Beilstein-Institut

Glyco-Bioinformatics – *Bits 'n' Bytes of Sugars*
October 4th – 8th, 2009, Potsdam, Germany

Bioinformatics – Key to the Future of Chemical Glycomics

Peter H. Seeberger

Max Planck Institute for Colloids and Surfaces, Potsdam, Germany,
Free University of Berlin, Arnimallee 22, 14195 Berlin, Germany, and
The Burnham Institute, La Jolla, CA, U.S.A.

E-Mail: peter.seeberger@mpikg.mpg.de

Received: 28th May 2010 / Published: 10th December 2010

Abstract

The glycome is more complex than either the genome or the proteome. Efforts to understand glycomics are producing information regarding the structure and function of carbohydrates. Branching and stereochemistry of the glycosidic linkage renders carbohydrates much more complex than oligonucleotides and proteins. Bioinformatics is a key technology to extract the information relayed via glycans. Three major classes of mammalian carbohydrates, glycolipids, O- and N-linked glycans, were analyzed based on the largest available database. The average oligosaccharide is composed of about eight monosaccharide units and while about a quarter of all oligosaccharides are strictly linear, the remainder are branched at least once. Glucosamine, galactose and mannose are dominating and comprise about 75% of the monosaccharides within mammalian oligosaccharide frameworks. α-linked sialic acid, α-linked fucose and β-linked galactose decorate the majority of reducing termini. Glucose as the most abundant carbohydrate in mammals plays only a very minor role within these structures. Particular emphasis was placed on analyzing the way the monosaccharide units are linked within the oligomeric framework. Just eleven monosaccharide connections account for more than 75% of all linkages. Thus, the number of structural combinations found in nature – the part of the occupied mammalian glycospace – is much smaller than expected. Only 36 monosaccharide building blocks are required to construct 75% of the 3299 mammalian oligosaccharides.

Introduction

Three major repeating biomacromolecules, polynucleotides, polypeptides and carbohydrates, are responsible for much of the information transfer in biological systems. Encoding and transmission of information relies on the construction of diverse macromolecules that contain the message. Polynucleotides serve as the blueprint of life in form of DNA; polypeptides carry out most of reactions in living cells. Both polymers are strictly linear and derived biosynthetically via reliable templated syntheses. DNA is composed of four nucleotides and mammalian proteins have 20 proteinogenic amino acids that determine polymer diversity: 4096 (4^6) hexanucleotides and 64 million (20^6) hexapeptides are possible. Posttranslational modifications such as phosphorylation, glycosylation, and lipidation further increase protein complexity.

The term ‘‘carbohydrates’’ describes a host of different bio-oligomers composed of monosaccharides. Oligosaccharides are almost always part of glycoconjugates, i.e. the combination of a sugar chain with a protein (glycoprotein), a lipid (glycolipid) or both lipid, and protein (glycosylphosphatidylinositol (GPI) anchored proteins) [1]. Carbohydrate chains can be branched, since each monosaccharide provides different positions around the ring that can be connected. In contrast to amide or phosphate diester linkages, the formation of each glycosidic linkage creates one new stereogenic centre. Carbohydrate complexity is increased by the stereocentres that constitute the ring in addition to ring size, linkage position, branching as well as further attachments such as sulfation, methylation, and phosphorylation.

Not surprisingly, carbohydrate complexity dwarves that of both DNA and proteins but to date has been assessed on a purely theoretical level [2]. We performed calculations regarding the diversity of mammalian carbohydrate structures, based only on the ‘‘ten mammalian monosaccharides’’ (Glc, Gal, Man, Sia, GlcNAc, GalNAc, Fuc, Xyl, GlcA, IdoA) not considering any further attachments. The number of structural combinations encountered in nature – the part of the glycospace that is actually occupied – has not yet been elucidated. A systematic structural analysis of mammalian oligosaccharide structures deposited in glycan databases will aid our understanding of carbohydrate diversity and help to identify a putative set of monosaccharide building blocks for efficient carbohydrate assembly.

Table 1. Diversity space of oligonucleotides, peptides and mammalian oligosaccharides. The numbers for the mammalian oligosaccharides are based on the ‘‘ten mammalian monosaccharides’’: D-Glc [4], D-Gal [4], D-Man [4], D-Sia [4], D-GlcNAc [3], D-GalNAc [3], L-Fuc [3], D-Xyl [3], D-GlcA [3], L-IdoA [3]. The number of substitutable OH groups (excluding the anomeric one) is given in square brackets. Commonly, only the pyranose ring forms, but not the furanose ring forms of the above mentioned monosaccharides are found in mammals [20].

Oligomer size	Numbers of different oligomers		
	Nucleotides	Peptides	Carbohydrates
1	4	20	20
2	16	400	1 360
3	64	8 000	126 080
4	256	160 000	13 495 040
5	1 024	3 200 000	1 569 745 920
6	4 096	64 000 000	192 780 943 360

Access to pure, structurally defined carbohydrates remains difficult at a time when the automated synthesis of oligonucleotides [3] and oligopeptides [4] is common. While biologically relevant oligosaccharides can be assembled from monosaccharides in a linear fashion on an automated synthesizer [5, 6], no general method for non-specialists to draw from a set of commercially available building blocks exists yet. The structural complexity of carbohydrates may complicate a comprehensive synthesis approach in case too many building blocks are needed.

A better understanding of the structures actually found in nature can guide the selection of building blocks needed for assembly. A database of reliable structures is required to analyze carbohydrate diversity. Relatively limited data sets exist, since carbohydrate isolation and structure elucidation are formidable challenges. The systematic collection of carbohydrates in databases is lagging far behind genomics and proteomics. Currently no database provides access to all published glycan structures although several commercial and publicly funded initiatives are working to make glycan structures available in a well-structured and annotated digital representation [7]. These databases contain mainly information about O- and N-linked glycans since their isolation and sequencing is more tractable than that of glycosaminoglycans for example. Reported here is a detailed statistical analysis of the GLYCOSCIENCES.de database [8] to elucidate the mammalian glycospace with a focus on the oligosaccharide portions of O- and N-glycans.

Glycan Database Analysis

The complexity analysis was based on 3299 oligosaccharides from 38 mammals (9). Non-carbohydrate portions such as amino or fatty acids at the reducing terminus were not considered. Most oligosaccharides (2128 of 3299, 64%) are of human origin, the rest are derived from cow, rat, pig, mouse and other species. Since the statistical analyses did not reveal any relevant difference between the human and the mammalian set of oligosaccharide structures we focused our analysis on the mammalian sugars.

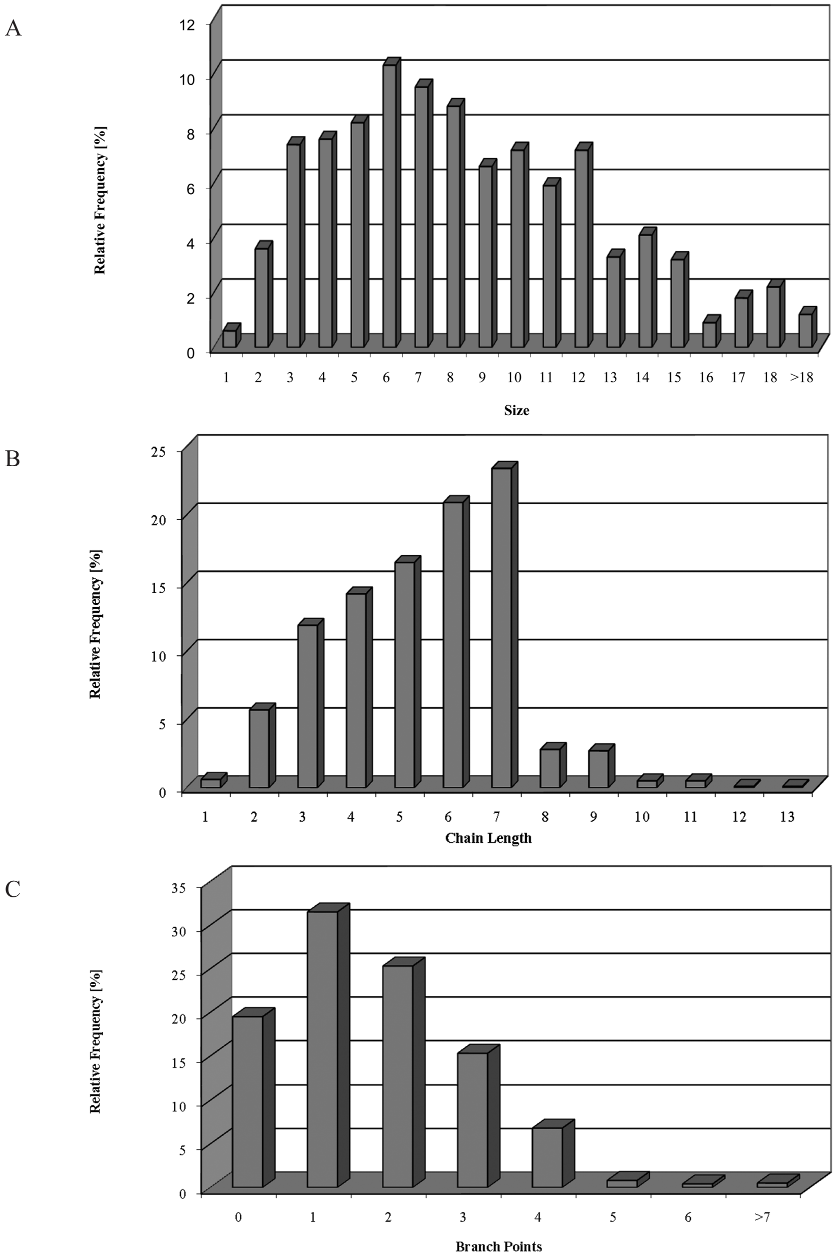

Figure 1. (A) Size of mammalian O- and N-glycans. The relative frequency is given in %. The largest oligosaccharide contains 37 monosaccharides; (B) Chain length of mammalian O- and N-glycans. The longest chain contains 13 monosaccharide moieties; (C) Branching complexity of mammalian O- and N-glycans.

Initially we addressed some basic questions: What is the most common size of an oligosaccharide? What is the typical chain length within a branched oligosaccharide? What portion of oligosaccharides is linear, what portion is branched? The size of an oligosaccharide is described as the number of monosaccharide units that make up the oligomer. The chain length is the longest path of monosaccharide units from the reducing end to the non-reducing terminus of the chain. The number of terminal residues was calculated as well. It differs by one from the number of branch points.

The average oligosaccharide is composed of about eight monosaccharide units whereas the sugars in the database vary in size from one (*e.g.* T_N antigen) to 37 monosaccharide units (Figure 1a). Most mammalian structures (about 95%) have a shorter chain length than eight residues (Figure 1b). The longest mammalian carbohydrate structure in the database has a maximum chain length of 13. About one fifth of the oligosaccharide structures in the database are linear. More than half of all oligosaccharides are branched once or twice, while 22% of the structures are branched three or four times. Few carbohydrates are branched five times or more than five times with a maximum of nine branch points (Figure 1c).

N-Acetyl-D-glucosamine (32%), D-galactose (25%) and D-mannose (19%) comprise more than 75% of all monosaccharide units found in mammalian oligosaccharides. Sialic acid and L-fucose are found less frequently (8% each). The most abundant monosaccharide in nature, D-glucose, that makes up cellulose, starch and glycogen, astonishingly plays only a minor role within mammalian O- and N-glycans. Three different monosaccharides dominate the non-reducing terminus, the site most often recognized by carbohydrate-binding proteins: D-galactose, sialic acid and L-fucose each cap about a quarter of oligosaccharides. D-Mannose and *N*-acetyl-D-glucosamine each terminate about 8% of mammalian oligosaccharides.

In planning a comprehensive, general and linear synthetic approach a set of building blocks containing the proper protective groups to install all possible connectivities and stereogenic centres is mandatory. Different protective groups mark hydroxyl groups that serve as nucleophiles during chain extension and those that remain latent. The protective groups used should also control the stereochemical outcome of glycosylation reactions. To further complicate matters, the sterics, electronics and conformation of the monosaccharides are fundamentally influenced by the choice of protective groups [10]. One aim of this study is to derive a minimal set of putative monosaccharide building blocks required to assemble the majority of mammalian oligosaccharides in a strictly linear fashion. Procurement of monosaccharide building blocks is a formidable challenge [11] and a defined number of reliable standard components for oligomer construction would help in synthetic planning and for practical reasons. These building blocks would be utilized in the build-up of linear and branched molecules by solution- and solid-phase methods.

Each of the ten mammalian monosaccharide units can in principle be connected to its neighbours in a variety of different ways including different anomeric configurations as well as branching once or more often. To construct all theoretically possible mammalian oligosaccharides by linear chemical synthesis 224 different building blocks would be required (for further details see SI). Due to this large number, special care was taken to elucidate the stereochemistry at the anomeric position (α or β) and the position of linkages within mammalian O- and N-glycans. The results illustrated in Figure 2 are stunning: 80% of the monosaccharide linkages within the oligomers can be constructed using only 13 building blocks. The most frequently occurring connections are 4-linked β-GlcNAc, capping α-Sia, capping α-Fuc, capping β-Gal, 2-linked α-Man and 3-linked β-Gal.

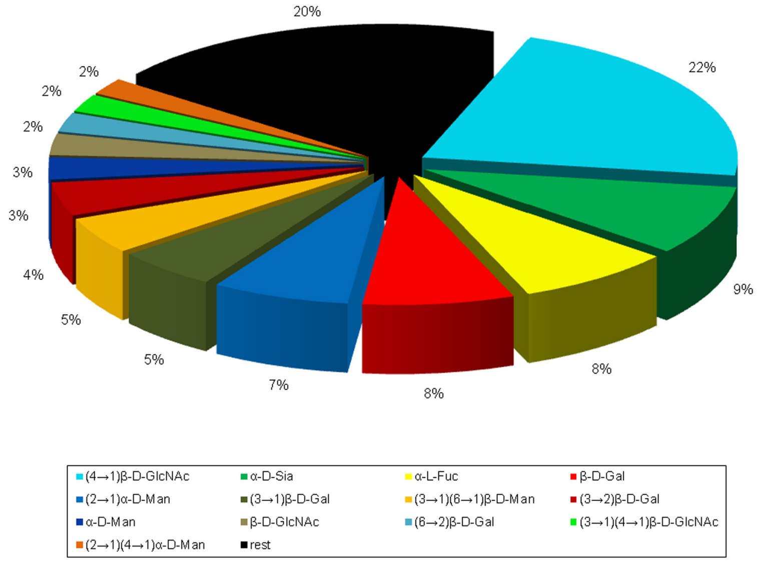

Figure 2. The 13 most abundant monosaccharide units (with linkage mode and position) found in mammalian O- and N-glycans.

Based on this analysis, key building blocks needed for the construction of mammalian oligosaccharides can be designed. Twenty putative building blocks **1 – 20** have been postulated to obtain the most abundant linkages. For this purpose, benzyl groups (Bn) were selected for the permanent protection of hydroxyls. Where participating groups are needed to ensure anomeric specificity, pivaloyl (Piv) [12], acetyl (Ac) or benzoyl (Bz) groups are placed for permanent protection. Fluorenylmethoxycarbonyl (Fmoc) was selected as temporary protecting group and also serves as a participating group in the C2 position for temporary protection [13]. To install branched carbohydrate structures two or more temporary protecting groups that can be cleaved chemoselectively are necessary. Levulinoyl ester

(Lev) and *p*-methoxybenzyl (PMB) were selected as other temporary protecting groups [14]. Similar protecting group schemes can be selected alternatively. However, the large majority of the building blocks presented in Figure 3 have been tested successfully for their utility in solution- and solid-phase oligosaccharide syntheses [15]. The sialic acid and β-mannose building blocks **2** and **7** represent a special challenge concerning selective glycosidic bond construction. So far, α-sialic acid and β-mannosidic bonds were constructed using disaccharide modules in place of monosaccharide units [16]. These challenges are currently being addressed.

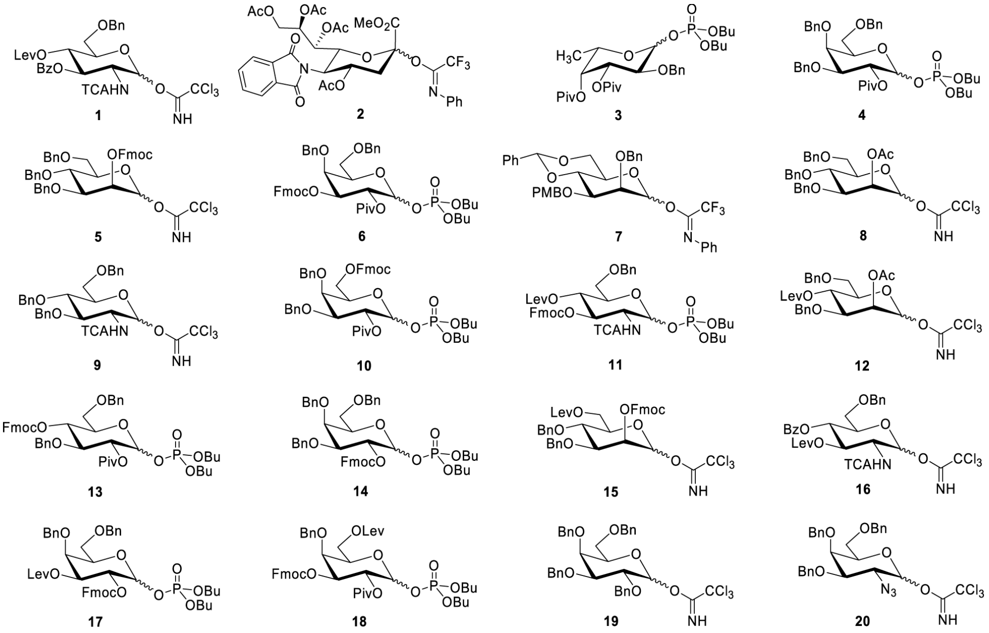

Figure 3. Putative monosaccharide building blocks **1 – 20** sorted by their relative abundance inmammalian oligosaccharides. Fmoc, Lev and PMB serve as temporary protecting groups, whereas Bn, Ac, Piv, Bz and TCA serve as permanent protecting groups.

With a set of putative building blocks in hand, the construction of mammalian carbohydrates contained in the database GLYCOSCIENCES.de was simulated. The results of these calculations are impressive: about 60% of the 3299 mammalian oligosaccharides are accessible with only 25 building blocks (Figure 4)! Just eleven more building blocks are needed to construct 75% of oligosaccharides. To produce 90% of all structures, a set of 65 building blocks is required. The number of building blocks to access the last 10% of mammalian oligosaccharides increases tremendously. The occurrence of rare monosaccharide units commonly not found in mammals such as D-fucose, L-arabinose, L-rhamnose and D-galacturonic acid as well as unusual linkages of L-fucoses and sialic acids are likely the result of erroneous assigned databank entries. Microorganisms that live in mammals express a much broader variety of carbohydrate moieties and linkages and may be the source of the additional sugars.

Evaluating the accessibility of the different classes reveals that for each class even fewer building blocks are required to reach a certain number of structures. Glycolipids in comparison show a greater variety of different linkages than N- and O-linked glycans (Figure 4).

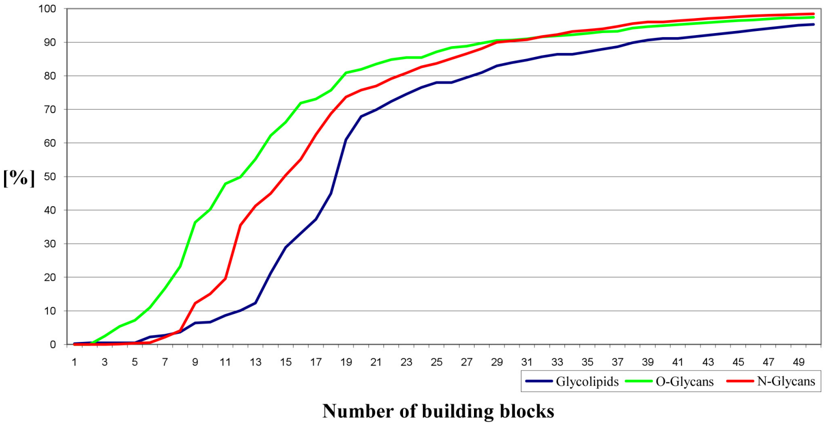

Figure 4. Percentage of accessible mammalian carbohydrates split into different classes (glycolipids, N- and O-linked glycans) and correlated to the number of building blocks.

A rather small number of building blocks is sufficient to access the majority of the mammalian glycospace. In many cases the reducing terminal units can be introduced by building blocks containing a temporary protecting group to further reduce this number. However, in the case of branched structures such an approach may be problematic. Therefore, our synthetic strategy is based on general principles with special capping building blocks.

Conclusions and Outlook

This review describes mammalian oligosaccharide diversity (“glycospace”) based on a glycan databank. Carbohydrate sizes, chain lengths and branching complexity were examined. Analysis of monosaccharide connectivities within the oligomeric structures guided us to identify a set of putative monosaccharide building blocks suitable for the linear solution- and solid-phase assembly of mammalian oligosaccharides. This potential building block set was correlated with the accessible 3299 mammalian carbohydrate structures in the GLYCOSCIENCES.de databank. Only 36 building blocks are needed to construct 75% of the 3299 mammalian oligosaccharides.

Acknowledgements

I am grateful to all present and past members of my research group as well as our collaborators at a variety of institutions. I thank the Max-Planck Society for generous support.

References

[1] Seeberger, P.H, Werz, D.B. (2005) Automated Synthesis of Oligosaccharides as a Basis for Drug Discovery. *Nat. Rev. Drug Discovery* **4**:751 – 763.
doi: http://dx.doi.org/10.1038/nrd1823.

[2] a) Schmidt, R.R. (1986) New methods of glycoside and oligosaccharide syntheses – are there alternatives to the Koenigs-Knorr method? *Angew. Chem. Int. Ed. Engl.* **25**:212 – 235.
doi: http://dx.doi.org/10.1002/anie.198602121.
b) 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. *Glycobiol*ogy **4**:759 – 767.
doi: http://dx.doi.org/10.1093/glycob/4.6.759.

[3] a) Caruthers, M.H. (1985) Gene synthesis machines: DNA chemistry and its uses, *Science* **230**:281 – 285.
doi: http://dx.doi.org/10.1126/science.3863253.
b) Caruthers, M.H. (1991) Chemical synthesis of DNA and DNA analogs. *Acc. Chem. Res.* **24**:278 – 284.
doi: http://dx.doi.org/10.1021/ar00009a005.

[4] Atherton, E., Sheppard, R.C. (1989) *Solid-Phase Peptide Synthesis: A Practical Approach*. Oxford Univ. Press, Oxford.

[5] Sears, P., Wong, C.-H. (2001) Toward automated synthesis of oligosaccharides and glycoproteins. *Science* **291**:2344 – 2350.
doi: http://dx.doi.org/10.1126/science.1058899.

[6] a) Plante, O.J., Palmacci, E.R., Seeberger, P.H. (2001) Automated solid-phase synthesis of oligosaccharides, *Science* **291**:1523 – 1527.
doi: http://dx.doi.org/10.1126/science.1057324.
b) Love, K.R., Seeberger, P.H. (2004) Automated solid-phase synthesis of protected tumor-associated antigen and blood group determinant oligosaccharides. *Angew. Chem. Int. Ed.* **43**:602 – 605.
doi: http://dx.doi.org/10.1002/anie.200352539.

c) Werz, D.B., Castagner, B., Seeberger, P.H. (2007) Automated synthesis of tumor-associated carbohydrate antigens Gb-3 and Globo-H: Incorporation of α-galactosidic linkages. *J. Am. Chem. Soc.* **129**:2770 – 2771.
doi: http://dx.doi.org/10.1021/ja069218x.

[7] von der Lieth, C.-W. (2004) An endorsement to create open access databases for analytical data of complex carbohydrates. *J. Carbohydr. Chem.* **23**:277 – 297.
doi: http://dx.doi.org/10.1081/CAR-200030093.

[8] a) Lutteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M., von der Lieth, C.-W. (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research, *Glycobiology* **16**:71R-81R.
doi: http://dx.doi.org/10.1093/glycob/cwj049.
b) Less complex oligosaccharides may dominate due to experimental difficulties in sequencing larger structures. This problem is common to all the glycan databases.

[9] About 47% of the 3299 structures are derived from N-linked glycoproteins, 19% of them from O-linked glycoproteins and 17% from glycolipids. The rest is not assigned.

[10] a) Mootoo, D.R., Konradsson, P., Udodong, U., Fraser-Reid, B. (1988) "Armed" and "disarmed" n-pentenyl glycosides in saccharide couplings leading to oligosaccharides. *J. Am. Chem. Soc.* **110**:5583 – 5584.
doi: http://dx.doi.org/10.1021/ja00224a060.
b) Ye, X.S., Wong, C.-H. (2000) Anomeric reactivity-based one-pot oligosaccharide synthesis: A rapid route to oligosaccharide libraries. *J. Org. Chem.* **65**:2410 – 2431.
doi: http://dx.doi.org/10.1021/jo991558w.
c) Orgueira, H.A., Bartolozzi, A., Schell, P., Seeberger, P.H. (2002) Conformational locking of the glycosyl acceptor for stereocontrol in the key step in the synthesis of heparin. *Angew. Chem. Int. Ed.* **41**:2128 – 2131.
doi: http://dx.doi.org/10.1002/1521-3773(20020617)41:12<2128::AID-ANIE2128>3.0.CO;2-V.

[11] Boons, G.J., Ed. (1998) *Carbohydrate Chemistry.* Blackie, London, UK.

[12] Kunz, H., Harreus, A. (1982) Glycosidsynthese mit 2,3,4,6-tetra-*O*-pivaloyl-α-D-glucopyranosylbromid. *Liebigs Ann.* 41 – 48.

[13] Roussel, F., Knerr, L., Grathwohl, M., Schmidt, R.R. (2000) *O*-glycosyl trichloroacetimidates bearing Fmoc as temporary hydroxy protecting group: A new access to solid-phase oligosaccharide synthesis. *Org. Lett.* **2**:3043 – 3046.
doi: http://dx.doi.org/10.1021/ol006081l.

[14] Koeners, H.J., Verhoeven, J., van Boom, J.H. (1980) Synthesis of oligosaccharides by using levulinic ester as an hydroxyl protecting group. *Tetrahedron Lett.* **21**:381 – 382.
doi: http://dx.doi.org/10.1016/S0040-4039(01)85479-4.

[15] The presented building blocks (BBs) were successfully used:
a) **1**, **16**: Love, K.R., Seeberger, P.H. (2005) Solution syntheses of protected type II Lewis blood group oligosaccharides: Study for automated synthesis. *J. Org. Chem.* **70**:3168 – 3177.
doi: http://dx.doi.org/10.1021/jo047723b.
b) **2**: Tanaka, K., Goi, T., Fukase, K. (2005) Highly efficient sialylation towards α(2 – 3)- and α(2 – 6)-Neu5Ac-Gal synthesis: Significant 'fixed dipole effect' of N-phthalyl group on alpha-selectivity. *Synlett* **19**:2958 – 2962.
doi: http://dx.doi.org/10.1055/s-2005-921889.
c) **3**, **6**, **11**, **13**, **14**: Ref. 6b.
d) **4**, **8**, **12**: Hewitt, M.C., Seeberger, P.H. (2001) Automated solid-phase synthesis of a branched Leishmania cap tetrasaccharide. *Org. Lett.* **3**:3699 – 3702.
doi: http://dx.doi.org/10.1021/ol016631v.
e) **5**: Wu, X., Grathwohl, M., Schmidt, R.R. (2002) Efficient solid-phase synthesis of a complex, branched N-glycan hexasaccharide: Use of a novel linker and temporary-protecting-group pattern. *Angew. Chem. Int. Ed.* **41**:4489 – 4493.
doi: http://dx.doi.org/10.1002/1521-3773(20021202)41:23<4489::AID-ANIE4489>3.0.CO;2-X.
f) **7**; **10**, **15**, **17**, **18**: Kröck, L., Oberli, M., Werz, D.B., Seeberger, P.H., *unpublished results*.
g) **9**: Blatter, G., Beau, J.M., Jacquinet, J.C. (1994) The use of 2-deoxy-2-trichloroacetamido-D-glucopyranose derivatives in syntheses of oligosaccharides. *Carbohydr. Res.* **260**:189 – 202.
doi: http://dx.doi.org/10.1016/0008-6215(94)84038-5.
h) **19**: Wang, C.C., Lee, J.C., Luo, S.Y., Fan, H.F., Pai, C.L., Yang, W.C., Lu, L.D., Hung, S.C. (2002) Synthesis of biologically potent α1→2-linked disaccharide derivatives via regioselective one-pot protection-glycosylation. *Angew. Chem. Int. Ed.* **41**:2360 – 2362.
doi: http://dx.doi.org/10.1002/1521-3773(20020703)41:13<2360::AID-ANIE2360>3.0.CO;2-R.
i) **20**: Belén Cid, M., Bonilla, J.B., Alfonso, F., Martín-Lomas, M. (2003) Synthesis of new hexosaminyl D- and L-chiro-inositols related to putative insulin mediators. *Eur. J. Org. Chem.* 3505 – 3515.

[16] Ratner, D.M., Swanson, E.R., Seeberger, P.H. (2003) Automated synthesis of a protected N-linked glycoprotein core pentasaccharide. *Org. Lett.* **5**:4717 – 4720.
doi: http://dx.doi.org/10.1021/ol035887t.

[17] Doubet, S., Albersheim, P. (1992) Carbbank. *Glycobiology* **2**:505.
doi: http://dx.doi.org/10.1093/glycob/2.6.505.

[18] Doubet, S., Bock, K., Smith, D., Darvill, A., Albersheim, P. (1989) The complex carbohydrate structure database. *Trends Biochem. Sci.* **14**:475 – 477.
doi: http://dx.doi.org/10.1016/0968-0004(89)90175-8.

[19] Bohne-Lang, A., Lang, E., Forster, T., von der Lieth, C.-W. (2001) LINUCS: Linear Notation for Unique Description of Carbohydrate Sequences. *Carbohydr. Res.* **336**:1 – 11.
doi: http://dx.doi.org/10.1016/S0008-6215(01)00230-0.

[20] a) de Lederkremer, R.M., Colli, W. (1995) Galactofuranose-containing glycoconjugates in trypanosomatids. *Glycobiology* **5**:547 – 552.
doi: http://dx.doi.org/10.1093/glycob/5.6.547.
b) Marlow, A.L., Kiessling, L.L. (2001) Improved Chemical Synthesis of UDP-Galactofuranose. *Org. Lett.* **3**:2517 – 2519.
doi: http://dx.doi.org/10.1021/ol016170d.