Carbohydrate Research 344 (2009) 322–330

Contents lists available at ScienceDirect

Carbohydrate Research

journal homepage: www.elsevier.com/locate/carres

SOACS index: an easy NMR-based query for glycan retrieval

Emmanuel Maes *, Fanny Bonachera, Gerard Strecker, Yann Guerardel

Unité de Glycobiologie Structurale et Fonctionnelle, UMR 8576 CNRS, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq cedex, France

ARTICLE INFO

Article history:
Received 14 October 2008
Received in revised form 5 November 2008
Accepted 5 November 2008
Available online 12 November 2008

Keywords:
Query
Database
O-Glycans
N-Glycans
^{1}H NMR

ABSTRACT

^{1}H NMR is now a standard method to determine de novo primary sequence of all sorts of glycans. These last 30 years, tens of thousands of oligosaccharide sequences have been elucidated by NMR spectroscopy in conjunction with other physico-chemical methods including mass spectrometry and gas chromatography. Most of these sequences are now compiled and available in several web databases recently unified in publicly available *GlycomeDB*, along with sets of experimental data. However, because the search for an exact sequence exclusively based on proton chemical shifts is sometimes delicate for NMR non-specialists, we worked out a new type of query, named SOACS, which allows the easy retrieval of existing sequences. This query is based on the readily distinguished ^{1}H chemical shifts from any ^{1}H NMR spectrum, and was designed to be usable to the widest scientist community.

1. Introduction

Liquid NMR spectroscopy is nowadays a standard method for primary structure determination of a large panel of organic and biological molecules, including glycoconjugates (O- and N-linked glycans, glycolipids, polysaccharides, etc.). NMR spectroscopy is the only single method that gives insight on all parameters of the primary structure of native glycans—with the notable exception of D/L isomery- including epimery, ring size and anomeric configuration of monosaccharides, sequence, and linkage position. Its versatility poised NMR as the method of choice for definitive structural deciphering of glycans, which permitted to elucidate the structure of thousands of glycans in the past 30 years.[1–3] Data collection is essentially based on the measurement and comparison of the chemical shifts of relevant nuclei ^{1}H, ^{13}C, and ^{31}P that will be influenced by direct electronic environment. Classical pulse programs such as homonuclear proton–proton COSY, TOCSY, and NOESY experiments give access to configuration of each monosaccharide through both proton–proton vicinal coupling and dipolar coupling constants. Then, their linkage patterns are identified on the basis of heteronuclear (H/X) pulse programs such as HMQC, HSQC, and HMBC experiments through the observation of direct, geminal, and vicinal heteronuclear coupling constants. Moreover, a combination of pulse program such as HMQC-TOCSY or ROESY-TOCSY can be performed in order to obtain more information in a single experiment. The relatively low sensitivity of the method coupled to limited amount of biological material is often the only limitation to the data collection. However, recent advances in data handling, coupled with availability of very high field magnets (up to 23.5 T) and cryogenic probes dramatically decreased the quantities of material required for high quality analyses down to micromoles.

Despite these technical improvements, the shear structural heterogeneity of glycans as well as the difficulties in interpreting NMR spectra hinders further development of NMR technology for non-specialized laboratories. Two mains approaches have been separately developed over the years to overcome this problem and render NMR spectroscopy accessible to a wider community of glycobiologists. The first one is based on the comparison of NMR signals (^{1}H or ^{13}C) chemical shifts measured in ppm to experimental data compiled in a database and permits to retrieve existing glycan structures from a list of chemical shifts. Such approaches have been successfully integrated to different web-based databanks such as SugaBase, BCSDB (Bacterial Carbohydrate Structure DataBase), and more recently to Glycosciences. de portal.[1–3] A second approach relies on the estimation of ^{1}H and ^{13}C NMR chemical shifts from a predicted glycan structure. Two main programs presently allow such estimation, GLYNEST and CASPER, which are available through the common interface of Glycome DB.[4,5]

Although the retrieval systems integrated in the different glycan-oriented databases are nowadays extremely discriminative and permit robust structure predictions from experimental data, their general use by non-specialized scientists is still limited, mainly because of upstream manual data collection. Indeed, automated data collection from ^{1}H NMR experiments has so far never been successfully implemented on complex glycans, which greatly limits the potential use of databank for high-throughput analyses. Despite the fact that chemical shifts are invariant for a given set of experimental conditions (pH, solvent, and temperature), the

* Corresponding author. Tel.: +33 320336347; fax: +33 320436555.
E-mail address: emmanuel.maes@univ-lille1.fr (E. Maes).

0008-6215/$ - see front matter
doi:10.1016/j.carres.2008.11.001

coupling of signals generated by the presence of multiple vicinal protons splits signals of all coupled protons and thus prevents the automatic determination of each proton chemical shift from experimental data. Furthermore, variability in the coupling constant values that depend on the dihedral angles between vicinal protons, the magnet field intensity, and eventually the local homogeneity of magnetic field renders the automatic determination of each signal chemical shift very uncertain. In conclusion, we believe that accurate determination of most chemical shifts of glycan-associated protons to efficiently use most databanks retrieving systems presently requires basic knowledge of spectra interpretation.

We propose here a very simple retrieval system of glycan structures based on the calculation of a single value from a reduced set of informative and easily identified signals from a 1D ^{1}H NMR spectrum. We named it SOACS index that stands for Sum Of Anomeric Chemical Shifts. This method that may be applied to O- and N-linked glycans was tested on several hundreds of compounds from our own glycan library and on data reported by Vliegenthart's group, and proved to be a very discriminative query to search for referenced molecules.[6–9] On this basis, SOACS index was implemented as a query on our home-based glycan data base, GLYCOBASE. Furthermore, it has the potential to be used in other databases in addition to existing queries and in support to the classical structural reporter groups developed by Vliegenthart and co-workers.[6–10]

2. Results and discussion

2.1. Basic concept

SOACS stands for sum of anomeric chemical shifts. The basic principle is to sum the chemical shifts (δ) of well-defined and easily identified ^{1}H NMR signals extracted from the spectrum of any given oligosaccharide, and then use the calculated number to identify it. A second index, called SOACS-ol, is a complementary index to use when the glycan is reduced, as often is the case for mucin-type O-linked glycans released by reductive β-elimination. Compared to SOACS, it integrates an additional parameter representative of the reduced terminal monosaccharide. Despite their extreme simplicities, both indexes exhibited a surprisingly high capacity to discriminate close isomers, and may be used as a query for database searching.

2.2. Origin of glycans

During the course of this study we validated the concept on both O-linked mucin-type glycans (OGs) and N-linked glycans (NGs). All used OGs had in common the presence of a *N*-acetyl galactosamine residue in terminal reducing position presumably attached to the protein backbone on a serine or threonine amino acid, and are members of the most common and widely distributed family of O-linked glycans in mammals. Outside this well-known family, a wide range of glycans may be associated to protein backbone through an O-linked bond involving different reducing monosaccharides (Man, Gal, Glc, GlcNAc, Fuc, and Xyl) and amino acids (Ser, Thr, OHLys, and OHPro). With the exceptions of α-*N*-acetylgalactosaminidases from *Diplococcus pneumonia* and from *Alcaligenes* sp, which hydrolyze Gal(β1,3)GalNAc-(α1-O)Ser/Thr O-glycosidic linkage, and *endo*-α-*N*-galactosaminidase S from *Streptomyces* sp OH-11242, which releases Gal(β1,3)GalNAc-(α1-O)Ser/Thr and Gal(β1→4)GlcNAc(β1→6)[Gal(β1→3)]GalNAc-(α1-O)Ser/Thr, most OGs can only be released from glycoproteins by chemical means, either as reduced oligosaccharides (reductive β-elimination) or as unreduced oligosaccharides (non-reductive β-elimination, mild hydrazinolysis).[11–16] However, considering the relatively low yields of the non-reducing methods of release, reductive β-elimination is still the method of choice for OGs release. So, all the OGs used in the present study have been purified as reduced oligosaccharides according to published procedure.[17] N-linked oligosaccharides are associated to protein backbone via β-N-glycosidic type bonds between GlcNAc residues and the amide groups of asparagins. They all have in common the presence of a $Man_3GlcNAc_2$ core structure, with the exception of the members of the so-called pauci-mannosylated family in which the mannose moiety may be reduced to two Man residues or a single Man residue.[18–20] They may be released in different forms from the protein backbone either by enzymatic digestion with *endo*-*N*-acetyl-β-D-glucosaminidase or peptide-*N*-glycanase, or by chemical methods such as reductive alkalinolysis or hydrazynolysis.[21–24] Peptide-*N*-glycanase and hydrazynolysis essentially release N-glycans as intact, reducing oligosaccharides, whereas *endo*-*N*-acetyl-β-D-glucosaminidase releases them as truncated oligosaccharide lacking the terminal GlcNAc residue and reductive alkalinolysis as reduced oligosaccharides. Alternatively, N-glycans may also be prepared as glycopeptide, attached to either a single Asn or an oligopeptide, by extensive protease digestion.

2.3. Data collection and calculation of SOACS

For OGs and NGs, ^{1}H spectral width is about 6 ppm independently of the NMR magnet field. Whatever be the glycan structure, three regions that contain distinct signals associated to oligosaccharide non-exchangeable protons are easily observed on a one-dimension spectrum (Fig. 1):

- Between 0 and 3 ppm; region that contains signals associated to methyl groups (associated to 6-deoxysugars from 1.1 to 1.3 ppm and to *N*-acetyl and *O*-acetyl groups from 2.0 to 2.1 ppm) and methylene groups ($H3_{ax}$ of sialic acids around 1.7 ppm and $H3_{eq}$ of sialic acids around 2.7 ppm).
- Between 3 and 4.4 ppm; bulk region that contains signal associated to most ring protons except anomeric protons.
- Between 4.4 and 5.6 ppm; region that essentially contains anomeric proton resonances. In some rare cases, other signals may also be observed in this region, such as H4 of 4-O-sulfated HexNAc, H5 and H6 protons of 1,6 anhydroHexNAc, and H5 of Lewis

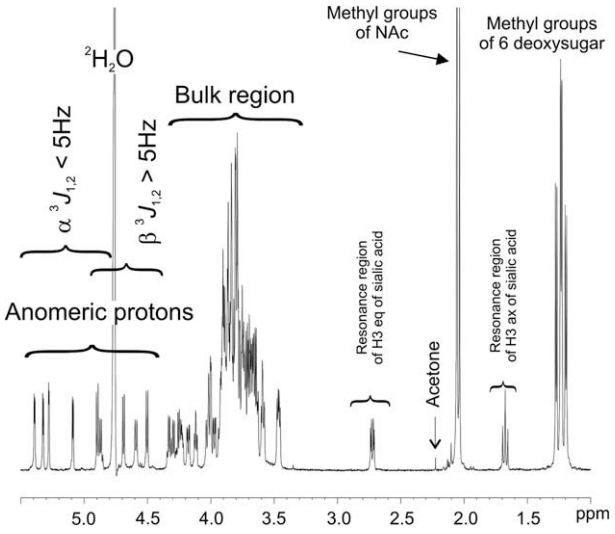

Figure 1. ^{1}H NMR spectrum of a O-glycan showing the different informative regions.

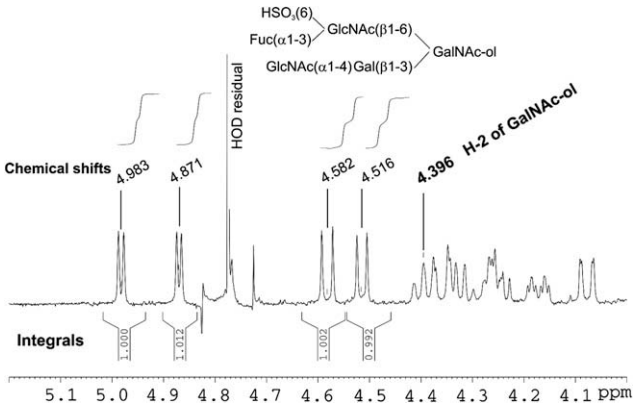

Figure 2. Details of sulfated OG-1 ^{1}H NMR spectrum showing signals used to calculate both SOACS (4.983 + 4.871 + 4.582 + 4.516 = 18.952) and SOACS-ol (18.952 + 4.395 = 23.347) indexes. Integration value of each anomeric signal (0.992–1.012) establishes that the analyzed molecule is pure.

type fucosyl residues.[25,9,10] None of these signals are doublets for they resonated as multiplet and thus cannot be confused with anomeric protons. Indeed, anomeric protons are in their vast majority observed as doublets because of the single vicinal coupling between H1 and H2 protons. Some notable exceptions are oligosaccharide-1-phosphate molecules in which the anomeric signal of the monosaccharide in reducing position is further split to a doublet of doublets (dd) because of an additional vicinal coupling of H1 with phosphorus atom. Moreover, H1 of 2-deoxyoses appear as dd since they are twice coupled through vicinal coupling constants ($^{3}J_{H1,H2ax}$ and $^{3}J_{H1,H2eq}$). However, both oligosaccharide-1-phosphate and 2-deoxyoses have never been observed in N- and O-glycans so far. Based on the anomeric configuration of hexopyranose residues, this region is typically divided into two sub-regions: one ranging from 4.4 to 5.0 ppm containing most β-anomeric protons with a large vicinal coupling constant ($^{3}J_{1,2} > 5$ Hz) and the other ranging from 5.0 to 5.6 ppm containing most α-anomer protons with a small constant constant ($^{3}J_{1,2} < 5$ Hz).[26]

SOACS of a given oligosaccharide is calculated by summing up chemical shifts values of the most easily recognized protons of the molecule. They include all anomeric protons between 4.4 and 5.6 ppm plus eventual $H3_{ax}$ (1.60–1.95 ppm) of sialic acid residues

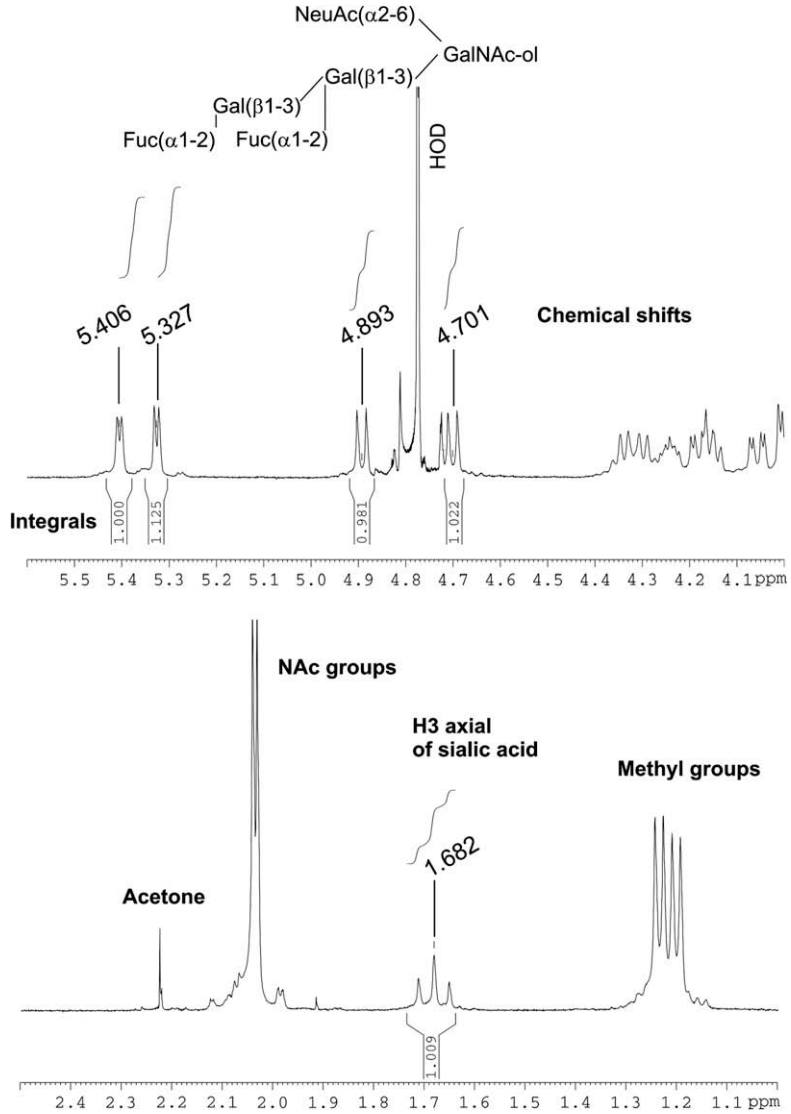

Figure 3. Details of sialylated OG-2 ^{1}H NMR spectrum showing informative data used to calculate SOACS index (5.406 + 5.327 + 4.893 + 4.701 + 1.682 = 22.009). Compared to OG-1, the easily identified H3ax value (1.682 ppm) is computed into the OG2 SOACS. Protons associated to methyl, methylene, and *N*-acetyl groups from Fuc, Neu5Ac, and HexNAc residues, respectively, are also identified in the same region but not computed.

for sialylated glycans. $H3_{ax}$ resonates between methyl groups and acetyl groups as a triplet instead of a doublet of doublets (H3eq) because of two identical values of geminal and vicinal coupling constants. Although, well separated from the other signals, we did not include in the calculation the methyl signals from the deoxysugars and from the *N*-acetyl groups because they, respectively, resonate in very narrow ranges and might be difficult to differentiate. For reduced OGs, an additional index, the SOACS-ol, may be calculated by adding to the SOACS the value of the GalNAc-ol H2. This signal is always observed as a pseudo-quadruplet with a value between 4.25 and 4.4 ppm, and is easily distinguished on a ^{1}H NMR spectrum. Its value is dependent on the substitution status of the GalNAc-ol, and thus very specific of the core type. The choice of SOACS or SOACS-ol depends on information which could be obtained from NMR spectra, SOACS-ol permitting to eventually discriminate two molecules with close SOACS values.

2.4. Calculation of SOACS and SOACS-ol

One dimension ^{1}H NMR experiment of any given purified OG provides all the data needed for SOACS index calculation. As shown on the 1D ^{1}H NMR spectrum of a sulfated reduced pentasaccharide (OG-1, Fig. 2), four isomeric signals at 4.516, 4.582, 4.871, and 4.983 ppm are easily recognized and computed to obtain a SOACS index of 18.952. The identical values of integration for each signal establish that they individually originate from single protons and are not the result of overlapping sub-signals. Similarly, four isomeric signals at 4.701, 4.893, 5.327, and 5.406 ppm are observed on the 1D ^{1}H NMR spectrum of OG-2 (Fig. 3). Furthermore, a clear $H3_{ax}$ signal is observed at 1.682 ppm, which permits to calculate a SOACS index of 22.009 for this sialylated reduced hexasaccharide. Two methyl groups associated to deoxyhexose residues and two associated to *N*-acetyl groups are also easily distinguished but not computed.

In some cases, signal overlaps prevent clear assignment of isomeric signals and calculation of SOACS index from 1D ^{1}H NMR analysis, as exemplified with the analysis of OG-3 that contains five β-anomers and one α-anomer (Fig. 4). Indeed, whereas three clear signals at 4.469, 4.681, and 5.382 ppm can be assigned to individual anomeric protons, a cluster of signals around 4.60 ppm cannot be differentiated from the one dimension spectrum. The relative integration of this signals cluster establishes that it contains several protons that may only be individually observed through $^3J_{H1-H2}$ on a COSY or a TOCSY 2D ^{1}H NMR spectrum (Fig. 4). Chemical shifts may be directly determined from 2D spectra, but their lower spectral resolution only permits a two-decimal measurement for a final SOACS value of 28.33. So it is preferable to use 1D ^{1}H NMR spectrum for value reporting after having identified the signals from the 2D ^{1}H NMR spectrum, which permits to calculate a three-decimal index of 28.339 (Fig. 4). Furthermore, a 2D experiment is often required to identify the GalNAc-ol H2 signal and use SOACS-ol. So, the combined use of 1D and 2D spectra permits to calculate SOACS and SOACS-ol indexes.

In order to evaluate their efficiency to discriminate close structural relatives, SOACS and SOACS-ol indexes were calculated from 165 OGs, from di- to deca-reduced saccharides, whose NMR parameters were compiled by Kamerling and Vliegenthart (Table 1).[9] Their individual structures have been summarized in Table S1. Similarly, SOACS and SOACS-ol indexes from amphibian OGs have been compiled in GB[6], and are tabulated in Table 2. As expected, values of SOACS and SOACS-ol indexes increase with increasing number of constituent monosaccharides (Fig. 5a and Fig. S1). Within each family characterized by the same number of monosaccharides, SOACS indexes values may vary by 3.403 units for reduced disaccharides

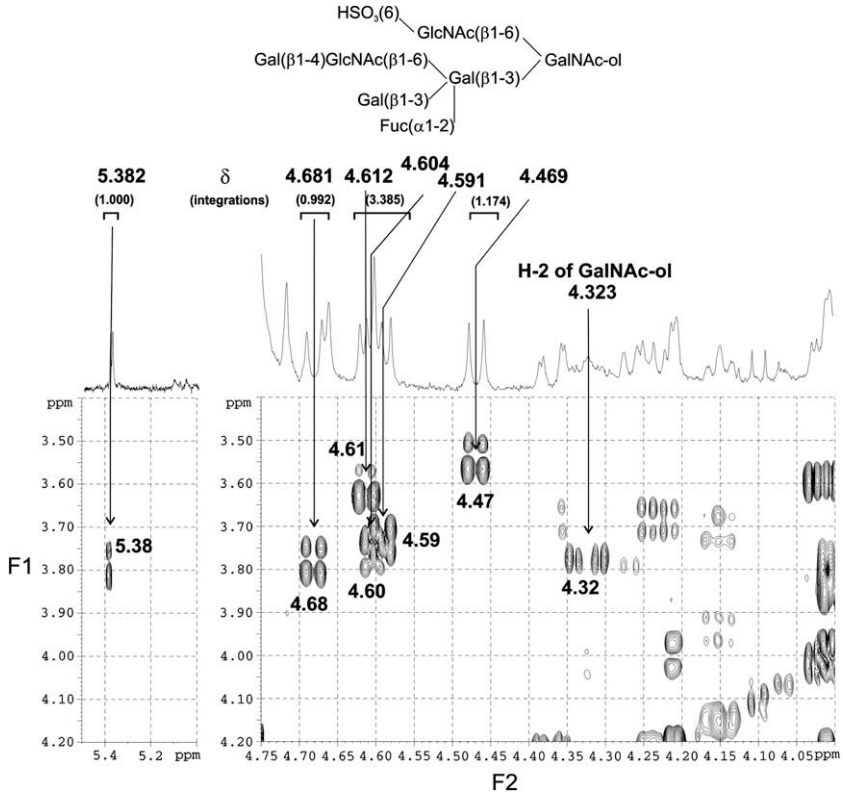

Figure 4. Details of 2D ^{1}H NMR COSY spectrum of OG-3 showing H1–H2 correlations. Two-dimension spectrum permits to resolve a cluster of otherwise overlapped anomeric signals at δ 4.59, 4.60, and 4.61 and to identify the H2 of *N*-acetylgalactosaminitol residue at δ 4.32. Calculation of SOACS directly from the COSY spectrum only permits a two-decimal precision because of lower spectral resolution compared to 1D ^{1}H NMR (5.38 + 4.68 + 4.461 + 4.60 + 4.59 + 4.47 = 28.33), whereas its calculation from 1D spectrum after signal identification permits to calculate a three-decimal SOACS of 28.339.

Table 1
Classification of SOACS indexes according to their increasing SOACS values and their corresponding SOACS-ol index

Reference number	SOACS	SOACS-ol	Reference number	SOACS	SOACS-ol	Reference number	SOACS	SOACS-ol
1	No	4.252	23	18.132	22.419	108	25.745	30.132
2	4.478	8.873	20	18.195	22.437	125	25.758	30.142
5	4.553	8.795	76	18.398	22.79	100	25.775	30.174
3	4.604	8.891	167	18.405	22.806	126	27.466	31.849
4	5.103	9.498	41	18.580	22.973	137	27.694	32.082
9	6.169	10.544	49	18.738	23.018	45	27.714	32.109
10	6.305	10.565	30	18.757	23.154	46	27.759	32.156
78	6.347	10.737	77	18.760	23.15	136	27.903	32.302
78A	6.364	10.753	75	18.824	23.215	54	28.507	32.777
11	6.790	11.183	39	18.833	23.238	58	28.814	33.074
11A	6.807	11.207	40	18.843	23.241	53	28.822	33.102
85	8.033	12.411	48	18.99	23.272	127	29.180	33.569
7	9.006	13.401	36	19.507	23.773	44	29.226	33.627
165	9.007	13.394	72	19.746	24.026	70	29.474	33.876
19	9.048	13.291	67	19.830	24.139	139	29.775	34.159
17	9.086	13.376	110	19.977	24.362	121	30.207	34.584
18	9.115	13.404	114	19.981	24.274	103	31.000	35.294
16	9.141	13.541	116	19.989	24.275	104	31.091	35.415
8	9.143	13.423	117	20.004	24.282	143	31.697	36.084
60	9.394	13.798	101	20.194	24.586	142	31.712	36.099
29	9.840	14.239	88	20.436	24.815	146	32.541	36.825
90	10.796	15.056	107	20.443	24.830	47	32.632	37.024
94	10.812	15.073	119	20.506	24.886	151	32.669	36.888
166	10.871	15.258	113	20.578	24.856	162	32.884	37.277
79	11.225	15.605	82	20.594	24.989	164	32.960	37.356
83	11.551	15.932	122	20.598	24.974	147	33.035	37.317
83A	11.571	15.957	99	20.711	25.111	157	33.148	36.532
96	12.666	16.928	106	20.714	25.102	154	33.158	37.437
86	12.915	17.725	93	21.216	25.466	74	33.214	37.614
25	13.495	17.889	123	22.303	26.687	159	33.318	37.705
133	13.583	17.968	141	22.565	26.954	32	33.442	37.828
13	13.615	18.016	27	22.634	27.040	145	33.594	37.982
12	13.631	18.027	26	22.636	27.033	59	33.853	38.143
21	13.636	17.918	28	22.658	27.050	56	33.865	38.135
24	13.636	18.039	15	22.695	27.084	55	33.877	38.142
61	13.932	18.337	14	22.790	27.191	71	34.952	39.222
35	14.213	18.485	62	23.369	27.774	144	36.968	41.357
38	14.344	18.745	57	23.464	27.728	149	37.522	41.804
34	14.420	18.724	51	23.466	27.749	65	37.523	41.915
37	14.446	18.713	52	23.466	27.736	64	37.547	41.939
66	15.288	19.592	168	23.727	28.135	152	37.738	41.997
87	15.320	19.700	63	23.775	28.181	153	37.743	42.022
80	15.354	19.752	33	23.872	28.272	160	37.840	42.235
98	15.359	19.751	43	23.922	28.323	150	38.235	42.636
105	15.364	19.754	124	24.021	28.405	148	38.265	42.561
134	15.446	19.830	50	24.026	28.308	155	38.463	42.742
112	15.503	19.782	31	24.063	28.459	158	38.509	42.908
81	15.516	19.910	69	24.112	28.504	163	38.842	43.243
92	15.919	20.166	42	24.189	28.589	161	38.912	43.182
135	16.043	20.293	68	24.237	28.629	156	43.666	48.068
89	16.057	20.332	138	24.517	28.902	73	43.827	48.229
95	16.129	20.366	140	24.530	28.916			
91	16.159	20.422	111	25.053	29.440			
84	17.009	21.300	115	25.069	29.350			
109	17.224	21.611	120	25.132	29.508			
97	17.766	22.015	118	25.334	29.590			
22	18.119	22.402	102	25.550	29.949			

The reference numbers correspond to Kamerling and Vliegenthart. [9] Structures are summarized in Table S1.

up to 16.200 units for reduced undecasaccharides (Fig. 5a). Survey of the structural heterogeneity of glycans within each family showed that the SOACS index value is dependent on the α-linked MS/β-linked MS/Sialic acid (α/β/S) ratio present in individual glycans; lower values being associated to high sialic acid contents and higher values being associated to a high content in α-linked MSs (Figs. 5c, d, 6a and b). This is easily rationalized by the increasing contribution to SOACS index of chemical shifts associated to sialic acid $H3_{ax}$, β-anomers and then α-anomers. As a result, molecules with identical α/β/S ratios exhibit reduced SOACS index dispersion within each family of glycans. For example, several rather homogeneous sub-families are easily distinguished in each family according to their α/β/S ratios: four in reduced tri-saccharides (α/β/S 0:1:1, 1:0:1, 0:2:0, and 1:1:0) (Fig. 5c), six in reduced tetra-saccharides (α/β/S 0:1:2, 0:2:1, 1:1:1, 0:3:0, 1:2:0, and 2:1:0) (Fig. 5d), and seven in reduced penta-saccharides (α/β/S 0:2:2, 0:3:1, 1:2:1, 2:1:1, 0:4:0, 1:3:0, 2:2:0) (Fig. 6a, b). Then, within sub-families, individual glycans exhibit SOACS index values different enough to be easily distinguished, even in the case of very close isomers (Fig 6c). Over the tested molecules, about 74% exhibit SOACS indexes with values superior by 0.02 units to their closest neighbors and 87% superior by 0.01 units, which permit to confidently use 0.002 ppm or 0.001 ppm error margin on chemical shifts ($\Delta\delta$) per measured signal on up to a decasaccharide without impairing the specificity of

Table 2
Classification of O-glycans from various amphibian species according to their SOACS values and their corresponding SOACS-ol

Nomenclature from literature	**SOACS**	SOACS-ol	Nomenclature from literature	**SOACS**	SOACS-ol	Nomenclature from literature	**SOACS**	SOACS-ol
rt NIII-1a[17]	**0,000**	4,251	xt 5C[35]	**15,289**	19,590	rr N-II-8d[30]	**24,621**	28,988
rt 100-4-a[17]	**1,660**	5,903	rtb 14[29]	**15,595**	23,933	rr 200-I-7[30]	**24,663**	29,017
rt NIII-1b[17]	**4,478**	8,873	rd A-19A[25]	**15,784**	20,176	rt 50-7a[17]	**24,770**	29,114
ra FNII-2[33]	**4,478**	8,873	xt 12A[35]	**15,846**	20,099	rd A15-B[25]	**24,892**	29,223
rt 200-3[17]	**4,592**	8,988	bv 11[36]	**16,378**	20,704	xt 9B[35]	**24,967**	29,270
rt NIII-2[17]	**5,170**	9,543	rr 100-5[30]	**16,381**	20,763	rt 100-6[17]	**25,156**	29,476
rt 50-5-a[17]	**6,124**	10,501	rd A-18[25]	**16,459**	20,828	rr N-II-10c[30]	**25,794**	30,096
xt 10B[35]	**6,165**	10,544	xt 9A[35]	**18,511**	22,902	bv 2[36]	**25,816**	30,197
rt 1[28]	**6,242**	10,625	rc C3[34]	**18,570**	22,969	bv 13[36]	**26,582**	30,818
bv 10[36]	**6,615**	10,544	rc C2[34]	**18,737**	23,121	rc C10-1[34]	**27,727**	32,339
rt 50-4[17]	**6,811**	11,176	rtb 13(4)[29]	**18,879**	23,277	rc C8-1[34]	**27,825**	32,212
ra FNII-4[33]	**9,001**	13,395	rd A-8[25]	**18,952**	23,347	rc C6[34]	**28,242**	32,628
rc D1[34]	**9,001**	13,395	ra 100-G[33]	**18,963**	23,355	ra 100-M[33]	**28,339**	32,667
xt 5A[35]	**9,002**	13,395	rtb 12[29]	**18,994**	23,393	bv 3[36]	**28,626**	33,027
rt 300-5[17]	**9,022**	13,411	rd A-4B[25]	**19,169**	23,564	rd A-19B[25]	**28,827**	33,160
rd A-5[25]	**9,022**	13,411	ra 100-H[33]	**19,244**	23,615	rr 100-12[30]	**28,879**	33,234
rtb 6[29]	**9,094**	13,483	rd A12-B[25]	**19,276**	23,645	rr 400-II-4[30]	**28,895**	33,246
ra 50-5[33]	**9,201**	13,596	ra 100-F[33]	**19,276**	23,645	rr 400-II-3[30]	**28,976**	33,340
rt 200-7[17]	**9,203**	13,592	rt 4[28]	**19,286**	23,675	bv 4[36]	**29,231**	33,231
rd N-4[25]	**9,413**	13,821	ra 100-D[33]	**19,320**	23,670	rr 100-11[30]	**29,288**	33,651
rp N-2[32]	**9,413**	13,821	ra 50-10[33]	**19,374**	23,728	bv 5[36]	**29,429**	33,662
rd N-3[29]	**9,820**	14,210	rr 100-7[30]	**19,381**	23,741	ra 100-L[33]	**29,508**	33,732
ra FNII-3[33]	**9,820**	14,210	rd A-7C[25]	**19,393**	23,785	rt 7[28]	**29,592**	33,924
rtb 4[29]	**10,429**	14,854	rr 200-II-6[30]	**19,393**	23,785	rd A-20[25]	**29,778**	34,050
rp 80-4a[32]	**11,073**	15,453	rd N-8A[25]	**19,650**	23,989	rtb 15(6)[29]	**30,522**	34,908
rd A-12A[25,31]	**11,073**	15,453	rt 100-4b[17]	**19,673**	24,032	rc C10-4[34]	**32,816**	37,264
rd 3[31]	**11,104**	15,499	xt 8[35]	**19,809**	24,099	rr 100-13[30]	**33,811**	38,176
rp 80-5[32]	**11,104**	15,499	xt 10A[35]	**19,863**	24,160	rr 200-I-10[30]	**33,871**	38,236
bv 8[36]	**11,223**	15,603	rc D3[34]	**19,956**	24,346	rr 400-II-5[30]	**33,888**	38,232
rr 100-2[30]	**11,249**	15,627	rtb 10[29]	**20,010**	24,388	rr 400-II-6[30]	**33,896**	38,257
bv 9[36]	**11,535**	15,912	rr N-II-6d[30]	**20,087**	24,450	rc C10-2[34]	**33,954**	38,488
rt 50-5-b[17]	**12,105**	16,256	rd N8-B[25]	**20,338**	24,624	rc C10-3A[34]	**33,979**	38,369
rtb 13(1)[29]	**13,539**	17,929	xt 7[35]	**20,431**	24,731	ra 50-14A[33]	**33,986**	38,198
ra 100-E[33]	**13,550**	17,945	rd 4[31]	**22.009**	26,335	rc C10-3B[34]	**33,987**	38,377
rd A-4A[25]	**13,819**	18,214	bv 12[36]	**22.009**	26.335	rd A-21[25]	**34,050**	39,039
rt 3[28]	**13,897**	18,284	xt 12B[35]	**22,126**	26,422	rt 300-11[17]	**34,326**	38,668
rd A-7A[25]	**13,941**	18,331	rc C7[34]	**23,161**	27,622	bv 6[36]	**34,448**	38,676
rc C1(D2)[34]	**13,993**	18,386	rc C5-3[34]	**23,238**	27,623	rc C8-2[34]	**34,924**	38,699
rd A-7B[25]	**14,015**	18,407	rc C5-1[34]	**23,298**	27,680	bv 14[36]	**36,184**	40,414
rr 200-I-2[30]	**14,015**	18,407	rc C5-4[34]	**23,476**	27,859	rr 400-II-7[30]	**38,488**	42,839
rd A-6A[25]	**14,069**	18,479	ra 100-J[33]	**23,833**	28,156	rr 200-I-8[30]	**38,904**	43,255
xt 6[35]	**14,145**	18,423	bv 1[36]	**23,856**	28,237	bv 7[36]	**38,976**	43,199
rtb 8[29]	**14,344**	18,747	rr 100-10[30]	**23,909**	28,254	rc C11-4[34]	**39,188**	43,577
ra FNII-5[33]	**14,344**	18,747	ra 50-14B[33]	**23,909**	28,254	bv 15[36]	**41,390**	45,613
xt 5B[35]	**14,345**	18,749	rd A-12C[25]	**24,018**	28,363	rr 200-I-9[30]	**43,252**	47,591
rd A-6B[25]	**14,395**	18,791	rt 5[28]	**24,130**	28,518	rc C12[34]	**44,434**	48,821
rt 200-9[17]	**14,397**	18,790	rd A-15A[25]	**24,223**	28,554	rc C14[34]	**49,237**	53,775
rtb 7[29]	**14,485**	18,878	rt 100-8[17]	**24,233**	28,568	rc C15-6[34]	**49,375**	53,959
rt 200-8[17]	**14,594**	18,981	rr 100-9[30]	**24,287**	28,650	rc C16[34]	**53,744**	58,132
rtb 9[29]	**14,672**	19,009	rt 6[28]	**24,421**	28,753	rc C17-3[34]	**58,143**	62,529
ra FNII-6[33]	**14,672**	19,009	rt 2[28]	**24,435**	28,783	rc B1(C18)[34]	**68,452**	72,840
ra 100-A[33]	**14,760**	19,122	rc D5(C2)[34]	**24,493**	28,883	rc B2(C19)[34]	**68,455**	72,841
rd A-3[30]	**14,762**	19,119	rt 50-7b[17]	**24,557**	28,954			

These values are reported in Glycobase. rt: *Rana temporaria*, ra: *Rana arvalis*, xt: *Xenopus tropicalis*, bv: *Bufo viridis*, rc: *Rana clamitans*, rr: *Rana ridibunda*, rd: *Rana dalmatina*.

the index (Fig. 5b). From these molecules, isomers that differ by the presence of either H2 epitopes [Fuc(α1,2)Gal(β1,4)GlcNAc] (compounds **40**, **42**, **48**, and **106**) or Le^x epitopes [Gal(β1,4) Fuc(α1,3)GlcNAc] (compounds **41**, **43**, **49**, and **107**) are easily distinguished according to their SOACs indexes that differ by 0.252–0.271 units (Table 1, Table S1). Also, O-glycan isomers that exclusively differ by the linkage position of one of their monosaccharides (e.g., **12**–**13**, **17**–**18**, **138**–**140**, and **142**–**143**) typically exhibit SOACS indexes differing by 0.013–0.029 units, which is discriminative enough for using a $\Delta\delta$ of 0.001 to 0.002 ppm without impairing the specificity of the index. A notable exception was observed for **26**–**27** and **152**–**153** isomers that are characterized by the presence of either terminal type-1 or type-2 LacNAc motifs and only differ by 0.002 and 0.005 units, respectively. Thus, SOACS indexes closeness prevents the use of any experimental error margin. In such cases, additional use of SOACS-ol may help to discriminate between two closely related structures such as compounds **152** and **153** for which SOACS-ol index appears as much more discriminative than SOACS index with a value difference of 0.025 units. However, in the case of **26** and **27**, SOACS-ol does not bring any noticeable improvement. Finally, compounds **51** and **52** (SOACS indexes = 23.466, Table S1 (5/5) noted 51* and 52*), two very close isomers, possess the same SOACS index. These two compounds only differ by the position of the O-2 linked Fuc residue either on the upper 6-linked (**51**) or on lower 3-linked branch (**52**) of the oligosaccharide-alditols. However, they are nonetheless easily distinguished by their respective SOACS-ol indexes that differ by 0.013 units.

A similar approach may also be applied to the assignment of N-linked glycans. Calculation of SOACS index for N-Glycans is essentially the same than for O-glycans. However, the core 3,

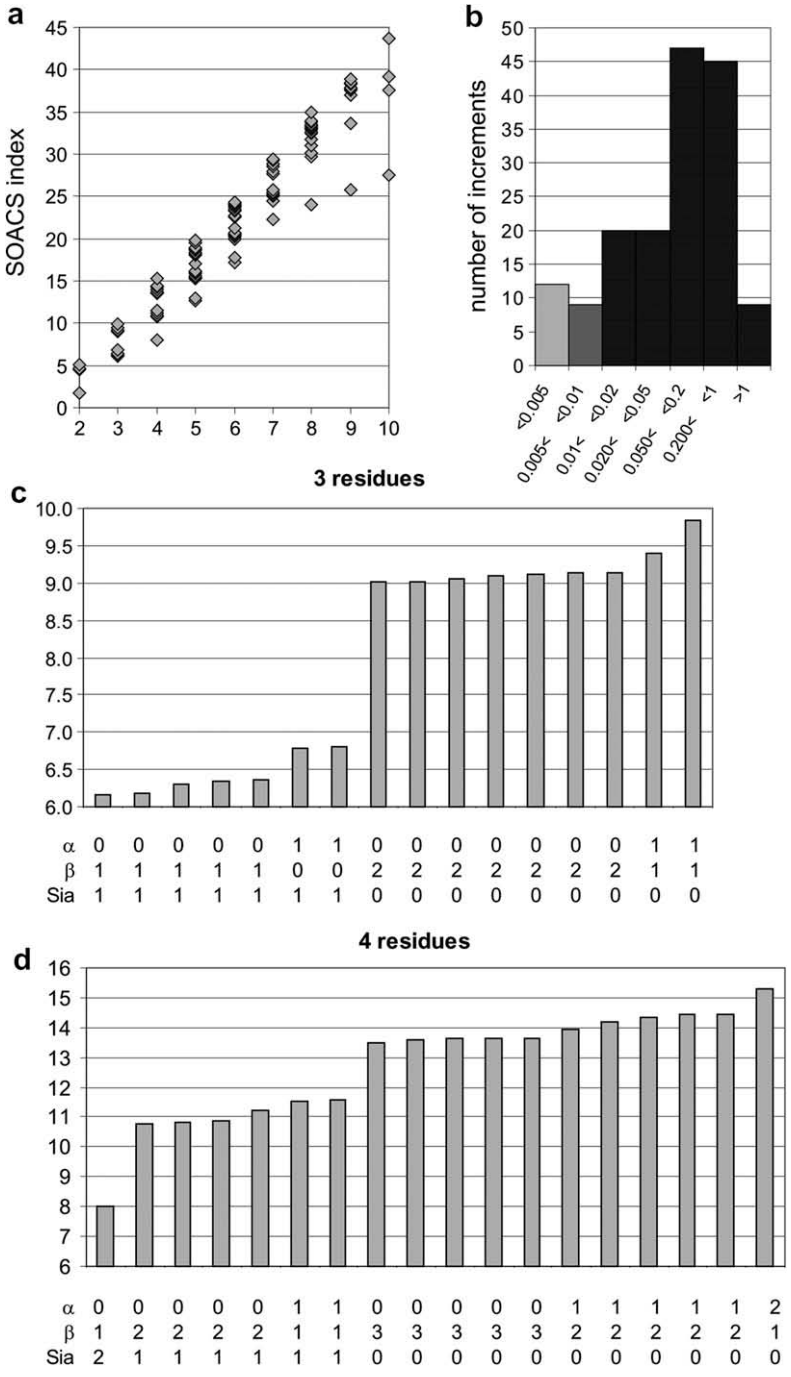

Figure 5. SOACS index of O-glycans from referenced O-glycans.[9] (a) When plotted against oligosaccharide size, SOACS index values are strictly size dependent. (b) The value increments between adjacent SOACS indexes have been distributed in five categories from <0.005 to >1 units. 87% of them are superior to 0.01 units and 74% are superior to 0.02 units. SOACS index values within homogenous families of O-glycans with (c) 3 and (d) 4 monosaccharides are dependent on the α-linked MS/β-linked MS/sialic acid (α/β/S) ratio of individual oligosaccharides.

6-linked β-Man*p* (residue 3, Fig. S3) H-1 (H1 Man3) value is not computed because its exact reading is rendered hazardous by the proximity of residual HOD signal at 300 K (~4.76 ppm) and is invisible in most cases. Then, the collection mode will vary according to the form of the N-glycan.

- For reduced N-glycans, the chemical shift of the reduced GlcNAc H-2 is not computed. In contrast to O-glycans, SOACS-ol is never informative due to the homogenous GlcNAc-ol parameters.
- For intact N-glycans linked to an aglycon moiety—either an oligopeptide or a chemical group (methyl, dansyl, fluorescent tag)—H-1 signal of the GlcNAc in reducing position is computed.
- For free N-glycans, with either an intact chitobiose or a single reducing GlcNAc residue, both α- (70%) and β-anomers (30%) should be taken into consideration without discrimination. Other anomeric signals of N-glycan core under the influence of reducing GlcNAc residue may be split into two different signals and should also be computed in a single SOACS index.

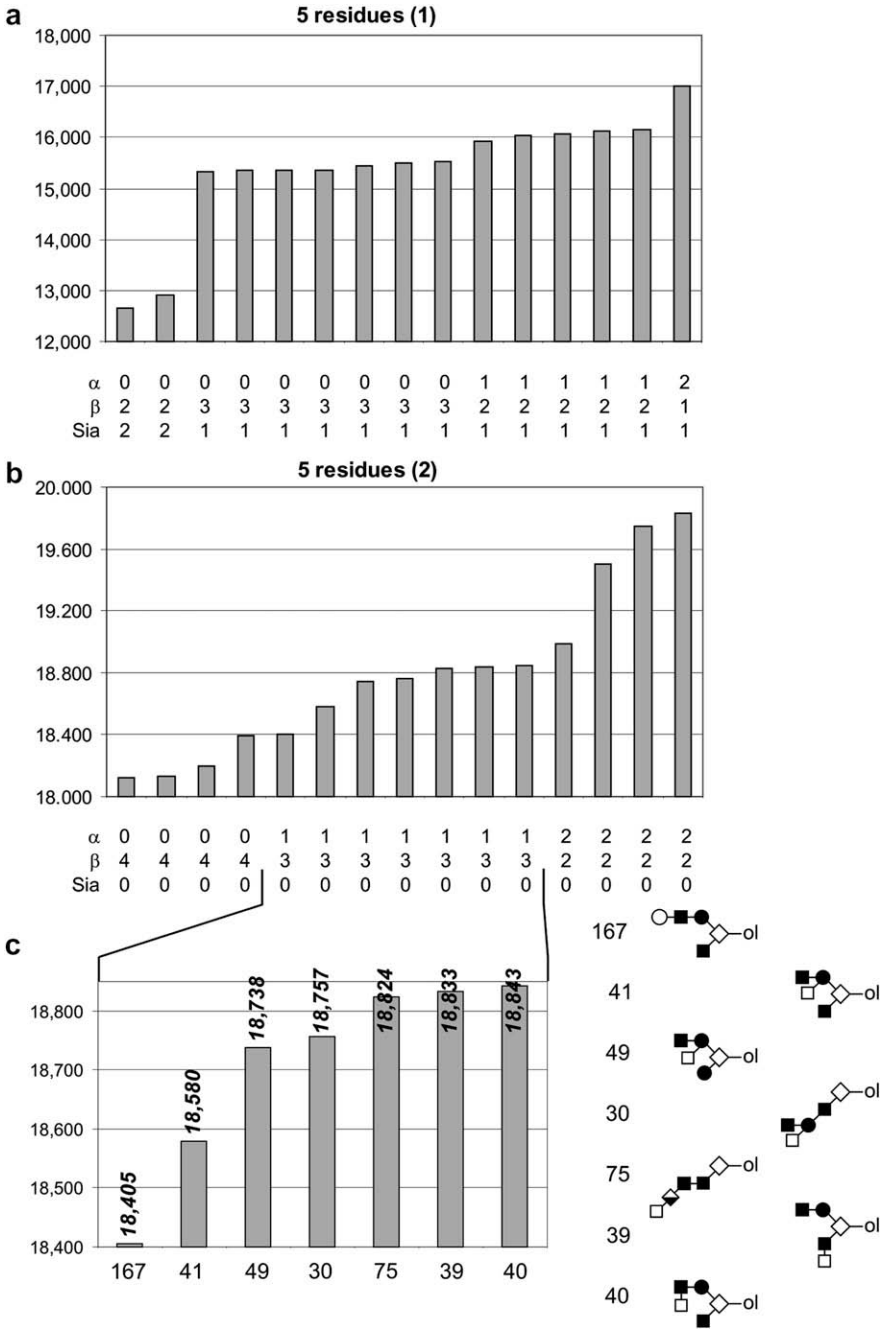

Figure 6. (a) and (b) SOACS index distribution of referenced O-glycans with five monosaccharides.[9] (c) Fine structure and SOACS index of pentasaccharide alditols with a 1:3:0 α/β/S ratio. Reference numbers correspond to Kamerling and Vliegenthart.[9]

A survey on a panel of 79 known N-Glycan structures from literature gave very similar results in term of values specificity than with O-glycans (Table S2, Fig. S2).[8,27]

Finally, SOACS index may be computed from simple mixtures of oligosaccharides (usually less than three oligosaccharides), as long as the different molecules are not present in equimolar proportions. Indeed, considering that NMR signal integration is quantitative, anomeric signals from two oligosaccharides whose relative quantities differ by more than 10% may be easily distinguished from each other and their chemical shifts values individually computed into distinct SOACS indexes. However, oligosaccharides that represent less than 10% of the major molecule cannot be confidently computed due to deterioration of signal to noise ratio.

2.5. Application to databank query

The SOACS and SOACS-ol indexes have been successfully implemented into the public database GLYCOBASE (available at http://glycobase.univ-lille1.fr/base/) as a query for molecule retrieval. GLYCOBASE is aimed to give access to NMR parameters of referenced glycans by providing fully interpreted NMR spectra as well as NMR acquisition files. So far, it references molecules that have been purified from a wide range of organisms in the Institute for Structural and Functional Glycobiology (Unité de Glycobiologie Structurale et Fonctionnelle, UGSF). Along with SOACS and SOACS-ol indexes, retrieval system includes a search by trivial names of glyco-epitopes, species of origin, nature of the charge, molecular weight,

and nature of constituting monosaccharides.[6] Research query may be either a combination of all these attributes or a single attribute, including SOACS or SOACS-ol indexes. When using SOACS/SOACS-ol-based research, user may enter either the individual observed NMR chemical shifts from any given spectrum or the already calculated indexes. In each case, he may choose to introduce an error margin whose value is the sum of estimated reading errors on individual signals. This can be as low as 0.001–0.002 ppm for any good quality ^{1}H NMR spectrum.

SOACS and SOACS-ol indexes have been developed as a simple query aimed to retrieve complex glycans from extensive databanks based on ^{1}H NMR parameters. It does not require in-depth knowledge of carbohydrate NMR spectra interpretation since indexes computation is based on easily recognized signals, mainly in the anomeric region of the spectra. Being a sum of chemical shifts, the indexes do not contain exploitable structural information, but as demonstrated by the use of several hundreds of molecules from the literature and existing database, these are specific enough to retrieve with a good confidence any given glycan structure. As such, it proved to be a valuable tool to easily check if a newly analyzed molecule has already been described and previously stored in the GLYCOBASE database. Then, in conjunction with other queries, its implementation in existing database is thought to increase the specificity of glycan retrieval by allowing the differentiation of close isomers.

3. Experimental

All purified glycans from amphibian species were analyzed by one- and/or two-dimension ^{1}H NMR spectroscopy using basic BRUKER® pulse programs. Samples were analyzed in 200 × 5 mm BMS-005B SHIGEMI® or standard 5 mm tubes. Prior to analyses, oligosaccharides were twice exchanged with 99.97% ^2H_2O and finally solubilized in ^2H_2O where the pD must be adjusted around 7. The chemical shifts were expressed relative to acetone methyl signals (δ_{1H} 2.225). When very low quantity of material is available, protons of methyl group from residual acetate salt may be used for calibration (δ_{1H} 1.909 ppm). Finally all spectra were recorded at 300 K.

In order to obtain a three-decimal precision for chemical shifts measurements high quality spectra are required. These may only be obtained from highly purified oligosaccharide samples exempt of salts, with a neutral pH solution, and uncontaminated by residual peptides. Finally, a manual fine adjustment of the spectrometer is necessary (i.e., tuning frequencies, optimizing field homogeneities, calibrating both pulse and delay calibrations, and proper temperature regulation).

Acknowledgments

This work was supported by the CNRS (Unité Mixte de Recherche CNRS/USTL 8576, Dir J.C. Michalski), by the Ministère de l'Enseignement Supérieur et de la Recherche, by the Institut Fédératif de Recherche 147 de Lille (Dir. J. Mazurier), and by the 'PPF Bio-informatique de Lille'.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.carres.2008.11.001.

References

1. http://www.boc.chem.uu.nl/sugabase/sugabase.html.
2. http://www.glyco.ac.ru/bcsdb/start.shtml.
3. http://www.glycosciences.de/sweetdb/.
4. Loss, A.; Stenutz, R., Schwarzer, E.; von der Lieth, C. W.; GlyNest and CASPER: Two Independent Approaches to Estimate ^{1}H and ^{13}C NMR Shifts of Glycans Available Through a Common Web-interface. Nucleic Acids Res., 2006, July 1; 34 (Web Server issue): W733-7.
5. http://www.glycome-db.org/About.action.
6. http://glycobase.univ-lille1.fr/base/.
7. Vliegenthart, J. F. G.; van Halbeek, H.; Dorland, L. *Pure Appl. Chem.* **1981**, *53*, 45–77.
8. Vliegenthart, J. F. G.; Dorland, L.; van Halbeek, H. *Adv. Carbohydr. Chem. Biochem.* **1983**, *41*, 209–374.
9. Kamerling, J. P.; Vliegenthart, J. F. G. *Biol. Magn. Res.* **1992**, *10*, 1–288.
10. Vliegenthart, J. F. G.; Kamerling, J. *Compr. Glycosci.* **2007**, *2*, 133–191.
11. Umemoto, J.; Bhavanandan, V. P.; Davidson, E. A. *J. Biol. Chem.* **1977**, *252*, 8609–8614.
12. Fan, J. Q.; Yamamoto, K.; Kumagai, H.; Tochikura, T. *Agric. Biol. Chem.* **1990**, *54*, 233–234.
13. Ishii-Karakasa, I.; Iwase, H.; Hotta, K.; Tanaka, Y.; Omura, S. *Biochem. J.* **1992**, *288*, 475–482.
14. Ishii-Karakasa, I.; Iwase, H.; Hotta, K. *Eur. J. Biochem.* **1997**, *247*, 709–715.
15. Spiro, R. G. *Methods Enzymol.* **1972**, *28*, 32–34.
16. Patel, T.; Bruce, J.; Merry, A.; Bigge, C.; Wormald, M.; Jaques, A.; Parekh, R. *Biochemistry* **1993**, *32*, 679–693.
17. Maes, E.; Florea, D.; Delplace, F.; Lemoine, J.; Plancke, Y.; Strecker, G. *Glycoconj. J.* **1997**, *14*, 127–146.
18. Märtz, L.; Altmann, F.; Staudacher, E.; Kubelka, V.; Protein Glycosylation in Insects. In '*Glycoproteins*', Montreuil, J.; Schachter, H., Vliegenthart, J. F. G., Eds.; 1995; pp 543–563.
19. Altmann, F. *Glycoconj. J.* **1997**, *14*, 643–646.
20. Lerouge, P.; Cabanes-Macheteau, M.; Rayon, C.; Fischette-Lainé, A. C.; Gomord, V.; Faye, L. *Plants Mol. Biol.* **1998**, *38*, 31–48.
21. Koide, N.; Muramatsu, T. *J. Biol. Chem.* **1974**, *249*, 4897–4904.
22. Hirani, S.; Bernasconi, R. J.; Rasmussen, J. R. *Anal. Biochem.* **1987**, *162*, 485–492.
23. Lee, Y. C.; Scocca, J. R. A. *J. Biol. Chem* **1972**, *247*, 5733–5738.
24. Bayard, B.; Roux, D. *FEBS-Lett.* **1975**, *55*, 206–211.
25. Florea, D.; Maes, E.; Haddad, M.; Strecker, G. *Biochimie* **2002**, *84*, 611–624.
26. Koerner, T. A. W.; Prestegard, J. H.; Yu, R. K. *Methods Enzymol.* **1987**, *138*, 38–59.
27. Haeuw, J. F.; Strecker, G.; Wieruszeski, J. M.; Montreuil, J.; Michalski, JC. *Eur. J. Biochem.* **1991**, *202*, 1257–1268.
28. Florea, D.; Maes, E.; Strecker, G. *Carbohydr. Res.* **1997**, *302*, 179–189.
29. Coppin, A.; Maes, E.; Morelle, W.; Strecker, G. *Eur. J. Biochem.* **1999**, *265*, 94–104.
30. Mourad, R.; Morelle, W.; Neveu, A.; Strecker, G. *Eur. J. Biochem.* **2001**, *268*, 1990–2003.
31. Morelle, W.; Guyétant, R.; Strecker, G. *Carbohydr. Res.* **1998**, *306*, 435–443.
32. Maes, E.; Florea, D.; Coppin, A.; Strecker, G. *Eur. J. Biochem.* **1999**, *264*, 301–313.
33. Coppin, A.; Maes, E.; Flahaut, C.; Coddeville, B.; Strecker, G. *Eur. J. Biochem.* **1999**, *266*, 370–382.
34. Delplace, F.; Maes, E.; Lemoine, J.; Strecker, G. *Biochem. J.* **2002**, *363*, 457–471.
35. Guerardel, Y.; Petit, D.; Madigou, T.; Guillet, B.; Maes, E.; Maftah, A.; Boujard, D.; Strecker, G.; Kol, O. *FEBS-Lett.* **2003**, *554*, 330–336.
36. Coppin, A.; Maes, E.; Strecker, G. *Carbohydr. Res.* **2002**, *337*, 121–132.