeness of the searched glycan structures, the *de novo* methods suffer from the exponential explosion of the solution space when the mass searched increases. Accordingly, authors pragmatically constrained their algorithm to match only structures having a small number of monosaccharides or restrict the search space to specific classes of glycan-like *N*-glycans and take into account biosynthetic knowledge.

A database algorithm similar to the *MASCOT*[71] approach widely used for peptide identification in proteomics projects seems to be the most promising approach for the rapid identification of *N*- and *O*-glycans in high-throughput glycomics projects. Recently, the '*GlycoFragment*'[64] algorithm was used to create databases containing all theoretically possible fragments of about 5,000 *N*- and 1,200 *O*-glycans. The '*GlycoSearchMS*'[70] algorithm compares each peak of a measured MS spectrum with the calculated fragments of all entries contained in the database. The number of matched peaks within a certain tolerance is used to compute a score by which the best matching spectra are ranked. The reliability of results retrieved by search algorithms heavily depends on the completeness of the underlying data collection. Looking at the various worldwide activities to create glyco-related databases, there is a realistic hope that sufficiently large databases of glycan structures will soon be available that can be used to calculate comprehensive lists of theoretical fragments.

3D STRUCTURES OF GLYCANS

Knowledge of the 3D structure of a glycan and its dynamics is a prerequisite for a full understanding of the many biological processes oligosaccharides are involved in. Unfortunately, glycan structures are highly flexible structures, which are difficult to crystallise. Therefore, only a limited number of experimentally solved glycan structures are available. A recent study[28] detected 1,562 entries in the PDB containing a total of 5,397 carbohydrate chains. The majority of chains are found to be *N*-glycosidically bound. Non-covalently bound ligands are also frequent, while *O*-glycans form a minority. The *PDB2LINUCS* web interface provides an online access to all carbohydrate-related information contained in PDB and enables an analysis of all structural parameters defining the conformation of glycans.

GlycoMapsDB contains 700 free energy maps of disaccharide high-temperature simulation

***PDB2LINUCS* provides online access to carbohydrate-related information in PDB**

Small and medium-sized organic molecules can be found in the Cambridge Structural Database (CSD),[72,73] which contains about 4,000 entries classified as carbohydrates (about 1.5 per cent). However, a large number of these are cyclodextrin inclusion complexes and synthetic intermediates, which are of limited biological interest. The CSD is not freely available and does not have a web interface. Nevertheless, it is widely available, at least in the crystallographic community.

The '*Dynamic Molecules*' service uses 3D-structures generated by *SWEET-II*

COMPUTATIONAL APPROACHES TO EXPLORE THE CONFORMATIONAL SPACE OF GLYCANS

NMR-derived structural constraints in combination with computational methods are the most frequently used techniques to investigate the dynamic behaviour of the spatial structure of complex carbohydrates.[74,75] The most often used computational approaches are systematic searches where the relevant torsion angles are systematically changed and the associated energies are calculated, Metropolis Monte Carlo approaches and, increasingly, molecular dynamics simulations.[76–78] The results are often presented by the 3D structure of the lowest energy conformation or as conformational maps for each glycosidic linkage indicating iso-energetic areas as a function of the torsion angles Φ and Ψ. Several data collections providing energy surfaces are available, the more extensive one, '*GlycoMapsDB*' (see Table 1) contains about 700 Φ, Ψ free energy maps obtained from high-temperature simulations of disaccharides.

The well-established *SWEET-II*[48,49] web interface uses a comprehensive collection of rather crude conformational maps to rapidly generate one realistic glycan conformation from many. The extended alphanumeric nomenclature is required as input. The generated 3D structures are mainly thought to be used as starting points for further refinement using more comprehensive computational techniques. The '*Dynamic Molecules*' services[79] – the first internet portal which provides an interactive access to set up, perform and analyse molecular dynamic simulations – can take 3D structures created with *SWEET-II* and provides many features especially devoted to investigating and analysing the dynamic behaviour of complex carbohydrates. The interactive analysis of time dependencies of any interesting degrees of freedom, free energy conformational maps and support for the interpretation of experimental findings mainly derived from NMR spectroscopy are provided by the '*Dynamic Molecules*' web interface.

CONCLUSIONS AND PERSPECTIVES

Since glycans are secondary gene products and their structure cannot be easily predicted from DNA sequences, very few of the bioinformatics algorithms and techniques developed for genomics and proteomics research can be directly adapted for glycomics. Despite being macromolecules, glycan structures require informatics approaches more similar to those developed for small molecules than for proteins and nucleic acids.

Cross-referencing of glycorelated data leads to synergistic effects

In spite of recent improvements it is obvious that the analysis of the glycome research is still in its infancy compared with the extent of research on genes and proteins. This is probably mainly due to the complexity and heterogeneity of glycan structures, making them difficult to study. Glycan structures cannot be readily obtained and identified because they (a) cannot be amplified as nucleic acids, (b) show heterogeneous patterns of structures bound to a single protein and (c) often form non-linear, branched molecules. Moreover, there is currently no universal method to determine precisely their structure without taking into account biosynthetic considerations.

However, the progressing glycomics projects will dramatically accelerate the understanding of the roles of carbohydrates in cell communication and hopefully lead to novel therapeutic approaches for treatment of human disease. The MIT's magazine *Technology Review* (21st January 2003) has identified glycomics as one of the top ten technologies that will change the future. The development of new and advanced bioinformatics tools, algorithms and data collections for glycobiology is an absolute requirement to manage and analyse successfully the large amount of data that will be produced by the upcoming glycomics projects. To screen the glycan content of various tissues using high-throughput techniques and HPLC and/or MS to detect glycans, automatic and reliable procedures for spectra interpretation are required. The implementation and testing of suitable algorithms are currently the most active fields of the development of bioinformatics applications for glycobiology.

Glycomics is one of the top ten technologies supposed to change the future

Another important issue will be the cross-linking of glycobiology resources with existing genomic and proteomic data collections. The present glyco-related databases are not reciprocally cross-referenced like many gene and protein databases. There are currently no generally accepted standards how to normalise glycan structure and how to exchange related data. This unfavourable situation hampers an efficient cross-linking and the automatic exchange of distributed data. Additionally, the two largest projects to collect glycorelated data are run by biotechnology companies. Unfortunately, in the absence of any competing open access project, both companies have decided not to make their primary data available to the public. This development is in clear contrast to the experience gained from genomics and proteomics projects, which have demonstrated that unrestricted dissemination of scientific data accelerates scientific findings, guarantees better quality of data and initiates a number of new initiatives to explore the available experimental data under various scientific questions.

References

1. Taylor, M. E. and Drickamer, K. (2002), 'Introduction to Glycobiology', Oxford University Press, Oxford.
2. Varki, A., Esko, J. and Freeze H. (1999), 'Essential of Glycobiology', Cold Spring Harbor Laboratory Press, New York.
3. Drickamer, K. and Dell, A. (2001), 'Glycogenomics: The impact of genomics and informatics on glycobiology', in Dell, A., Ed., 'Biochemical Society Symposium (69th); 2001; University of York', Portland Press, London, p. 163.
4. Coutinho, P. and Henrissat, B. (1999), 'The

modular structure of cellulases and other carbohydrate-active enzymes: An integrated database approach', in Ohmiya, K., Hayashi, K., Sakka, K. *et al*., Eds, 'Genetics, Biochemistry and Ecology of Cellulose Degradation', Uni Publishers, Tokyo.

5. Coutinho, P. M., Deleury, E., Davies, G. J. and Henrissat, B. (2003), 'An evolving hierarchical family classification for glycosyltransferases', *J. Mol. Biol.*, Vol. 328, pp. 307–317.

6. Feizi, T. and Mulloy, B. (2003), 'Carbohydrates and glycoconjugates. Glycomics: the new era of carbohydrate biology', *Curr. Opin. Struct. Biol.*, Vol. 13, pp. 602–604.

7. Rudd, P. M., Colominas, C., Royle, L. *et al*. (2001), 'A high-performance liquid chromatography based strategy for rapid, sensitive sequencing of *N*-linked oligosaccharide modifications to proteins in sodium dodecyl sulphate polyacrylamide electrophoresis gel bands', *Proteomics*, Vol. 1, pp. 285–294.

8. Dell, A. and Morris, H. R. (2001), 'Glycoprotein structure determination by mass spectrometry', *Science*, Vol. 291, pp. 2351–2356.

9. Rudd, P. M., Elliott, T., Cresswell, P. *et al*. (2001), 'Glycosylation and the immune system', *Science*, Vol. 291, pp. 2370–2376.

10. Peracaula, R., Tabares, G., Royle, L. *et al*. (2003), 'Altered glycosylation pattern allows the distinction between prostate-specific antigen (PSA) from normal and tumor origins', *Glycobiology*, Vol. 13, pp. 457–470.

11. Peracaula, R., Royle, L., Tabares, G. *et al*. (2003), 'Glycosylation of human pancreatic ribonuclease: Differences between normal and tumour states', *Glycobiology*, Vol. 13, pp. 227–244.

12. Rudd, P. M., Merry, A. H., Wormald, M. R. and Dwek, R. A. (2002), 'Glycosylation and prion protein', *Curr. Opin. Struct. Biol.*, Vol. 12, pp. 578–586.

13. Butler, M., Quelhas, D., Critchley, A. J. *et al*. (2003), 'Detailed glycan analysis of serum glycoproteins of patients with congenital disorders of glycosylation indicates the specific defective glycan processing step and provides an insight into pathogenesis', *Glycobiology*, Vol. 13, pp. 601–622.

14. Sutton-Smith, M., Morris, H. R., Grewal, P. K. *et al*. (2002), 'MS screening strategies: investigating the glycomes of knockout and myodystrophic mice and leukodystrophic human brains', *Biochem Soc Symp*, Vol. 69, pp. 105–115.

15. Stanley, P. (2002), 'Biological consequences of overexpressing or eliminating *N*-acetylglucosaminyltransferase-TIII in the mouse', *Biochim Biophys Acta*, Vol. 1573, pp. 363–368.

16. Martin, L. T., Marth, J. D., Varki, A. and Varki, N. M. (2002), 'Genetically altered mice with different sialyltransferase deficiencies show tissue-specific alterations in sialylation and sialic acid 9-*O*-acetylation', *J. Biol. Chem.*, Vol. 277, pp. 32930–32938.

17. Feizi, T., Fazio, F., Chai, W. and Wong, C. H. (2003), 'Carbohydrate microarrays – a new set of technologies at the frontiers of glycomics', *Curr. Opin. Struct. Biol.*, Vol. 13, pp. 637–645.

18. Schadt, E. E., Li, C., Su, C. and Wong, W. H. (2003), 'Analyzing high-density oligonucleotide gene expression array data', *J. Cell. Biochem.*, Vol. 80, pp. 192–202.

19. Comelli, E. M., Amado, M., Head, S. R. and Paulson, J. C. (2002), 'Custom microarray for glycobiologists: Considerations for glycosyltransferase gene expression profiling', *Biochem. Soc. Symp.*, Vol. 69, pp. 135–142.

20. Kemmner, W., Roefzaad, C., Haensch, W. and Schlag, P. M. (2003), 'Glycosyltransferase expression in human colonic tissue examined by oligonucleotide arrays', *Biochim. Biophys. Acta*, Vol. 1621, pp. 272–279.

21. Dwek, R. A., Butters, T. D., Platt, F. M. and Zitzmann, N. (2002), 'Targeting glycosylation as a therapeutic approach', *Nat. Rev. Drug Discov.*, Vol. 1, pp. 65–75.

22. Venter, J. C., Adams. M. D., Myers, E. W. *et al*. (2001), 'The sequence of the human genome', *Science*, Vol. 291, pp. 1304–1351.

23. Apweiler, R., Hermjakob, H. and Sharon, N. (1999), 'On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database', *Biochim. Biophys. Acta*, Vol. 1473, pp. 4–8.

24. Ben-Dor, S., Esterman, N., Rubin, E. and Sharon, N. (2004), 'Biases and complex patterns in the residues flanking protein *N*-glycosylation sites', *Glycobiology*, Vol. 14, pp. 95–101.

25. Kui Wong, N., Easton, R. L., Panico, M. *et al*. (2003), 'Characterization of the oligosaccharides associated with the human ovarian tumor marker CA125', *J. Biol. Chem.*, Vol. 278, pp. 28619–28634.

26. Gupta, R. and Brunak, S. (2002), 'Prediction of glycosylation across the human proteome and the correlation to protein function', *Pac. Symp. Biocomput.*, Vol. 310–322.

27. Jensen, L. J., Gupta, R., Blom, N. *et al*. (2002), 'Prediction of human protein function from post-translational modifications and localization features', *J. Mol. Biol.*, Vol. 319, pp. 1257–65.

28. Lütteke, T., Frank, M. and von der Lieth, C.-W. (2004), 'Data mining the protein data

bank: automatic detection and assignment of carbohydrate structures', *Carbohydr. Res.*, Vol. 339, pp. 1015–1020.

29. Petrescu, A. J., Milac, A. L., Petrescu, S. M. *et al.* (2004), 'Statistical analysis of the protein environment of *N*-glycosylation sites: Implications for occupancy, structure and folding', *Glycobiology*, Vol. 14, pp. 103–114.

30. Gupta, R., Birch, H., Rapacki, K. *et al.* (1999), 'O-GLYCBASE version 4.0: A revised database of *O*-glycosylated proteins', *Nucleic Acids Res.*, Vol. 27, pp. 370–372.

31. Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997), 'Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites', *Protein Eng.*, Vol. 10, pp. 1–6.

32. Hansen, J. E., Lund, O., Tolstrup, N. *et al.* (1998), 'NetOglyc: Prediction of mucin type *O*-glycosylation sites based on sequence context and surface accessibility', *Glycoconj. J.*, Vol. 15, pp. 115–130.

33. Harvey, D. J. (2001), 'Identification of protein-bound carbohydrates by mass spectrometry', *Proteomics*, Vol. 1, pp. 311–328.

34. Küster, B., Krogh, T. N., Mortz, E. and Harvey, D. J. (2001), 'Glycosylation analysis of gel-separated proteins', *Proteomics*, Vol. 1, pp. 350–361.

35. Sagi, D., Peter-Katalinic, J., Conradt, H. S. and Nimtz, M. (2002), 'Sequencing of tri- and tetraantennary *N*-glycans containing sialic acid by negative mode ESI QTOF tandem MS', *Amer. Soc. Mass Spectrom.*, Vol. 9, pp. 1138–1148.

36. Geyer, H., Schmitt, S., Wuhrer, M. and Geyer R. (1999), 'Structural analysis of glycoconjugates by on-target enzymatic digestion and MALDI-TOF-MS', *Anal. Chem.*, Vol. 71, pp. 476–482.

37. Duus, J. O., Gotfredsen, C. H. and Bock, K. (2000), 'Carbohydrate structural determination by NMR spectroscopy: Modern methods and limitations', *Chem. Rev.*, Vol. 100, pp. 4589–4614.

38. Manzi, A. E., Norgard-Sumnicht, K., Argade, S. *et al.* (2000), 'Exploring the glycan repertoire of genetically modified mice by isolation and profiling of the major glycan classes and nano-NMR analysis of glycan mixtures', *Glycobiology*, Vol. 10, pp. 669–689.

39. Berteau, O. and Stenutz, R. (2004), 'Web resources for the carbohydrate chemist', *Carbohydr. Res.*, Vol. 339, pp. 901–1020.

40. von der Lieth, C. W. (2004), 'An appoval to create open access databases for analytical data of complex carbohydrates', *J. Carbohydr. Chem.*, in press.

41. Bartolozzi, A. and Seeberger, P.H. (2001), 'New approaches to the chemical synthesis of bioactive oligosaccharides', *Curr. Opin. Struct. Biol.*, Vol. 11, pp. 587–592.

42. Seeberger, P. H. (2003), 'Automated carbohydrate synthesis to drive chemical glycomics', *Chem. Commun.*, Vol. 10, pp. 1115–1121.

43. Hewitt, M. C., Snyder, D. A. and Seeberger, P. H. (2002), 'Rapid synthesis of a glycosylphosphatidylinositol-based malaria vaccine using automated solid-phase oligosaccharide synthesis', *J. Amer. Chem. Soc.*, Vol. 124, pp. 13434–13446.

44. Galperin, M. Y. (2004), 'The Molecular Biology Database Collection: 2004 update', *Nucleic Acids Res.*, Vol. 32, pp. D3–D22.

45. Laine, R. A. (1994), 'A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10(12) structures for a reducing hexasaccharide: The isomer barrier to development of single-method saccharide sequencing or synthesis systems', *Glycobiology*, Vol. 4, pp. 759–767.

46. URL: http://glycomics.scripps.edu/CFGnomenclature.pdf

47. McNaught, A. D. (1997), 'Nomenclature of carbohydrates (recommendations 1996)', *Adv. Carbohydr. Chem. Biochem.*, Vol. 52, pp. 43–177.

48. Schwardt, O., Kolb, H. and Ernst, B. (2003), 'Drug discovery today', *Curr. Top Med. Chem.*, Vol. 3, pp. 1–9.

49. Bohne, A., Lang, E. and von der Lieth, C. (1998), 'W3-SWEET: Carbohydrate modeling by internet', *J. Mol. Model.*, Vol. 4, pp. 33–43.

50. Bohne, A., Lang, E. and von der Lieth, C.W. (1999), 'SWEET – WWW-based rapid 3D construction of oligo- and polysaccharides', *Bioinformatics*, Vol. 15, pp. 767–768.

51. Bohne-Lang, A., Lang, E., Forster, T. and von der Lieth, C.W. (2001), 'LINUCS: Linear notation for unique description of carbohydrate sequences', *Carbohydr. Res.*, Vol. 336, pp. 1–11.

52. Banin, E., Neuberger, Y., Altshuler, Y. *et al.* (2002), 'A novel linear code nomenclature for complex carbohydrates', *Trends Glycosci. Glycotech.*, Vol. 14, pp. 127–137.

53. Doubet, S. and Albersheim, P. (1992), 'CarbBank', *Glycobiology*, Vol. 2, pp. 505.

54. Loss, A., Bunsmann, P., Bohne, A. *et al.* (2002), 'SWEET-DB: An attempt to create annotated data collections for carbohydrates', *Nucleic Acids Res.*, Vol. 30, pp. 405–408.

55. van Kuik, J. A., Hard, K. and Vliegenthart, J. F. (1992), 'Databases of complex carbohydrates', *Trends Biotechnol.*, Vol. 10, pp. 182–185.

56. Cooper, C. A., Joshi, H. J., Harrison, M. J. *et al.* (2003), 'GlycoSuiteDB: A curated

relational database of glycoprotein glycan structures and their biological sources. 2003 update', *Nucleic Acids Res*., Vol. 31, pp. 511–513.

57. Stenutz, R., Jansson, P. E. and Widmalm, G. (1998), 'Computer-assisted structural analysis of oligo- and polysaccharides: An extension of CASPER to multibranched structures', *Carbohydr. Res*., Vol. 306, pp. 11–17.

58. Marchal, I., Golfier, G., Dugas, O. and Majed, M. (2003), 'Bioinformatics in glycobiology', *Biochemie*, Vol. 85, pp. 75–81.

59. Kanehisa, M., Goto, S., Kawashima, S. *et al.* (2004), 'The KEGG resource for deciphering the genome', *Nucleic Acids Res*., Vol. 32, pp. D277–D280.

60. Goto, S., Okuno, Y., Hattori, M. *et al.* (2002), 'LIGAND: Database of chemical compounds and reactions in biological pathways', *Nucleic Acids Res*., Vol. 30, pp. 402–404.

61. Doubet, S., Bock, K., Smith, D. *et al.* (1989), 'The complex Carbohydrate Structure Database', *TIBS*, Vol. 14, pp. 475–477.

62. Appel, R. D., Bairoch, A. and Hochstrasser, D. F. (1994), 'A new generation of information retrieval tools for biologists: The example of the ExPASy WWW server', *Trends Biochem. Sci*., Vol. 19, pp. 258–260.

63. Cooper, C. A., Gasteiger, E. and Packer, N. H. (2001), 'GlycoMod – a software tool for determining glycosylation compositions from mass spectrometric data', *Proteomics*, Vol. 1, pp. 340–349.

64. Lohmann, K. K. and von der Lieth, C.-W. (2003), 'GLYCO-FRAGMENT: A web tool to support the interpretation of mass spectra of complex carbohydrates', *Proteomics*, Vol. 3, pp. 2028–2035.

65. Domon, B. and Costello, C.E. (1988), 'A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates', *Glycoconjugate*, Vol. 5, pp. 397–409.

66. Gaucher, S. P., Morrow, J. and Leary, J. A. (2000), 'STAT: A saccharide topology analysis tool used in combination with tandem mass spectrometry', *Anal. Chem*., Vol. 72, pp. 2331–2336.

67. Ethier, M., Saba, J. A., Ens, W. E. *et al.* (2002), 'Automated structural assignment of derivatized complex N-linked oligosaccharides from tandem mass spectra', *Rapid Commun. Mass Spectrom*., Vol. 16, pp. 1743–54.

68. Ethier, M., Saba, J. A., Spearman, M. *et al.* (2003), 'Application of the StrOligo algorithm for the automated structure assignment of complex *N*-linked glycans from glycoproteins using tandem mass spectrometry', *Rapid Commun. Mass Spectrom*., Vol. 17, pp. 2713–2720.

69. Clerens, S., Van den Ende, W., Verhaert, P. *et al.* (2004), 'Sweet Substitute: A software tool for in silico fragmentation of peptide-linked *N*-glycans', *Proteomics*, Vol. 4, pp. 629–632.

70. Lohmann, K. K. and von der Lieth, C. W. (2003), 'Glyco-Search-MS: A web-based tool to support the rapid identification of *N*- and *O*-glycans in MS spectra', *Glycobiology*, Vol. 13, abstract 74.

71. Perkins, D. N., Pappin, D. J., Creasy, D. M. and Cottrell, J. S. (1999), 'Probability-based protein identification by searching sequence databases using mass spectrometry data', *Electrophoresis*, Vol. 20, pp. 3551–3567.

72. Allen, F. H. (2002), 'The Cambridge Structural Database: A quarter of a million crystal structures and rising', *Acta Crystallogr. B*., Vol. 58, pp. 380–388.

73. URL: http:// www.ccdc.cam.ac.uk

74. von der Lieth, C. W., Siebert, H. C., Kozar, T. *et al.* (1998), 'Lectin ligands: New insights into their conformations and their dynamic behavior and the discovery of conformer selection by lectins', *Acta Anat. (Basel)*, Vol. 161, pp. 91–109.

75. Kogelberg, H., Solis, D. and Jimenez-Barbero, J. (2003), 'New structural insights into carbohydrate–protein interactions from NMR spectroscopy', *Curr. Opin. Struct. Biol*., Vol. 13, pp. 646–653.

76. Woods, R.J. (1988), 'Computational carbohydrate chemistry: What theoretical methods can tell us', *Glycoconj. J*., Vol. 15, pp. 209–216.

77. Bush, C. A., Martin-Pastor, M. and Imberty, A. (1999), 'Structure and conformation of complex carbohydrates of glycoproteins, glycolipids, and bacterial polysaccharides', *Annu. Rev. Biophys. Biomol. Struct*., Vol. 28, pp. 269–293.

78. Kozar, T. and von der Lieth, C. W. (1997), 'Efficient modelling protocols for oligosaccharides: From vacuum to solvent', *Glycoconj. J*., Vol. 14, pp. 925–933.

79. Frank, M., Gutbrod, P., Hassayoun, C. and von der Lieth, C. W. (2003), 'Dynamic molecules: Molecular dynamics for everyone. An internet-based access to molecular dynamic simulations: Basic concepts', *J. Mol. Model*., Vol. 9, pp. 308–315.