Improved Carbohydrate Structure Generalization Scheme for 1 H and 13 C NMR Simulations

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2015

Impact Factor: 5.64 · DOI: 10.1021/acs.analchem.5b01413 · Source: PubMed

2 AUTHORS:

Roman R. Kapaev
Russian Academy of Sciences
4 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Philip Toukach
Russian Academy of Sciences
75 PUBLICATIONS **945** CITATIONS
SEE PROFILE

analytical chemistry

Technical Note

pubs.acs.org/ac

Improved Carbohydrate Structure Generalization Scheme for ^{1}H and ^{13}C NMR Simulations

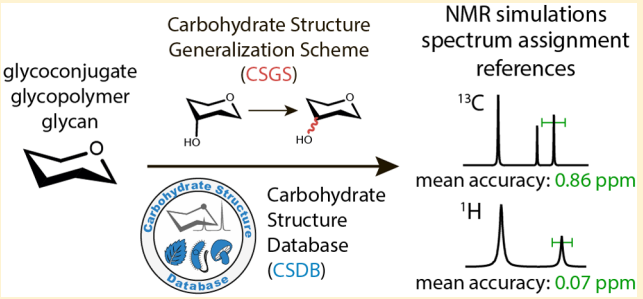
Roman R. Kapaev*,† and Philip V. Toukach*,‡

†Higher Chemical College of the Russian Academy of Sciences, Miusskaya sq. 9, Moscow 125047, Russia

‡N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky prosp. 47, Moscow 119991, Russia

Supporting Information

ABSTRACT: The improved Carbohydrate Structure Generalization Scheme has been developed for the simulation of ^{13}C and ^{1}H NMR spectra of oligo- and polysaccharides and their derivatives, including those containing noncarbohydrate constituents found in natural glycans. Besides adding the ^{1}H NMR calculations, we improved the accuracy and performance of prediction and optimized the mathematical model of the precision estimation. This new approach outperformed other methods of chemical shift simulation, including database-driven, neural net-based, and purely empirical methods and quantum-mechanical calculations at high theory levels. It can process structures with rarely occurring and noncarbohydrate constituents unsupported by the other methods. The algorithm is transparent to users and allows tracking used reference NMR data to original publications. It was implemented in the Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS) web service, which is freely available at the platform of the Carbohydrate Structure Database (CSDB) project (http://csdb.glycoscience.ru).



Despite a wide range of remarkable applications, such as discovery of new drugs and vaccines[1−5] and development of new biofuels,[6] glycoscience still suffers from the lack of structural data.[7] NMR spectroscopy is the major tool for structural studies of carbohydrates; however, the interpretation of NMR observables is still a tedious task. Fortunately, progress in informatics gave an opportunity to make the research easier. In particular, NMR spectrum simulators are quite useful tools, which makes it possible to confirm and rank a proposed structure and serve as a basis for solving the inverse problem, i.e., the automated elucidation of a structure from experimental NMR observables.[7−9] As compared to ^{1}H NMR, ^{13}C NMR spectroscopy provides a wider range and better reproducibility of chemical shifts and is indispensable in glycobiology. However, ^{1}H NMR spectra promote the computer-assisted structural elucidation by significant reduction of the number of structural hypotheses. Among existing methods of ^{13}C NMR simulation, the Carbohydrate Structure Generalization Scheme (CSGS) was shown to be the most suitable approach for saccharides and glycoconjugates.[10] However, it has not been reported to simulate ^{1}H NMR chemical shifts, although CSGS can potentially operate with any atomic observables, and the Carbohydrate Structure Database[11,12] (CSDB) used for ^{13}C NMR simulations contains thousands of ^{1}H NMR spectra. Moreover, CSGS had a number of drawbacks, such as nonoptimized algorithmic features and poor performance, in comparison with other empirical approaches.[10]

Here, we report an improved CSGS tuned for ^{1}H and ^{13}C NMR spectrum simulation of various carbohydrates, including those containing amino acids, lipids, phosphates, alditols, and other noncarbohydrate constituents. The algorithm was enhanced to evaluate the credibility of the predicted values, and the software implementation got new features.

EXPERIMENTAL SECTION

Carbohydrate Structure Generalization Scheme basic principles and used terms have been published recently.[10] Minor algorithmic improvements and user interface changes are listed in the Supporting Information, Section 1 (including Figures S-1 and S-2).

Tuning CSGS for ^{1}H NMR Simulations. Generally, the CSGS approach is applicable to the simulation of any atomic observables without major algorithmic changes: only the generalization weights need to be optimized. Consequently, tuning the weights for ^{1}H NMR predictions was a simple procedure, because the developed optimization technique was universal.

Weight Optimizations. The generalization weights for ^{13}C and ^{1}H NMR simulations were optimized using a method based on the modified Artificial Bee Colony[13] (ABC) algorithm. The ABC algorithm has been studied extensively and applied to solve optimization problems;[14] it was shown to outperform the evolutionary strategies[15] applied for the previous weight optimizations.[10] It was expected to be more

Received: April 15, 2015
Accepted: June 18, 2015
Published: June 18, 2015

DOI: 10.1021/acs.analchem.5b01413
Anal. Chem. 2015, 87, 7006−7010

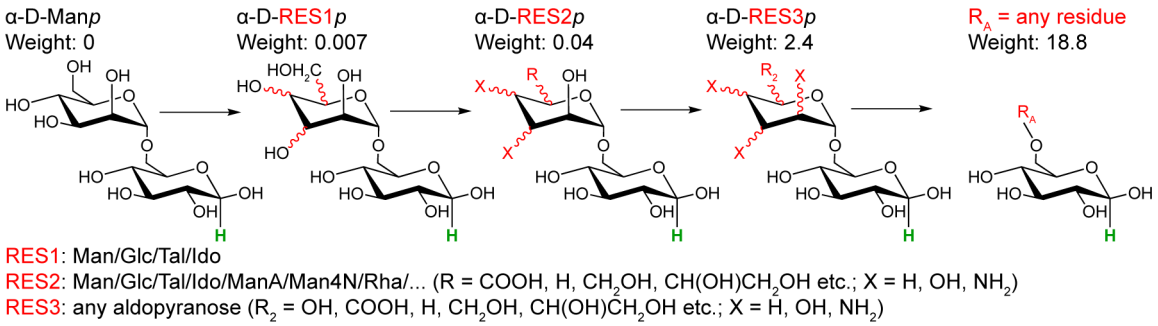


Figure 1. Four generalization steps of the α-D-Man*p* substituent in the α-D-Man*p*-(1 → 6)-β-D-Glc*p* disaccharide (*extreme* prediction mode, $n = 4$). Generalization weights are given for the prediction of β-D-Glc*p* H1 (marked in green). Wavy bonds indicate any stereo configuration. Generalized properties are shown in red.

Table 1. Structures Used for Comparison of Accuracy and Performance of Different ^{1}H Simulation Methods[a]

structure	peculiarities
(**1**)[16] β-D-Glc*p*NAc-(1→3)-α-D-Gal*p*	simple disaccharide
(**2**)[17] →3)-β-D-Gal*p*-(1→3)-α-D-Gal*f*-(1→ (D-galactan)	simple polymer repeating unit containing furanose
(**3**)[18] β-D-Fru*f*-(2→1)-α-D-Glc*p* (sucrose)	furanose residue, uncommon 2→1 glycosidic linkage
(**4**)[19] β-D-Glc*p*NAc-(1→P→P→5)-β-D-Rib*f*-(1→N)-uracil (UDP-β-D-Glc*p*NAc)	rarely occurring and poorly parametrized residues within uridine diphosphate
(**5**)[20] L-Ala-(2→1)-L-Glu-(2→6)-α-D-Gal*p*NAcA-(1→4)-D-GalNAc-ol	rarely occurring alditol and amino acid residues
(**6**)[21] →3)-β-D-Gal*p*-(1→1)-[β-D-Gal*p*-(1→2)]-D-Gro-(3→P→	polymer repeating unit, phosphate group, glycerol, bisubstitution at neighboring positions
(**7**)[22] →2)-[α-Abe*p*-(1→3)]-α-D-Man*p*-(1→4)-α-L-Rha*p*-(1→3)-α-D-Gal*p*-(1-	polymer repeating unit, bisubstitution at neighboring positions, deoxy sugars, rarely occurring pentose
(**8**)[23] →4)-4HOBut-(1→7)-β-Pse*p*-(2→	polymer repeating unit, higher sugar with flexible tail (pseudaminic acid), aliphatic residue

[a]References correspond to the published experimental ^{1}H NMR data in D_2O.

useful to find the best fitting weight sets. In addition, we added a separate weight set for alditol residues. Detailed description of the procedure and the resultant weights are provided in the Supporting Information (Section 1, Tables S-1−S-4).

Optimization of the Substituent Generalization Scheme. Previously, the generalization of substituents of the residue under prediction was based on the thresholds determining which substituent properties were generalized at every certain step.[10] Such an approach did not allow the strict control of the number of generalization steps.

This drawback led to the irregularity of the substituent generalization weights: the steps had small or large weights, while medium-weight steps were often missing, sometimes causing the suboptimal use of the database data.

To fix this issue, we made the number of generalization steps the determining factor instead of the thresholds. If the number of generalization steps is n, $\lfloor iD/n \rfloor$ (rounding iD/n down) descriptors of minimal weight are generalized at the i-th step of the generalization of a substituent with D structural descriptors (see Figure 1; for the list of structural descriptors, please refer to the recent publication[10]). We made n dependent on the prediction mode and the distance between the predicted atom and the linkage with a substituent residue (see Supporting Information, Table S-5).

Trustworthiness Estimation Algorithm. In our previous study,[10] it was possible to estimate the level of precision for every predicted chemical shift. Trustworthiness (T) was dependent on the weight of applied generalizations (W), the number of chemical shifts in a data set used for the prediction (N), and the standard deviation between values in the data set (σ). However, the formula used for the trustworthiness estimation has not been proven best. In the optimized algorithm, the trustworthiness is calculated according to the formula:

$$T = 100 - P_W(W) - P_N(1/N) - P_\sigma(\sigma)$$

where $P_X(X)$ is a square polynomial function:

$$P_X(X) = x_1X + x_2X^2$$

where x_1 and x_2 are factors reflecting the nature of the formal parameter X (W, N, or σ) and the type of residue (pyranose, furanose, etc.) to which the predicted atom belongs. The optimal x_1 and x_2 parameter sets were determined using an algorithm similar to the weight optimization procedure. In the case of trustworthiness, our purpose was to maximize the Pearson correlation coefficient between T and absolute deviation of the predicted values from the experimental ones ($\Delta = |\delta_{sim} - \delta_{exp}|$) for several structure samplings. Detailed description of the formula, optimization procedure, and the resultant x_1 and x_2 values are provided in the Supporting Information (Section 4, Tables S-6−S-8, and Figure S-3).

Verification. Estimation of the prediction accuracy was performed on the pool of various structures stored in the Bacterial CSDB (BCSDB), including pyranose, furanose, alditol, amino acid, lipid, and aliphatic residues (see Table S-6, Supporting Information, for details). A total number of experimental chemical shifts in the samplings was 5886 for ^{13}C and 6181 for ^{1}H NMR simulations. Predictions were carried out using BCSDB in the *accurate* mode[10] with unrestricted solvent, pH, and temperature parameters. Average time required for the simulation was measured using the web server hosting CSDB. Here and below, the program was not allowed to use database records containing the spectra of the structure being processed to avoid the prediction bias. In contrast to the

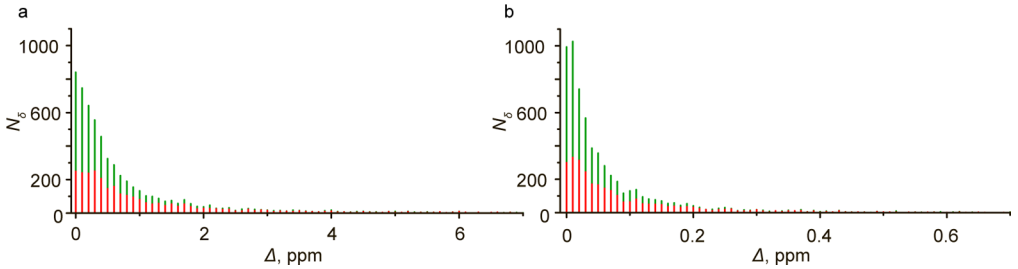
DOI: 10.1021/acs.analchem.5b01413
Anal. Chem. 2015, 87, 7006−7010

Table 2. Root-Mean-Square Deviations (Simulated vs Experimental) and Simulation Time (in Parentheses) for Various ^{1}H Simulation Methods Applied to Test Structures 1−8 (see Table 1)[a]

method[b]	structure							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
improved CSGS (GODESS)	0.07 (4.1 s), *0.05 (6.2 s)*	0.09 (8.2 s), *0.08 (41 s)*	0.11 (8.6 s), *0.11 (28 s)*	0.03 (1.4 s), *0.03 (1.4 s)*	0.102 (19.6 s), *0.107(345 s)*	0.12 (9.8 s), *0.10 (36 s)*	0.03 (3.2 s), *0.03 (3.2 s)*	0.08 (1.1 s), *0.08 (7.1 s)*
empirical (CASPER)[9]	0.07 (<2 s)	X	X	X	X	X	0.04 (<2 s)	X
HOSE + neural net (ACD/Labs 10)[24]	0.28 (<2 s)	X	0.12 (<2 s)	0.11 (<2 s)	0.31 (<2 s)	X	X	X
Modgraph (ChemBioDraw Ultra 14.0)	0.53 (<2 s)	X	0.30 (<2 s)	0.36 (<2 s)	0.38 (<2 s)	X	X	X
GIAO B3LYP/6-311G++(2d,2p) + COSMO (Gaussian 09)[25]	0.36 (152 h)	X	0.29(67.8 h)	0.70 (206 h)	no data	X	X	X

[a]For the CSGS approach, results in the *accurate* mode are in regular font, whereas results in the *extreme* mode are in italic; "X" indicates that the simulation is not supported by the software. [b]Software used for the simulations is specified in parentheses.

Chart 1. Distribution of the Absolute Deviation Δ (Predicted vs Experimental) for (a) ^{13}C and (b) ^{1}H NMR Simulations[a]



[a]All residue types were involved in these simulations (see Table S-6, Supporting Information, for sampling characteristics). Red bars correspond to chemical shifts, the simulation of which required structure generalizations; green bars correspond to atoms strictly represented in the database. N_δ is the number of simulated chemical shifts.

previous study,[10] the samplings did not include N- and O-linked acetic acid residues due to their exceptional abundance in the database; such high abundance led to very precise and fast predictions of their chemical shifts.

To compare the accuracy and performance of the ^{1}H NMR simulation with other methods, we selected eight structures covering various structural features of natural carbohydrates and their derivatives[10] (see Table 1). To evaluate the improvement of the trustworthiness estimation algorithm, we calculated Pearson correlation coefficients between T and Δ before and after optimizations using the same samplings as for accuracy and performance verification.

Materials and Methods. The NMR simulation software was implemented on top of the Carbohydrate Structure Database (CSDB), powered by the MySQL 5.0.70 relational database engine and PHP 5.5.14 scripts. The web interface uses DHTML 4, CSS 2, and JavaScript 1.2; it was tested in Google Chrome 41, Mozilla Firefox 36, and Internet Explorer 10. Statistical data processing was performed in OriginPro 9.0.0 and Microsoft Excel 2013. A personal computer (Intel Core 2 X4, 3.0 GHz) was used for quantum-mechanical calculations. All services are freely available under the "NMR simulation" menu item at the CSDB Web site (http://csdb.glycoscience.ru).

RESULTS AND DISCUSSION

Representative Simulation Examples. For all the structures listed in Table 1, the improved CSGS approach gave the most precise predictions of ^{1}H NMR chemical shifts (see Table 2). On the average, the CSGS simulations in the *extreme* mode were slightly more accurate than in the *accurate* mode. The accuracy of the CASPER[9] approach was similar to that of CSGS. However, the number of structural features supported by CASPER is limited, and it did not allow simulating the NMR observables for a relatively widespread furanose-containing galactan (example (**2**)). For a wide range of polymeric carbohydrates, CSGS is the only approach supporting the ^{1}H NMR simulations.

In most cases, the CSGS performance was close to that of other empirical approaches (Table 2). Slow processing (up to 5−10 min) usually took place in the *extreme* mode for multibranched structures or ones containing uncommon residues (example (**5**)).

Statistical Measurements. Accuracy and speed of the old[10] vs improved CSGS ^{13}C NMR predictor were measured as described in the Experimental Section. Algorithmic optimizations resulted in the 36% average accuracy increase on a test pool of various types of structures (from 1.17 to 0.86 ppm per resonance) and 4.6 times acceleration (from 3.4 to 0.7 s per resonance).

Overall distribution of the absolute deviation between the simulated and experimental chemical shifts is provided in Chart 1. For ^{1}H NMR, 95% of the predicted chemical shifts lay within 0.28 ppm from the experimental ones; the average deviation was 0.072 ppm. For ^{13}C NMR, 95% of the predicted values lay within 3.0 ppm from the experimental ones. For more details, see Charts S-1 and S-2, Supporting Information.

Trustworthiness Evaluation. Comparison of the optimized and unoptimized[10] formulas of trustworthiness estimation showed that the linear correlation coefficients between T and Δ increased by 20−100% depending on the nature of the central residue (see Table S-9, Supporting Information). For the optimized formulas, the correlation was from 0.24 to 0.61 depending on the nuclei and the central residue type. Aside from the linear coefficients, we calculated the linearization parameters to estimate the expected Δ from T (see Supporting

DOI: 10.1021/acs.analchem.5b01413
Anal. Chem. 2015, 87, 7006−7010

Information, Table S-8). Consequently, we allowed users to get comprehensible Δ values of expected error in addition to the abstract trustworthiness.

However, it should be noted that the expected Δ calculated from T by linear regression is approximation only. Aside from W, N, and σ, there appear to be a lot of obscure parameters affecting the prediction accuracy which are very difficult to consider.

Moreover, foreseeing the optimal dependency (linear, square, exponential, etc.) of the trustworthiness on W, N, and σ is problematic due to the diversity attributable to each of the input parameters; it is also unclear if the parameters should affect the trustworthiness independently. We expect that the improvement of the spectrum hybridization[10] algorithm, which benefits from both NMR simulators implemented in Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS), may help to level the above-listed ambiguities and to increase the correlation between T and Δ.

Future Prospects. Tuning CSGS for ^{1}H NMR simulations enables the prediction of 2D NMR spectra, such as ^{1}H−^{1}H COSY, ^{1}H−^{13}C HSQC, ^{1}H−^{13}C HMBC, and TOCSY. Implementation of the 2D NMR prediction service at the qualitative level is a subject of further study. Consequently, the 2D NMR support opens up an opportunity to develop a multifactorial algorithm of structure elucidation, based on user data from 1D (^{13}C, ^{1}H) and 2D NMR spectra. Due to its multifactorness, the automated NMR-based structure prediction service may become a worthy solution for unravelling various carbohydrate structures.

CONCLUSION

The glyco-tuned CSGS approach, applicable for simulations of any database-represented atomic observables, was applied for ^{1}H NMR chemical shift predictions. It showed higher accuracy or wider applicability than modern ^{1}H NMR prediction instruments, such as ACD/NMR, Modgraph NMRPredict, CASPER, and quantum-mechanical calculations at high theory levels with large basis sets. In contrast to the other methods, it supports most of the structural features present in natural carbohydrates, including uncommon residues and linkage types and noncarbohydrate constituents. It has perceptible performance superiority over quantum-mechanical methods and provides better accuracy similar to that of other glyco-tuned empirical approaches. The prediction algorithm was optimized, which resulted in better accuracy and speed of ^{13}C NMR simulations compared to the previous version of CSGS. The formulas to estimate the precision level for every predicted chemical shift were updated, and confidence intervals measured in ppm were introduced in addition to the abstract trustworthiness values. The algorithm is transparent to the user and allows tracking the generalizations applied (see minor improvements in the Supporting Information) and the reference data used for predictions down to original publications. The approach was implemented in the Glycan-Optimized Dual Empirical Spectrum Simulation (GODESS) web service, which is freely available at the platform of the Carbohydrate Structure Database (CSDB) project (http://csdb.glycoscience.ru).

ASSOCIATED CONTENT

Supporting Information

Minor algorithmic improvements and user interface changes (Section 1, Figures S-1 and S-2); weight optimization procedures and the resultant weight sets (Section 2, Tables S-1−S-4); the number of substituent generalization steps for various prediction modes (Section 3, Table S-5); detailed description of the trustworthiness estimation algorithm and the trustworthiness optimization procedure (Section 4, Figure S-3); sampling characteristics used for the trustworthiness evaluation optimizations, accuracy, and performance validations (Table S-6); the resultant parameters for the trustworthiness calculation (Table S-7); parameters required for calculation of the expected deviation from the trustworthiness (Table S-8); prediction accuracy verification for different residue types (Charts S-1 and S-2); linear correlation coefficients for the trustworthiness verification (Table S-9); explanation of the residue abbreviations (Section 7). The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b01413.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: kapaev_roman@mail.ru (R.R.K).
*E-mail: netbox@toukach.ru (P.V.T.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was funded by the Russian Foundation for Basic Research, grant 15-04-01065. The NMR data obtained using the reported tool can be used free of charge with reference to this paper. The logo for the Carbohydrate Structure Database was created by Philip Toukach, and he retains the copyright for this logo.

REFERENCES

(1) Gaidzik, N.; Westerlind, U.; Kunz, H. *Chem. Soc. Rev.* **2013**, *42*, 4421−4442.
(2) Astronomo, R. D.; Burton, D. R. *Nat. Rev. Drug Discovery* **2010**, *9*, 308−324.
(3) Boltje, T. J.; Buskas, T.; Boons, G.-J. *Nat. Chem.* **2009**, *1*, 611−622.
(4) Johnson, M. A.; Bundle, D. R. *Chem. Soc. Rev.* **2013**, *42*, 4327−4344.
(5) Alper, J. *Science* **2001**, *291*, 2338−2343.
(6) Schmidt, L. D.; Dauenhauer, P. J. *Nature* **2007**, *447*, 914−915.
(7) Toukach, F. V.; Ananikov, V. P. *Chem. Soc. Rev.* **2013**, *42*, 8376−8415.
(8) Rudd, T.; Yates, E.; Hricovíni, M. *Curr. Med. Chem.* **2009**, *16*, 4750−4766.
(9) Lundborg, M.; Widmalm, G. *Anal. Chem.* **2011**, *83*, 1514−1517.
(10) Kapaev, R. R.; Egorova, K. S.; Toukach, P. V. *J. Chem. Inf. Model.* **2014**, *54*, 2594−2611.
(11) Egorova, K.; Toukach, P. *Carbohyd. Res.* **2014**, *389*, 112−114.
(12) Toukach, P.; Egorova, K. In *Glycoscience: Biology and Medicine*; Taniguchi, N., Endo, T., Hart, G. W., Seeberger, P. H., Wong, C.-H., Eds.; Springer: Japan, 2014; pp 241−250.
(13) Gao, W.-f.; Liu, S.-y. *Comput. Oper. Res.* **2012**, *39*, 687−697.
(14) Karaboga, D.; Gorkemli, B.; Ozturk, C.; Karaboga, N. *Artif. Intell. Rev.* **2014**, *42*, 21−57.
(15) Karaboga, D.; Akay, B. *Appl. Math. Comput.* **2009**, *214*, 108−132.
(16) Parra, A.; Veraldi, N.; Locatelli, M.; Fini, M.; Martini, L.; Torri, G.; Sangiorgi, L.; Bisio, A. *Glycobiology* **2012**, *22*, 248−257.
(17) Katzenellenbogen, E.; Toukach, P. V.; Kocharova, N. A.; Korzeniowska-Kowal, A.; Gamian, A.; Shashkov, A. S.; Knirel, Y. A. *FEMS Immunol. Med. Microbiol.* **2008**, *53*, 60−64.

DOI: 10.1021/acs.analchem.5b01413
Anal. Chem. 2015, 87, 7006−7010

(18) Jamróz, M. K.; Paradowska, K.; Zawada, K.; Makarova, K.; Kaźmierski, S.; Wawer, I. *J. Sci. Food Agric.* **2014**, *94*, 246−255.
(19) Schoenhofen, I. C.; McNally, D. J.; Vinogradov, E.; Whitfield, D.; Young, N. M.; Dick, S.; Wakarchuk, W. W.; Brisson, J.-R.; Logan, S. M. *J. Biol. Chem.* **2006**, *281*, 723−732.
(20) Ovchinnikova, O. G.; Arbatsky, N. P.; Chizhov, A. O.; Kocharova, N. A.; Shashkov, A. S.; Rozalski, A.; Knirel, Y. A. *Carbohydr. Res.* **2012**, *349*, 95−102.
(21) Shashkov, A. S.; Potekhina, N. V.; Naumova, I. B.; Evtushenko, L. I.; Widmalm, G. *Eur. J. Biochem.* **1999**, *262*, 688−695.
(22) De Castro, C.; Lanzetta, R.; Leone, S.; Parrilli, M.; Molinaro, A. *Carbohydr. Res.* **2013**, *370*, 9−12.
(23) Tul'skaya, E. M.; Streshinskaya, G. M.; Shashkov, A. S.; Sof'ya, N. S.; Avtukh, A. N.; Baryshnikova, L. M.; Evtushenko, L. I. *Carbohydr. Res.* **2011**, *346*, 2045−2051.
(24) Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J. *J. Chem. Inf. Model.* **2008**, *48*, 128−134.
(25) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. *Gaussian 09*; Gaussian Inc.: Wallingford, CT, 2009.

DOI: 10.1021/acs.analchem.5b01413
Anal. Chem. 2015, 87, 7006−7010