Preprint

This is the submitted version of a paper published in *ChemPlusChem*.

Facile structural elucidation of glycans using NMR data and the program CASPER: application to the O-antigen polysaccharide of *E. coli* O155

Carolina Fontana,[a] Andrej Weintraub,[b] and Göran Widmalm*[a]

Carbohydrates, the most abundant biomolecules found in nature, are present in all forms of life playing essential roles in a wide range of biological processes. The structural complexity of these molecules confers them an immense capacity to carry information in biological systems, acting as efficient mediators in the interaction of the cell with the environment. Hence, in order to understand the role of these carbohydrates, it is of key importance to have knowledge of their structures and, considering the features of these molecules, the number of possible compounds that can be generated with a given number of different building blocks (monosaccharides) is by far larger than for any other biopolymer.[1–3] The description of the primary structure not only comprises the identification of the component monosaccharides and their sequence in the polymer, but also the characterization of the ring form of each residue, their anomeric configuration and the linkage positions, as well as the presence of additional modifications.[4] In this regard, NMR spectroscopy is one of the most powerful tools for structural determination of glycans, since each of these structural features can be elucidated using this technique,[5–8] but the limited spectral dispersion of both ^{1}H and ^{13}C nuclei in carbohydrates can make the process of assignments of certain resonances tedious, and this is still the most time consuming part of the process. The program CASPER (http://www.casper.organ.su.se/casper/), on the other hand, is a promising tool to help to overcome this problem, as the whole analysis of the NMR data can be carried out in an automated or semi-automated manner. This software uses liquid state NMR data to elucidate the structure of glycans based on their ^{1}H and ^{13}C chemical shifts as well as ^{1}H-^{1}H or ^{1}H-^{13}C correlations from 2D experiments;[9,10] a module for component and absolute configuration analysis has also been implemented, allowing the fully automated analysis of glycans using solely unassigned NMR data as input information.[11] Herein, we demonstrate that a previously unknown O-antigen

polysaccharide (PS) structure can successfully and rapidly be determined using the program CASPER.
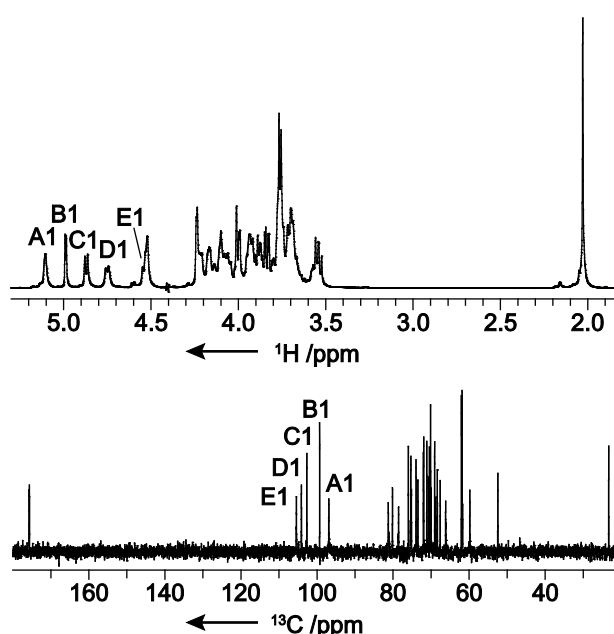
Figure 1. Diffusion-filtered ^{1}H and ^{13}C NMR spectra (*top* and *bottom*, respectively) of the O-antigen PS of *E. coli* O155.[12] The constituent sugar residues were denoted **A-E** in order of decreasing ^{1}H chemical shifts of their anomeric protons (resonances from anomeric atoms are annotated in the respective spectra)

An outline of the methodology used in this study is shown in Fig. S1, and one should note that all the information submitted to CASPER is from *unassigned* NMR data. The sugar components of the O-antigen PS of *E. coli* O155 and their absolute configuration were determined using a methodology previously developed in our laboratory.[11] The PS was hydrolyzed, followed by (+)-2-butanolysis of the sugar components. Unassigned ^{13}C chemical shifts from a ^{13}C spectrum and ^{13}C-^{1}H correlations from a multiplicity-edited ^{1}H,^{13}C-HSQC spectrum of the mixture of (+)-2-butyl glycosides were used as input information in the 'component analysis' module of the web interface of the CASPER program (Table S1), and the results (calculation time ~1 s) revealed three possible monosaccharide components: D-Gal, D-GalNAc and D-GalA (Table S2). The relative intensities of the cross-peaks in the ^{1}H,^{13}C-HSQC spectrum (Figure S2), revealed that the D-Gal monosaccharide is a major component and the resulting fit

[a]   *C. Fontana and Prof. G. Widmalm*
      Department of Organic Chemistry
      Stockholm University
      Arrhenius Laboratory, S-106 91 Stockholm, Sweden
      Fax: (+46) 815 49 08
      E-mail: gw@organ.su.se

[b]   *Prof. A. Weintraub*
      Department of Laboratory Medicine
      Division of Clinical Microbiology, Karolinska Institute
      Karolinska University Hospital, S-141 86 Stockholm, Sweden

      Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cplu.20xxxxxxx (Tables S1-S10 and Fig.S1-S4)

observed for the remaining components (D-GalNAc and D-GalA) could be attributed to their lower concentration in the sample (i.e. some of the correlations for the minor components of these butyl glycosides were not observed or at the noise level). The [1H] and [13C] NMR spectra of the PS (Fig. 1) revealed five resonances corresponding to anomeric atoms, which indicate that the PS is composed of pentasaccharide repeating units. In the [13C] NMR spectrum two signals corresponding to carbonyl atoms were observed at 175.64 and 175.75 ppm, and one resonance from the methyl group of an *N*-acetyl moiety at 23.28 ppm. Thus, the ratio of D-Gal, D-GalNAc and D-GalA in the repeating units was predicted to be 3:1:1, respectively.
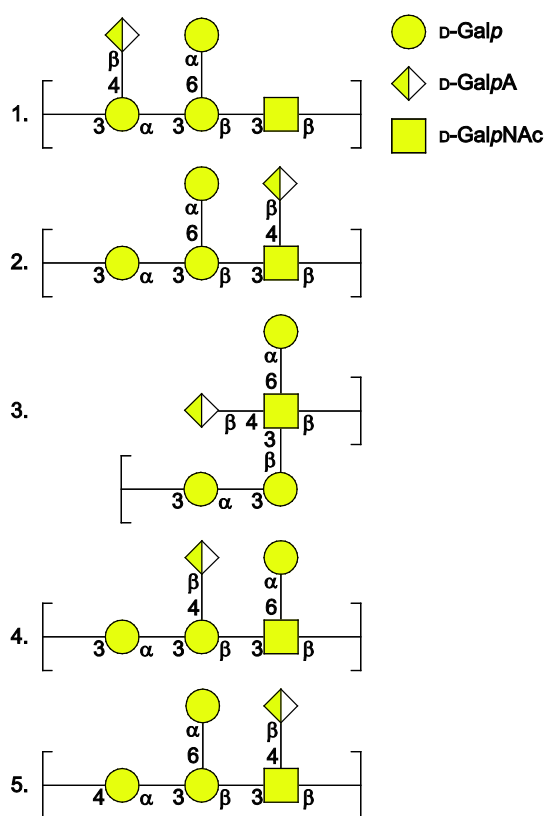

Figure 2. CASPER output of the five top-ranked structures, in CFG format, for the O-antigen PS of *E. coli* O155. The relative deviations for structures 1-5 are 1.00, 1.15, 1.18, 1.24 and 1.29, respectively. For standard carbohydrate listening of the ten top-ranked structures see Table S5, and for NMR assignments of the three top-ranked structures see Tables S6-8.

Subsequently, the CASPER program was utilized in the structural elucidation of the repeating unit of the O-antigen PS. In the first step, the analysis was exclusively focused on CASPER capacity to predict [13C] chemical shifts and the experimental data was obtained from a [1H]-decoupled [13C] NMR spectrum (Table S3). In addition, $^3J_{H1,H2}$ were obtained from a [1H] NMR spectrum and submitted to CASPER as follow: three couplings constants larger than 7 Hz and one in the range 2-7 Hz (the remaining unresolved coupling constant was not considered). The structural information was given as follow: one D-GalA, three D-Gal and one D-GalNAc, each with all the linkage positions marked as unknown. The WecA repeating unit restriction available in CASPER was also used, since it was known that the O-antigen PS of *E. coli* O155 is

biosynthetized by the Wzx/Wzy-dependent pathway.[13,14] The calculation (~4 min) returned a list of ten possible structures with a small score differences (see Table S4). All these structures (which are positional isomers where some common features can be recognized; e.g. the GalA residue is always terminal, none of the residues is 2-substituted, and no more than one Gal residue is 4-substituted) are difficult to discriminate from each other exclusively from the point of view of [13C] chemical shifts, as all the component monosaccharides have the same *galacto*-configuration. In order to increase the score differences between the candidate structures, additional information had to be submitted to CASPER, and thus data from [1H],[1H]-TOCSY, [1H],[13C]-HSQC, [1H],[13C]-HMBC and [1H],[13C]-H2BC were added, as well as information of $^1J_{C1,H1}$ (two > 169 Hz and three < 169 Hz) obtained from a non-decoupled [1H],[13C]-HSQC spectrum. The substitution patterns observed in the ten top-ranked structures listed in Table S4 were also used as restrictions, and submitted as follow: (*i*) a non-substituted GalA, (*ii*) a 3- and 4-substituted Gal, (*iii*) two 3- and 6-substituted Gal, and (*iv*) a 3-,4- and 6-substituted GalNAc; and the option to consider all the possible combinations within the specified linkage positions was selected for the GalNAc and Gal residues. These restrictions aimed to considerably reduce the calculation time, since a lower number of structures will have to be considered and evaluated by the software. On the other hand, it is of interest to prove that the correct structure was listed as one of the ten top-ranked structures during the first calculation, showing the significance of an approach that uses chemical shifts predictions in combination with the interpretation of unassigned NMR data. The calculation took ~3 s and the structure at the top of the new list (fourth in the previous calculation) showed a significant score difference with respect to the second and third top-ranked structures (Fig. 2 and Table S5). The structure at the top of the list could readily be confirmed as the correct structure using an additional [1H],[1H]-NOESY spectrum (Fig. S4) and the [1H] chemical shifts assignments given by CASPER (Table S6-8). [1H] and [13C] chemical shifts assignments of the PS, as well as inter-residue correlations from [1H],[1H]-NOESY and [1H],[13C]-HMBC experiments are compiled in Table S9. Once the structure was determined, the 'calculate chemical shifts' module of the CASPER program was used to predict the [1H] chemical shifts of the terminal non-reducing end moiety of the PS; facilitating the identification of the respective cross-peaks of lower intensity in the NMR spectra (Table S10) and confirming that the structure of the biological repeating unit of the PS is as shown in Fig. 3. Integration of the anomeric protons of residue **C** and **C'** (terminal sugar residue) in the [1H] NMR spectrum revealed that the PS preparation consisted of ~20 repeating units on average.

In conclusion, these results demonstrate that the CASPER program can successfully be used, in combination with NMR data, to rapidly determine the structure of the biological repeating unit of a new PS structure. Even though this polymer is composed of regular monosaccharide residues, its complexity lays in the fact that all five components have the same *galacto*-configuration and, consequently, very similar [1H] and [13C] NMR chemical shifts are expected. The two-step protocol proposed herein eludes the use of information from [1H] chemical shifts in the first calculation (which are expected to be very similar in most of the possible positional isomers). The manual interpretation of such NMR data can become very

tedious and time consuming, in particular for researchers with little experience in the interpretation of NMR of carbohydrates. Using the CASPER program, the structural elucidation process could then be reduced from several hours (or even days) of manual interpretation to ~4 min of automated analysis.
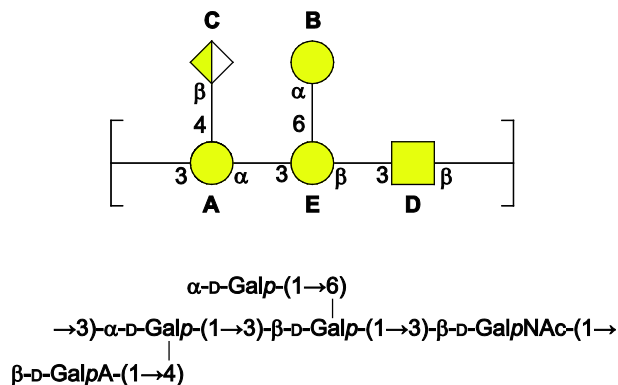
Figure 3. Structure of the biological repeating unit of the O-antigen PS of *E. coli* O155 in CFG notation (*top*) and standard nomenclature (*bottom*).

Experimental Section

Preparation of the LPS. *E. coli* O155 (CCUG 36524) was obtained from the Culture Collection, University of Gothenburg. The bacterium was grown and the LPS extracted as previously described.[15]

Preparation of the PS. The lipid-free PS (5.7 mg) was obtained from the LPS material (26.5 mg) as described previously.[16]

Preparation of (+)-2-butyl glycosides from the PS. The O-antigen PS of *E. coli* O155 (3.1 mg) was hydrolyzed with 2M TFA (1 ml, 120 °C, 24 h) and the (+)-2-butyl glycosides were prepared as described before.[11]

NMR spectroscopy. The spectra of the (+)-2-butyl glycosides of the PS hydrolyzate were recorded in D_2O (0.55 mL in a 5 mm NMR tube) at 70 °C using a Bruker Avance III 700 MHz spectrometer equipped with a 5 mm Z-gradient TCI ($^1H/^{13}C/^{15}N$) CryoProbe. The spectra of the O-antigen PS (2.6 mg) were recorded in D_2O (0.55 mL in a 5 mm NMR tube) at 60 °C using a Bruker Avance 500 MHz spectrometer equipped with a 5 mm Z-gradient TCI ($^1H/^{13}C/^{15}N$) CryoProbe. Chemical shifts are reported in ppm relative to internal sodium 3-trimethylsilyl-(2,2,3,3-$^2H_4$)propanoate (TSP, δ_H 0.00) or 1,4-dioxane in D_2O (δ_C 67.40).

Acknowledgements

[1] C. R. Bertozzi, D. Rabuka, in *Essentials of Glycobiology* (Eds.: A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler), Cold Spring Harbor Laboratory Press, New York, **2008**, pp. 23–36.

[2] U. Aich, K. J. Yarema, in *Carbohydrate-Based Vaccines and Immunotherapies*, John Wiley & Sons, Inc., **2008**, pp. 1–53.

[3] H. Ghazarian, B. Idoni, S. B. Oppenheimer, *Acta Histochem.* **2011**, *113*, 236–247.

[4] J. P. Kamerling, in *Comprehensive Glycoscience* (Eds.: J.P. Kamerling, G.-J. Boons, Y.C. Lee, A. Suzuki, N. Taniguchi, A.G.J. Voragen), Elsevier Ltd, Oxford, **2007**, pp. 1–38.

[5] G. Widmalm, in *Comprehensive Glycoscience* (Eds.: J.P. Kamerling, G.-J. Boons, Y.C. Lee, A. Suzuki, N. Taniguchi, A.G.J. Voragen), Elsevier Ltd, Oxford, **2007**, pp. 101–132.

[6] J. F. G. Vliegenthart, in *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances* (Eds.: J.F.G. Vliegenthart, R.J. Woods), American Chemical Society, Washington, DC, **2006**, pp. 1–19.

[7] J. Ø. Duus, C. H. Gotfredsen, K. Bock, *Chem. Rev.* **2000**, *100*, 4589–4614.

[8] J. Jiménez-Barbero, T. Peters, Eds., *NMR Spectroscopy of Glycoconjugates*, Wiley-VCH, Weinheim, **2003**.

[9] P.-E. Jansson, L. Kenne, G. Widmalm, *Carbohydr. Res.* **1989**, *188*, 169–191.

[10] M. Lundborg, G. Widmalm, *Anal. Chem.* **2011**, *83*, 10–13.

[11] M. Lundborg, C. Fontana, G. Widmalm, *Biomacromolecules* **2011**, *12*, 3851–3855.

[12] I. Ørskov, F. Ørskov, B. Rowe, *Acta Pathol. Microbiol. Scand., Sect. B: Microbiol. Immunol.* **1973**, *81B*, 59–64.

[13] H. Guo, Q. Kong, J. Cheng, L. Wang, L. Feng, *FEMS Microbiol. Lett.* **2005**, *248*, 153–161.

[14] R. Stenutz, A. Weintraub, G. Widmalm, *FEMS Microbiol. Rev.* **2006**, *30*, 382–403.

[15] M. V. Svensson, A. Weintraub, G. Widmalm, *Carbohydr. Res.* **2011**, *346*, 449–453.

[16] P. Chassagne, C. Fontana, C. Guerreiro, A. Gauthier, A. Phalipon, G. Widmalm, L. A. Mulard, *Eur. J. Org. Chem.* **2013**, 4085–4106.

Entry for the Table of Contents

COMMUNICATION

C. Fontana, A. Weintraub and G. Widmalm*

Page No. – Page No.

Facile structural elucidation of glycans using NMR data and the program CASPER: application to the O-antigen polysaccharide of *E. coli* O155

The program CASPER was successfully employed to rapidly elucidate a new O-antigen polysaccharide structure (obtained from a strain of *E. coli* serogroup O155), using solely *unassigned* NMR data as input information. Thus, what is considered the most tedious and time-consuming part of the structural elucidation process has been reduced from several hours (or even days) of manual interpretation to ~4 min of automated analysis.