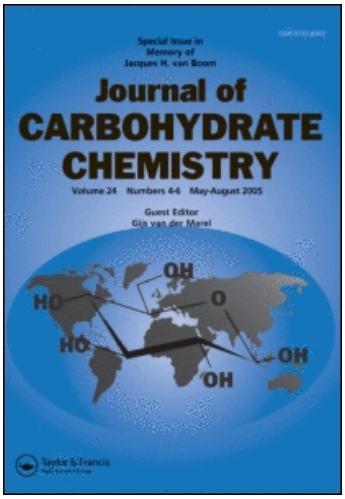

Journal of Carbohydrate Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713617200

An Endorsement to Create Open Access Databases for Analytical Data of Complex Carbohydrates

Claus-W. von der Lieth[a]
[a] Central Spectroscopic Department B090, German Cancer Research Center, Heidelberg, Germany

Online publication date: 28 November 2004

JOURNAL OF CARBOHYDRATE CHEMISTRY
Vol. 23, No. 5, pp. 277–297, 2004

An Endorsement to Create Open Access Databases for Analytical Data of Complex Carbohydrates

Claus-W. von der Lieth*

German Cancer Research Center, Central Spectroscopic Department B090, Heidelberg, Germany

CONTENTS

*Correspondence: Claus-W. 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.

DOI: 10.1081/CAR-200030093
0732-8303 (Print); 1532-2327 (Online)

ABSTRACT

One of the aims of the emerging glycomics projects is to create a cell-by-cell catalogue of detected glycan structures. Mass spectrometry (MS) and NMR in combination with separation techniques are the most intensively applied experimental methods for the analysis of carbohydrates. Unlike genome and proteome databases, development of carbohydrate databases has gained a broader attention only recently. However, no spectral libraries of suitable pure and homogeneous standards have been compiled so far. The difficulties to describe complex carbohydrate structures are discussed and an overview of currently available data collections and applications is given. The current situation in glycobiology is characterized by a nearly complete loss of all primary analytical data. The Internet has fundamentally changed the practical and economic realities to collect and distribute scientific data. Four suitable approaches how to organize the updating process for analytical data collections in the field of glycosciences are discussed. It is anticipated that open access data collections provide a better dissemination of scientific data, quicken scientific findings, guarantee better quality of data and initiate a number of new initiatives to explore the available experimental data under various scientific questions. Therefore, any new initiative in glycosciences should be established under the open access philosophy.

Key Words: Carbohydrate databases; Carbohydrate structures; Glycomics; Open access; Dissemination of primary scientific data.

1. INTRODUCTION

The human genome seems to encode for not more than 30,000–40,000 proteins. 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 proteins in health and disease. More than half of all the proteins in the human body have carbohydrate molecules attached.[1,2] The term "*glycomics*" describes the scientific attempt of

identifying and studying all of 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 HPLC techniques are currently applied to provide information on the glycan repertoire of cells, tissues, and organs.[3,4] One of the aims of the emerging glycomics projects is to create a cell-by-cell catalogue of glycosyltransferase expression and detected glycan structures.

In recent years, MS has become the method of choice for high sensitive protein[5,6] and glycan[4,7–10] identification and characterization. Fundaments to enable a rapid identification of peptides are the availability of large protein sequence databases and the development of efficient algorithms for MS/MS fragment identification techniques. NMR techniques can lead to a full structural characterization of oligosaccharides including the monosaccharide stereochemistry, the anomeric configuration, the linkage type and the complete sugar sequence.[11] NMR-derived structural constraints in combination with computational methods are the most often used techniques to investigate the dynamical behavior of the spatial structure of complex carbohydrates.[12,13] However, NMR spectroscopy is relatively insensitive with respect to the amount of sample needed in order to obtain good-quality structural data.[14] 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 needed additional steps to scale up the amount of produced oligosaccharides and their purification normally exclude NMR techniques to be applied in high-throughput sequencing projects.

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.[15–17] Such investigations benefit pivotally from the detailed structural information provided by NMR spectroscopy.

MS- and NMR-spectra of glycans can be complicated and difficult to interpret without reference data. Unfortunately, no MS spectral libraries of suitable pure and homogeneous standards have been compiled so far. With *SugaBase*,[18] a database of about 1500 ^{1}H and ^{13}C NMR spectra is freely available. However, updating this database has ended in 1998. Several authors pointed out,[11,19,20] that the rapid and automatic detection of glycan structures as required for high-throughput projects depends heavily on sufficiently large high quality collections of reference spectra. Regrettably, after the failure of *CarbBank*[21] and *SugaBase*,[18] the two major database projects in glycobiology during the nineties, only recently have new attempts been proposed to overcome this obvious lack of spectrometric reference data. The US Consortium for Functional Glycomics (http://web.mit.edu/glycomics/consortium/) has adumbrated to develop a carbohydrate database which will contain various data sets pertaining to glycan structures. According to the announcement, the database will include biochemical data such as chemical structure, MS profiles, NMR spectra, and HPLC profile. However, no detailed information is currently available. Additionally, two larger projects to collect glyco-related data are run by biotech companies. Unfortunately, in the absence of any competing public project, both companies have decided not to provide open access to their data. This development is in contrast to the genomics and proteomics projects, where it was clear from the beginning, that all sequence data will be accessible freely for everyone, even those sequences determined by companies. This development underspins the imperative to start new attempts to collect scientific data that will be freely available since otherwise the developement of bioinfomatic tools for glycomics will have a hard time to reach a similar level of acceptance and achievement as has been attained for genomics and proteomics.

The aim of this paper is to survey the already existing isolated approaches available on the web to collect and interpret analytical data for application in glycosciences and to discuss some new ideas how to assemble analytical data in the light of the new prospects provided by Internet technologies.

2. CARBOHYDRATE STRUCTURES, DESCRIPTIONS, AND NOMENCLATURE

All existing digital data collections in glycobiology[18,22–26] use a representation of the chemical structure as the primary key to access related bibliographic, biological, chemical, or physical data. However, a general representation of carbohydrates accepted likewise by scientists from all disciplines as the one-letter code of amino acid sequence for proteins does not exist in glycosciences.

Unlike oligonucleotides and proteins, oligosaccharides are not just linear oligomers, they are often branched and can be linked in a number of different ways. The number of monomers detected in nature for all classes of carbohydrates, not counting the numerous chemical modifications, approaches one hundred. These residues can be connected using not only classical glycosidic linkages (acetals), but also by phosphodiesters, either directly or through an alditol. Many carbohydrates also show varying degrees of heterogeneity. However, for specific scientific questions only certain classes of glycans showing a limited number of residues and linkages have to be examined. For example, the analysis of *N*-linked oligosaccharides revealed that normally only nine monosaccharides units are present in mammalian cells. Therefore, it is not surprising that for different scientific questions specific descriptions to characterize carbohydrates are used.

The most general characterization of a carbohydrate is its total mass or less frequently used the gross molecular formula. Since MS as the most often used experimental method to identify glycans cannot distinguish between isomeric monosaccharides, which have the same mass, like glucose, galactose, or mannose, often the composition of an oligosaccharide like $Hex_5HexNAc_2dHex_1$ is reported (see Fig. 1). *N*-Glycan patterns are often characterized using the three main categories (high mannose, hybrid, and complex) and/or other topological descriptions as, e.g., *neutral complex diantennary structures, with smaller amounts of tri- and tetra-antennary compounds* or *fucosylated hybrid and complex diantennary glycans with GalNAc–GlcNAc antennae*. Since the exact linkage information is often not available and can only be estimated on calling in 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.

A full structural characterization has to include the complete sugar sequence, the monosaccharide stereochemistry, the anomeric configuration, and the linkage information. In 1996, IUPAC–IUBMB "Nomenclature of Carbohydrates"[27] specified several ways how to uniquely describe complex oligosaccharides based on a three-letter code to characterize monosaccharide units (gal = galactose, 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

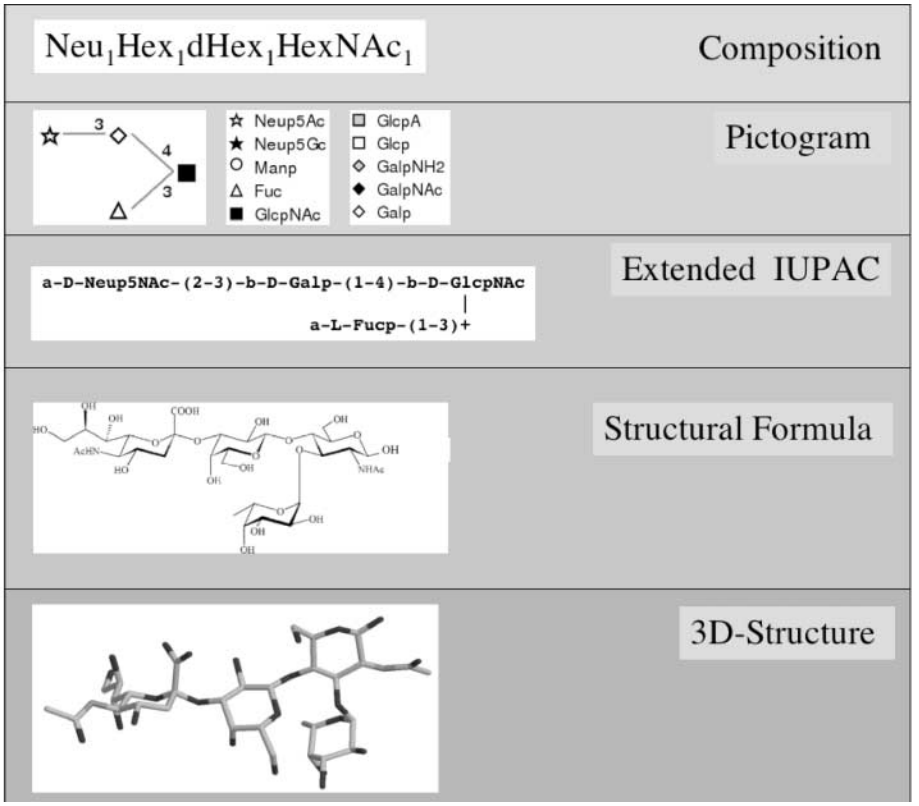

Figure 1. Several representations are used to describe carbohydrate structures. MS spectrometry normally provides information on the composition of glycans. Pictograms are often used for classes of oligosaccharides, where only a limited number of residues exist. The extended IUPAC nomenclature enables full structural characterization of regular oligosaccharides based on a three-letter code for each monosaccharide residue. This alphanumeric representation is also well suited for a computer-readable coding of oligosaccharides. Chemists prefer the structural formula in a pseudo 3D representation. The spatial structure is required for drug design studies.

locants of the linkage are given in parentheses between the symbols; an arrow indicates a linkage between two anomeric positions. The branches are displayed in separate lines (see Fig. 2) and are connected by the linkage information. In such a way, long carbohydrate sequences can thus be adequately described in abbreviated form. The extended nomenclature has intensively been used in many biochemically and some chemically oriented papers. Since this alphanumeric description is well suited for computer processing a similar representation has been employed by the carbohydrate databank *CarbBank* and *SugaBase*. Databases containing experimental NMR data require a complete structural description, where each resonance can be assigned to a specific atom.

However, to derive a linear, unique identifier from carbohydrate structures in order to efficiently link glyco-related data from various data collections, the rules provided in 1996

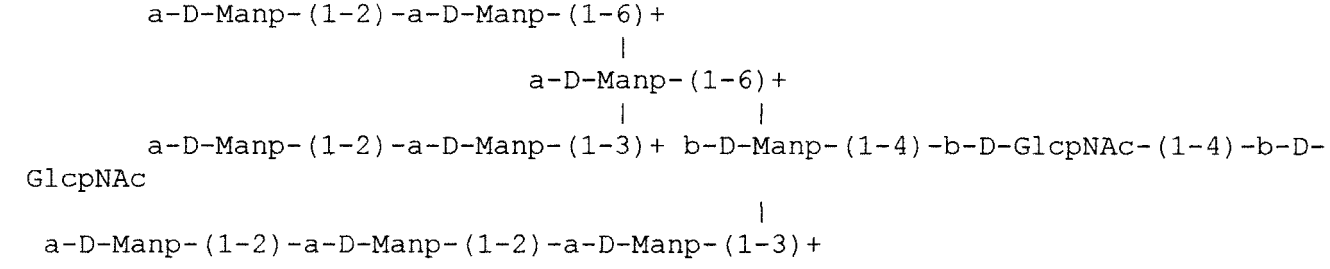

Figure 2. Full structural characterization of a complex carbohydrate using the ASCII description as used in *CarbBank* and *Glycosciences.de*. Monosaccharides are characterized by a three-letter code (e.g. Glc = glucose, Man = mannose). Each monosaccharide uit is preceded by the anomeric descriptor (a for α and b for β) and the configuration symbol (D or L). The ring size is indicated by an italic f for furanose or p for pyranose. The locants of the linkage are given in parentheses between the symbols. The branches are displayed in separate lines and are connected by the linkage information.

IUPAC recommendation are not sufficiently comprehensive to cover all ambiguities especially regarding the ordering of branches. Therefore, two linear codes have been developed. The "*LINUCS*" notation,[28] *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. Thus, the extended nomenclature contains all required information, and no additional rules or lookup tables to define the hierarchy of branches are necessary.

The company "Glycominds" has built their database of glycan structures on the "*Linear Code*™",[25] a simple one to two-letter representation of saccharide units and linkages, hoping to initiate a similar development of bioinformatics applications as the one-letter code for oligonucleotides and proteins has prepared. The ordering of glycan branches is established using a special lookup table, where the hierarchy of monosaccharide structures is defined. The commercial database *GlycoSuite*[24] uses the extended IUPAC nomenclature and has introduced some additional rules to establish the hierarchy of the various branches.[29]

Chemists prefer the commonly used schematic all-atom and bond drawings of molecules where the stereochemistry is indicated graphically at each individual stereo center. These pseudo 3D-pictures are just another representation of the structural information contained in extended alphanumeric IUPAC description. However, there is currently no software available which is capable of converting one representation into the other. Full systematic IUPAC names of carbohydrate structures may become rather complicated.

Although the extended form of the alphanumeric description of carbohydrates allows to indicate a specific substitution at each C-atom and thus a large variety of chemically modified oligosaccharides can be specified, this nomenclature is restricted to a scaffold of regular sugar units which are defined by the three-letter code. Any chemical modification of the sugar ring would require the definition of a new residue. Therefore, the alphanumeric description may not be suitable for characterizing carbohydrate mimetics where non-regular or sugar-like structures are present.

In drug discovery projects, virtual screening approaches are increasingly applied to find molecules, which promise to exhibit 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 publicly available *SWEET-II*[30,31] server provides for most types of carbohydrates an efficient conversion of the extended alphanumeric IUPAC nomenclature into a reliable spatial structure.

3. AVAILABLE ANALYTICAL DATA COLLECTIONS AND PREDICTION OF ANALYTICAL DATA

In a suitable analytical strategy for the structural determination of glycans attached to proteins or lipids, normally the following steps are involved: the sugar moieties are released using either chemical (e.g., hydrazine for *O*-glycans) or enzymatic (PNGase F for *N*-glycan) reagents. Subsequently, exoglycosidases are often applied to remove terminal monosaccharides from the non-reducing end of the oligosaccharide chains and

the produced free sugars will be labeled, if required for detection. The type of label depends on the applied detection method that follows. HPLC needs normally the attachment of a fluorescent or chromatic compound. This is done through a reductive amination of the reducing end of the sugar chain. For the detection with MS techniques, the carbohydrate is often linked to a reagent that contains a functional group which can be easily protonated. Reductive amination with an aromatic amine is one of the commonly applied derivatization techniques. 2-Aminopyridine (2AP) and 2-Aminobenzamidine (2AB) are often used fluorophore, which are suitable for HPLC as well as MS detection. Persubstituted derivatives of carbohydrates are sometimes created to increase their volatility and to improve the sensitivity of MS-ionization techniques. Such derivatives can also be used to acquire linkage information. Fragmentations of sugar rings showing up above the mass of a certain glycosidic fragments provide information about the glycosidic linkage position as a hydroxyl group rather than a methoxy group is left at the site of cleavage following a glycosidic residue loss.

NMR spectroscopy yields almost complete information on the glycan's structure and is, therefore, the most powerful tool for complete structural analysis of an oligosaccraide. Structural information deduced from NMR experiments include (a) the stereochemistry of the sugar residue, (b) the nature of the anomeric linkage, (c) the type of linkage between two sugar residues, and (d) the type and nature of substitutions.

3.1. Interpretation of MS-Spectra

The structural complexity and diversity of carbohydrates makes structure elucidation a challenging task despite the progress in experimental techniques. In recent years, MS has become the method of choice for high sensitive protein as well as glycan identification and characterization. Matrix-assisted laser-desorption ionization–time of flight (MALDI–TOF) and electrospray ionization (ESI) MS experiments are today the most intensively used techniques and have mostly replaces fast atom bombardment (FAB)-MS because of their higher sensitivity, high mass measurement capabilities,and the ability in some cases to analyze underivatized oligosaccharides.

The release of sub-picomole levels of N-linked oligosaccharides directly from 1–5 μg of protein in a band on an SDS–PAGE gel, coupled with MALDI-TOF mass spectrometer and HPLC, is an established technique for the analysis of biologically important glycoproteins that are difficult to purify or are available only in limited amounts.[19,32,33] The approach allows analyzing large numbers of samples at high speed in an automatic procedure. For the automatic identification of proteins the search engine *MASCOT*[34] is routinely used in most proteomics projects. The algorithm implemented in the *MASCOT* software matches MS data against calculated fragmentation pattern derived online from primary sequences contained in protein databases. A comparable search engine for carbohydrates is currently missing and also no libraries of suitably pure and homogeneous standards have been compiled so far. Therefore, the identification of glycan structures is performed manually most of the time with the help of printed tables or a pocket calculator. However, a few web tools to support spectrum interpretation are available. "*GlycanMass*"[35] is a simple web tool which allows calculating the mass of an oligosaccharide composition. "*GlycoMod*"[36] is designed to find all possible compositions of a glycan structure from its experimentally determined molpeak. The program searches for all com-

binations of composition agreeing with the mass read in. It can be used to predict the composition of any glycoprotein-derived oligosaccharide comprised of either underivatized, methylated or acetylated monosaccharides, or with a derivatized reducing terminus. "*GlycoMod*" can apply compositional constraints to the output, if the user supplies either known or suspected monosaccharide constituents. Since a list of experimental masses can be entered, "*GlycoMod*" can generate all possible glycan compositions for each mass.

The appearance of several papers in recent months[37–42] describing algorithms for automatic MS peak assignments mirrors the 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.

The "*GlycoFragment*"[42] tool allows to generate all theoretically possible A-, B-, C-, X-, Y-, and Z- fragments of oligosaccharides according to the definitions of Domon and Costello.[43] The extended IUPAC nomenclature[27] is used to input structures. Several forms of derivatization and substitution of the reducing end are implemented. Recently, the "*GlycoFragment*" algorithm was used to create databases containing all theoretically possible fragments of about 5000 *N*-glycans and 1200 *O*-glycans. Additionally, the masses of inner fragments, two independent glycosidic fragmentations or a single glycosidic and a cross-ring fragmentation, are included. The "*GlycoSearchMS*"[41] 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. For each matched experimental peak, the structure of the associated fragment can be displayed. The reliability of results retrieved by "*GlycoSearchMS*" depends heavily on the comprehensiveness of the data collection searched. Since the database needs only theoretically calculated lists of fragments, the completion of missing structures will be relatively easy. This approach seems to be applicative for the rapid identification of known *N*- and *O*-glycans in high-throughput projects since the procedure is similar to the *MASCOT*[34] approach for peptide identification.

3.2. HPLC

Although HPLC techniques are in use in many labs especially to detect small amount of glycan quantities, there are currently only few resources available on the Internet, which offer access to HPLC data characteristic for specific glycan structures. Probably the main reason for this poverty of available data originates from the fact that experimental procedures and conditions applied as well as the hardware systems used vary considerably from lab to lab. The only way to overcome this problem is to calibrate the applied experimental setup using external standards and fit the measured retention time for individual

glycans to the standard curve. *GlycoBase*[44,45] (maintained by the Oxford Glycobiology Institute) provides access to list of about 300 *O*- and *N*-linked glycans using a symbolic representation to describe glycan structures and normalized values, so-called glucose units (GU), characterizing the retention time. For calibration of the HPLC system hydrolyzed and 2AB-labeled glucose oligomers are used as external standard. The calibration curve for each experimental setup is obtained by plotting the number of glucose residues contained in each peak against the retention time. The obtained GU values for individual glycans are highly reproducible and do only slightly depend on the selected columns and HPLC systems.

3.3. NMR Resources

NMR techniques to analyze oligosaccaride structures may be divided into those which are required for a complete analysis of the primary structure and those needed to elucidate their three-dimensional structure. Here, only tools for primary structure elucidation will be discussed. ^{1}H and ^{13}C NMR chemical shifts are the most often used data for primary structure assignment. The limited dispersion of the chemical shifts requires high accuracy. Determination of the primary structure of a saccharide may still be a time-consuming process and computerized approaches are useful in order to speed up the analysis. Therefore, several computational methods have been proposed for assistance in the assignment of the primary structure of glycans. All published approaches depend heavily on the availability of good experimental NMR data for selected reference compounds.

3.4. Databases

Vliegenthart and co-workers[18,46] developed a ^{1}H and ^{13}C NMR database called *SugaBase*, where chemical shifts and coupling constants are stored. The search is based on the use of ^{1}H chemical shifts from the structural reporter groups. This concept is based on the fact that it is often sufficient to inspect only certain areas of a spectrum to ascertain the primary structure of a common glycoprotein carbohydrate structure. In the structural reporter group approach, the crowded region between 3 and 4 ppm is ignored and only the regions between 4–5.6 and 1–3 ppm are inspected. The anomeric protons, methyl protons, protons attached to a carbon atom in the direct vicinity of a linkage position, and protons attached to deoxy carbon atoms are considered relevant structural reporter groups. Unfortunately, *SugaBase* is no longer being updated. However, the interfaces are still available and the database content has been included into *GlycosciencesDB*.[23] The implemented NMR tools allow retrieving NMR spectra based on (sub)structural search, for atoms in a specific chemical environment and a spectral search where all peaks of the library search are compared with an experimental peak list. The number of matched peaks within a certain tolerance is used to compute a score by which the best matching library spectra are ranked. When comparing chemical shift values, it is important that the reference data is measured at the same temperature and that the data are based on the same internal reference or one that can be correlated in a simple manner.

3.5. Prediction of NMR Shifts

Normally, oligosaccharides do not exhibit a well-defined secondary structure. Therefore, long-range interactions affecting chemical environment of a specific atom are rare. Thus, it is possible to achieve high accuracy for the prediction of chemical shifts using simple additivity schemes.

Computer Aided SPectrum Evaluation of Regular polysaccharides (*CASPER*)[47–49] was originally developed for the analysis of the primary structures of oligosaccharides and for polysaccharides with repeating units. The now available WWW tool is devoted to the assignment of ^{1}H and ^{13}C NMR spectra.[26] It allows both the simulation of the NMR spectra of a specific structures and the sequence determination of unknown structures using unassigned NMR spectra and information from chemical analyses (sugar-, configuration-, and methylation-analysis).

CASPER spectrum estimation of a user definable structure is based on the chemical shifts of stored experimental monosaccharide resonances and the induced chemical shift displacements for disaccharides. More complex spectra are observed for branched structures, but the analysis can be performed using additional correction sets for vicinally disubstituted sugar residues.

To determine a primary oligosaccharide sequence on the basis of NMR shifts, the results from a sugar component analysis and linkage analysis are required. Furthermore, the available ^{1}H and ^{13}C chemical shifts and homo-/hetero-nuclear coupling constants have to be fed in. Starting from this information, all possible structures are generated and their ^{1}H and ^{13}C chemical shifts are estimated. The program sorts the results and removes structures incompatible with the input coupling constants. Subsequently, the simulated spectra are compared with the experimental data and the structures are ranked according to the lowest average total difference in chemical shifts.

4. AN ENDORSEMENT TO CREATE OPEN ACCESS DATABASES

Several authors have pointed out the importance of publishing good analytical data for selected reference compounds. However, journals tend to abolish primary analytical data from being part of the actual publication making it if at all available as supplementary material. Further on, only a filtered version of the primary data, e.g., the list of NMR shifts and not the complete spectrum, is normally reported. The lesson to learn from open access data collections like the human genome sequencing project and the protein data collections is, that a better dissemination of scientific data quickens scientific findings, guarantees better quality of data and initiates a number of new initiatives to explore the available experimental data under various scientific questions. Also, open source software projects like the operating system Linux and other programs available under the General Public License have produced evidence that better and safer computer programs are resulting which can be adopted faster to actual needs.

4.1. How to Collect and Disseminate Scientific Data in the Internet Area?

The traditional way to collect spectroscopic data and build-up specific data collections was to extract data manually from published papers. *SpecInfo*[50,51] and *CSearch*[52]

including several hundred thousands of ^{13}C NMR spectra were compiled in this way. Also *CarbBank*[22] and *SugaBase*[18,46] are typical representatives of this procedure. However, the Internet has fundamentally changed the practical and economic realities of collecting and distributing scientific data. For the first time ever, the Internet now offers the chance to constitute a global and interactive peer-to-peer communication for scientific data. In the light of these new technological developments also new ways to collect and disseminate scientific data are emerging. Such potentialities will here be discussed for the spread of analytical data in the field of glycobiology.

Table 1 summarizes the advantages and disadvantages of the various thinkable options to build-up scientific databases. The traditional way of a centralized database with central input of new data from publications by well-trained scientists is a well-established approach but quite expensive. The fates of *CarbBank* and *SugaBase* suggest that this attempt seems to be not a practical way for highly specialized data collections with a limited amount of possible users.

The availability of scientific abstracts and original articles in a digital form has opened the chance to automatically scan these resources and filter out the desired articles using appropriate text mining tools. Searching regularly and in fully automatic manner through all new abstracts contained in *PubMed*[53] allows identifying those abstracts, which are relevant for a specific topic like, e.g., glycobiology.[54,55] However, an unsupervised inclusion of the retrieved abstracts into the master collection seems to be only reasonable in a limited number of cases, where the vocabulary used to classify an abstract unambiguously identifies its relevance. Therefore, a two-step procedure is normally used, where first the retrieved abstracts are automatically stored into a working database and are included in the master collection after inspection and annotation by an expert. *SwissProt/Trembl*[56] and *Auto-Sweet-DB/GlycosciencesDB* are examples of this procedure. At present, text mining approaches are efficient tools to comprehensively scan the literature in an automatic way and to detect relevant publications. Thus, this approach is appropriate to configure this necessary work in an efficient way (Table 1).

The decentralized input done by scientists who recorded and evaluated experimental data into a centralized database is a well-established procedure to update genomic and proteomic sequence collections. Also, the spatial structures of small molecules, proteins, and nucleotides determined by x-ray crystallography have to be deposited prior to publication. However, a similar generally accepted procedure does not exist for NMR- and MS-spectra, which are important to determine the primary structure of oligosaccharides. Since many journals tend to abolish analytical data from being reported, and in the absence of any centralized data collections where spectra can be deposited, there is a high risk, that the recorded spectra may be lost for future scientific use. With *JCAMP-DX*[57,58] (http://www.jcamp.org/protocols.html) suitable data formats exist how to exchange NMR- as well as MS-spectra. However, there has been no or only little success to encourage scientists to build-up larger open access libraries of NMR- and MS-spectra for small organic molecules. *NMRShiftDB* (http://www.nmrshiftdb.org/) is the first open source, open access, open submission, and open content web database for chemical structures and their associated nuclear magnetic resonance data. However, since *NMRShiftDB* does not encode stereochemistry, it cannot be used to deposit oligosaccharide structures.

The Internet offers the chance to constitute a worldwide and interactive peer-to-peer communication for scientific data. Peer-to-peer is a communication model in which each party has the same capabilities and either party can initiate a communication session. In

Table 1. Advantages and disadvantages of strategies to collect scientific data for glycosciences.

	Central DB with central input	Central DB with automatic input	Central DB with local input	Local DBs with local input
Advantages	Consistent data, high quality, comprehensive, complex relations	Inexpensive, quick, automatic (comprehensive)	High quality, primary data, consistent data	Inexpensive, primary data, high quality, consistent data
Disadvantages	Expensive, no primary data	No quality check, no primary data, no complex data (inconsistent data)	Expensive	Difficult to maintain (loss of data), not comprehensive
Examples	*GlycoSuite*, *CarbBank*, *SugaBase*, *SwissProt*	*Trembl*, *Auto-Sweet-DB*	*PDB*, *GenBank* (*CarbBank*)	? (The future ?)

Table 2. Glyco-related web-tools and data collections.

Name	Description	URL
Releated carbohydrate information in protein databases		
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/
Suga Base	^{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/
GlycoBase	Normalized HPLC data	www.bioch.ox.ac.uk/glycob/glycoimmunology/glycobase.htm
Graphical representations and nomenclature		
IUPAC	Nomenclature	www.chem.qmw.ac.uk/iupac/2carb/
LINUCS	Linear encoding of sugars	www.dkfz.de/spec/linucs/
LiGraph	Graphical representation	www.dkfz.de/spec/ligraph/
Consortium Functional Glycomics	Glycan Nomenclature	glycomics.scripps.edu/CFGnomenclature.pdf
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
Glydict	Ensemble of glycan conformations	www.dkfz.de/spec/glydict/
GlycoMaps DB	Conformation maps	www.dkfz.de/spec/glycomaps/
GlyProt	*In silico* glycosylation of Proteins	www.dkfz.de/spec/glyprot/php/main.php
Dynamic Molecules	Molecular dynamics of glycans	www.md-simulations.de
Name	**Organization**	**URL**
Carbohydrate databases		
CarbBank	Complex Carbohydrate Research Center, Athens	www.boc.chem.uu.nl/sugabase/carbbank.html
GlycociencesDB	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

the light of these new technological developments also new ways to collect scientific data have to be considered. Creating a grid of distributed local databases that interact on the basis of peer-to-peer communication will encourage people to input their recorded spectra into a local database and keep it private until it is published. Thus, the open architecture of the database will enable that mirrors can be easily installed at various locations. Inexpensive hardware platforms and the availability of free software tools will favor this process. Experiences from open source projects indicate that also the quality of the provided data will increase for obvious reasons. People who have recorded them will also input and maintain the data. The correction of faulty information will be quick, since nobody likes to bear the blame having made available bad data. Looking at the development of other open access data collections, it can be anticipated that free access to analytical data will encourage scientists to create their own applications based on the data provided. The availability of standard exchange format will also accelerate the exchange of information and knowledge. Publishers may encourage scientists to use these data collections to include the NMR- and MS-spectra, which otherwise may be lost for the scientific community.

4.2. Which Way to Go?

In principle, all four discussed data input models are eligible for collecting scientific data. However, the traditional way to extract data manually from published papers is far too expensive and error-prone. At present, the automatic text mining tools cannot replace the manual extraction of scientific data, but they are quite an efficient option to scan automatically many journals and to identify publications which are relevant for a specific subject.

The direct input of experimental data into a centralized database is a well-established procedure, which works quite efficiently in the genomic and proteomic area. Prerequisites for the success of this approach is that the database is located and maintained at a universally accepted institution, continuity is guaranteed and that access to the deposited data is free. Another important motivation to include scientific data is that the publishers force scientists to deposit their experimental data prior to publication. For the glycomics area, there seems to be currently no prestigious and widely accepted institution which would be able to take over this task. The Carbohydrate Database of the US Consortium for Functional Glycomics, which may be an appropriate organization, still seems to be at a very early stage and not prepared to take over this task soon.

The peer-to-peer network of distributed data collections is an attractive idea, which, based on a challenging informatics concept taking advantage of new prospects provided by internet technologies, has to be ascertained in the near future. Currently, there is only a very first prototype, developed in our group, for the input and dissemination of NMR spectral data connected by peer-to-peer communication which still has to be tested in large networks. Nevertheless, such an early stage bottom-up approach may be appropriate to evaluate the requirements of input options for glyco-related data collections under varying scientific questions and to optimize and standard exchange formats. It is also clear that for the wide acceptance of a peer-to-peer network there must also be one place where all data provided by the various peers are routinely stored and archived so that the no loss of data is guaranteed. Although this seems technically to be a minor problem which can be done on a routine basis, the question still remains how to find a well-accepted institution which is generally trusted by glycoscientists.

5. OUTLOOK

Unlike genome and proteome databases, development of carbohydrate databases has gained a broader attention only recently. The current situation in glycobiology is characterized by a nearly complete loss of all primary experimental data. Most of these data are either not published (MS-spectra, HPLC profiles, often NMR spectra) or are distributed in various journals using different notations, assignments, and data formats. With the *LINUCS*[28] and the *LinearCode*®,[25] two encoding schemes generating a unique description of complex structures are reported, but both have yet to find wider acceptance.

There is also no general agreement among publishers which information should be included in which notation in publications or as supplementary material. It can be anticipated that the definition of rules and standards for the input of primary data into the database will have a similar structuring effect in this field as has been demonstrated when establishing a world wide repository for spatial structures of proteins and DNA solved by x-ray crystallography or NMR techniques.

The existence of an accepted repository for glyco-related analytical data will guarantee that the loss of primary data will be considerably reduced. The agreement to quality standards and notations will not only raise the scientific usefulness of the stored data, but also it will open the opportunity to apply various data-mining approaches including multivariate statistics and artificial neural network algorithms to extract new information and to derive new knowledge which until now is hidden in unstructured data. Several authors pointed out that artificial neural network algorithms which are trained on a subset of spectra with known structures have a high potential to be used efficiently in the evaluation of similar structures. The ability of such algorithms to deduce a structure or suggest parts of a structure, which is recognizable in a given spectrum, is dependent on the spectra used in the training set.[59,60] The development of appropriate tools for glycosciences has suffered from the lack of available experimental NMR- and MS-spectra for complex oligosaccharides.

The calculation of the NMR chemical shifts for carbohydrates based on ab initio quantum chemical methods[61] seems to be still in its infancy and suffers from insufficient accuracy for known structures. However, with increasing power of computers, this tech-

nique could potentially be an alternative to the above-mentioned methods especially for structural determination of compounds not previously assigned. Thus, this approach may hold significant potential for the future.

The new technological possibilities to disseminate scientific and analytical data via the Internet should be developed under the open access paradigm for the needs of relevant glycoscience applications. However, the two large projects to collect glyco-related data are run by biotech 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 and underspins the imperative to start a new attempt to collect scientific data that will be freely available.

It should be the intention of such an initiative to create a repository for carbohydrate related data similar to the intensively used data collections for other biological macromolecules (*GenBank*,[62] *Swiss Prot*,[56] or *PDB*[63]). Looking at the great success and broad acceptance and use of these data collections, it is obvious that the availability of a similar facility for carbohydrate structures will have a large practical impact for the daily work of glycoscientists who want to correlate their findings with already known facts.

REFERENCES

1. 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* (1), 4–8.
2. Ben-Dor, S.; Esterman, N.; Rubin, E.; Sharon, N. Biases and complex patterns in the residues flanking protein N-glycosylation sites. Glycobiology **2004**, *14* (1), 95–101.
3. Rudd, P.M.; Colominas, C.; Royle, L.; Murphy, N.; Hart, E.; Merry, A.H.; Hebestreit, H.F.; Dwek, R.A. 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 **2001**, *1* (2), 285–294.
4. Dell, A.; Morris, H.R. Glycoprotein structure determination by mass spectrometry. Science **2001**, *291* (5512), 2351–2356.
5. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. Nature **2003**, *422* (6928), 198–207.
6. Lin, D.; Tabb, D.L.; Yates, J.R.R. Large-scale protein identification using mass spectrometry. Biochim. Biophys. Acta **2003**, *1646* (1-2), 1–10.
7. Harvey, D.J. Identification of protein-bound carbohydrates by mass spectrometry. Proteomics **2001**, *1* (2), 311–328.
8. Küster, B.; Krogh, T.N.; Mortz, E.; Harvey, D.J. Glycosylation analysis of gel-separated proteins. Proteomics **2001**, *1* (2), 350–361.
9. Sagi, D.; Peter-Katalinic, J.; Conradt, H.S.; Nimtz, M. Sequencing of tri- and tetra-antennary *N*-glycans containing sialic acid by negative mode ESI QTOF tandem MS. J. Am. Soc. Mass. Spectrom. **2002**, *13* (9), 1138–1148.

10. Geyer, H.; Schmitt, S.; Wuhrer, M.; Geyer, R. Structural analysis of glycoconjugates by on-target enzymatic digestion and MALDI-TOF-MS. Anal. Chem. **1999**, *71* (2), 476–482.
11. Duus, J.O.; Gotfredsen, C.H.; Bock, K. Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. Chem. Rev. **2000**, *100* (12), 4589–4614.
12. von der Lieth, C.W.; Siebert, H.C.; Kozar, T.; Burchert, M.; Frank, M.; Gilleron, M.; Kaltner, H.; Kayser, G.; Tajkhorshid, E.; Bovin, N.V.; Vliegenthart, J.F.; Gabius, H.J. Lectin ligands: new insights into their conformations and their dynamic behavior and the discovery of conformer selection by lectins. Acta Anat. (Basel) **1998**, *161* (1–4), 91–109.
13. Kogelberg, H.; Solis, D.; Jimenez-Barbero, J. New structural insights into carbohydrate-protein interactions from NMR spectroscopy. Curr. Opin. Struct. Biol. **2003**, *13* (5), 646–653.
14. Manzi, A.E.; Norgard-Sumnicht, K.; Argade, S.; Marth, J.D.; van Halbeek, H.; Varki, A. 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 **2000**, *10* (7), 669–689.
15. Bartolozzi, A.; Seeberger, P.H. New approaches to the chemical synthesis of bioactive oligosaccharides. Curr. Opin. Struct. Biol. **2001**, *11* (5), 587–592.
16. Seeberger, P.H. Automated carbohydrate synthesis to drive chemical glycomics. Chem. Commun. **2003**, *10*, 1115–1121.
17. Hewitt, M.C.; Snyder, D.A.; Seeberger, P.H. Rapid synthesis of a glycosylphosphatidylinositol-based malaria vaccine using automated solid-phase oligosaccharide synthesis. J. Am. Chem. Soc. **2002**, *124* (45), 13434–13446.
18. van Kuik, J.A.; Hard, K.; Vliegenthart, J.F. Databases of complex carbohydrates. Trends Biotechnol. **1992**, *10* (6), 182–185.
19. Rudd, P.M.; Mattu, T.S.; Zitzmann, N.; Mehta, A.; Colominas, C.; Hart, E.; Opdenakker, G.; Dwek, R.A. Glycoproteins: rapid sequencing technology for N-linked and GPI anchor glycans. Biotechnol. Genet. Eng. Rev. **1999**, *16*, 1–21.
20. Marchal, I.; Golfier, G.; Dugas, O.; Majed, M. Bioinformatics in glycobiology. Biochemie **2003**, *85* (1–2), 75–81.
21. Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. The complex carbohydrate structure database. Trends Biochem. Sci. **1898**, *14*, 475–477.
22. Doubet, S.; Albersheim, P. CarbBank. Glycobiology **1992**, *2* (6), 505.
23. Loss, A.; Bunsmann, P.; Bohne, A.; Schwarzer, E.; Lang, E.; von der Lieth, C.W. SWEET-DB: an attempt to create annotated data collections for carbohydrates. Nucl. Acids Res. **2002**, *30* (1), 405–408.
24. Cooper, C.A.; Joshi, H.J.; Harrison, M.J.; Wilkins, M.R.; Packer, N.H. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 Update. Nucl. Acids Res. **2003**, *31* (1), 511–513.
25. Banin, E.; Neuberger, Y.; Altshuler, Y.; Halevi, A.; Inbar, O.; Dotan, N.; Dukler, A. A novel linear code nomenclature for complex carbohydrates. Trends Glycosci. Glycotech. **2002**, *14*, 127–137.
26. Stenutz, R.; Jansson, P.E.; Widmalm, G. Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures. Carbohydr. Res. **1998**, *306* (1–2), 11–17.

27. McNaught, A.D. Nomenclature of carbohydrates (recommendations 1996). Adv. Carbohydr. Chem. Biochem. **1997**, *52*, 43–177.
28. Bohne-Lang, A.; Lang, E.; Forster, T.; von der Lieth, C.W. LINUCS: linear notation for unique description of carbohydrate sequences. Carbohydr. Res. **2001**, *336* (1), 1–11.
29. Cooper, C.A.; Harrison, M.J.; Webster, J.M.; Wilkins, M.R.; Packer, N.H. Data standardisation in GlycoSuiteDB. Pac. Symp. Biocomput. **2002**, 297–309.
30. Bohne, A.; Lang, E.; von der Lieth, C. W3-SWEET: carbohydrate modeling by internet. J. Mol. Model. **1998**, *4* (1), 33–43.
31. Bohne, A.; Lang, E.; von der Lieth, C.W. SWEET-WWW-based rapid 3D construction of oligo- and polysaccharides. Bioinformatics **1999**, *15* (9), 767–768.
32. Rudd, P.M.; Dwek, R.A. Rapid, sensitive sequencing of oligosaccharides from glycoproteins. Curr. Opin. Biotechnol. **1997**, *8* (4), 488–497.
33. Rudd, P.M.; Guile, G.R.; Kuster, B.; Harvey, D.J.; Opdenakker, G.; Dwek, R.A. Oligosaccharide sequencing technology. Nature **1997**, *388* (6638), 205–207.
34. Perkins, D.N.; Pappin, D.J.; Creasy, D.M.; Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **1999**, *20* (18), 3551–3567.
35. Appel, R.D.; Bairoch, A.; Hochstrasser, D.F. A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. Trends Biochem. Sci. **1994**, *19*, 258–260.
36. Cooper, C.A.; Gasteiger, E.; Packer, N.H. GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics **2001**, *1* (2), 340–349.
37. Gaucher, S.P.; Morrow, J.; Leary, J.A. STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. Anal. Chem. **2000**, *72* (11), 2331–2336.
38. Ethier, M.; Saba, J.A.; Ens, W.E.; Standing, K.G.; Perreault, H. Automated structural assignment of derivatized complex N-linked oligosaccharides from tandem mass spectra. Rapid Commun. Mass Spectrom. **2002**, *16* (18), 1743–1754.
39. Ethier, M.; Saba, J.A.; Spearman, M.; Krokhin, O.; Butler, M.; Ens, W.E.; Standing, K.G.; Perreault, H. 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. **2003**, *17* (4), 2713–2720.
40. Clerens, S.; Van den Ende, W.; Verhaert, P.; Geenen, L.; Arckens, L. Sweet substitute: a software tool for in silico fragmentation of peptide-linked *N*-glycans. Proteomics **2004**, *4* (3), 629–632.
41. Lohmann, K.K.; von der Lieth, C.W. GlycoFragment and GlycoSearch MS: web tools to support the interpretation of mass spectra of complex carbohydrates. Nucleic Acids Res. **2004**, July 1; 30 (web server issue): w 261–266.
42. Lohmann, K.K.; von der Lieth, C.-W. GLYCO-FRAGMENT: a web tool to support the interpretation of mass spectra of complex carbohydrates. Proteomics **2003**, *3* (10), 2028–2035.
43. Domon, B.; Costello, C.E. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconjugate **1988**, *5*, 397–409.

44. Guile, G.; Rudd, P.; Wing, D.; Prime, S.; Dwek, R. A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. Anal. Biochem. **1996**, *240* (5), 210–226.
45. Royle, L.; Mattu, T.S.; Hart, E.; Langridge, J.I.; Merry, A.H.; Murphy, N.; Harvey, D.J.; Dwek, R.A.; Rudd, P.M. An analytical and structural database provides a strategy for sequencing *O*-glycans from microgram quantities of glycoproteins. Anal. Biochem. **2002**, *304* (1), 70–90.
46. van Kuik, J.A.; Hard, K.; Vliegenthart, J.F. A ^{1}H NMR database computer program for the analysis of the primary structure of complex carbohydrates. Carbohydr. Res. **1992**, *235*, 53–68.
47. Jansson, P.E.; Kenne, L.; Widmalm, G. Computer-assisted structural analysis of oligosaccharides using CASPER. Anal. Biochem. **1991**, *199* (1), 11–17.
48. Jansson, P.E.; Widmalm, G. CASPER: a computer program used for structural analysis of carbohydrates. J. Chem. Inf. Comput. Sci. **1991**, *31* (4), 508–516.
49. Jansson, P.E.; Widmalm, G.; Stenutz, R. Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures. Carbohydr. Res. **1998**, *306* (1–2), 11–17.
50. Barth, A. SpecInfo: an integrated spectroscopic information system. J. Chem. Inf. Comput. Sci. **1993**, *33*, 52.
51. Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using neural networks for (13)C NMR chemical shift prediction–comparison with traditional methods. J. Magn. Reson. **2002**, *157* (2), 242–252.
52. Kalchhauser, H.; Robien, W. CSEARCH: a computer program for identification of organic compounds and fully automated assignment of carbon-13 nuclear magnetic resonance spectra. J. Chem. Inf. Comput. Sci. **1985**, *25* (2), 103–108.
53. Wheeler, D.L.; Church, D.M.; Federhen, S.; Lash, A.E.; Madden, T.L.; Pontius, J.U.; Schuler, G.; Schriml, L.M.; Sequeira, E.; Tatusova, T.A.; Wagner, L. Database resources of the National Center for Biotechnology. Nucl. Acids Res. **2003**, *31* (1), 28–33.
54. de Bruijn, B.; Martin, J. Getting to the (c)ore of knowledge: mining biomedical literature. Int. J. Med. Inf. **2002**, *67* (1–3), 7–18.
55. Nenadic, G.; Spasic, I.; Ananiadou, S. Terminology-driven mining of biomedical literature. Bioinformatics **2003**, *19* (8), 938–943.
56. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.C.; Estreicher, A.; Gasteiger, E.; Martin, M.J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. Nucl. Acids Res. **2003**, *31* (1), 365–370.
57. Lampen, P.; Davis, A.N. JCAMP-DX for NMR. Appl. Spectrosc. **1993**, *47* (8), 1093–1098.
58. Lampen, P.; Hillig, H.; Davis, A.N.; Linscheid, M. JCAMP-DX for mass spectrometry. Appl. Spectrosc. **1994**, *48* (12), 1445–1452.
59. Meyer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. Identification of the 1H-NMR spectra of complex oligosaccharides with artificial neural networks. Science **1991**, *251* (4993), 542–544.
60. Radomski, J.P.; van Halbeek, H.; Meyer, B. Neural network-based recognition of oligosaccharide ^{1}H-NMR spectra. Nat. Struct. Biol. **1994**, *1* (4), 217–218.

61. Helgaker, T.; Jaszunski, M.; Ruud, K. Ab initio methods for the calculation of NMR shielding and indirect spinminus sign spin coupling constants. Chem. Rev. **1999**, *99* (1), 293–352.
62. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Wheeler, D.L. Genbank. Nucl. Acids Res. **2003**, *31* (1), 23–27.
63. Westbrook, J.; Feng, Z.; Chen, L.; Yang, H.; Berman, H. The Protein Data Bank and structural genomics. Nucl. Acids Res. **2003**, *31* (1), 489–491.

Received November 29, 2003
Accepted March 17, 2004