ELSEVIER

www.elsevier.com/locate/carres

Carbohydrate Research 335 (2001) 101–114

CARBOHYDRATE RESEARCH

Computer-assisted structural analysis of regular glycopolymers on the basis of ^{13}C NMR data

Filip V. Toukach,* Alexander S. Shashkov

N.D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky Prospect 47, 117913 Moscow, Russian Federation

Received 6 October 2000; received in revised form 27 June 2001; accepted 20 July 2001

Abstract

A computer-assisted approach to the prediction of the primary structures of regular glycopolymers is described. The analysis is based on comparing the calculated ^{13}C NMR spectra of all the possible structures of the repeating unit (for the given monomeric composition) to an experimental ^{13}C NMR spectrum. The spectra generation is based on the spectral database containing information on the ^{13}C chemical shifts of monomers, di- and trimeric fragments. If the required data are missing from this database, the special database for average glycosylation effects is used. The analysis reveals those structures with the calculated ^{13}C NMR spectrum most close to observed. The structures of repeating units of any topology containing up to six residues linked by glycosidic, amidic or phospho-diester bridges can be predicted. Unambiguous selection of the proper structure from the output list of possible structures may require additional experimental data. Testing the created program and databases on bacterial polysaccharides and their derivatives containing up to three non-sugar residues (alditols, amino acids, phosphate groups etc.) per repeating unit revealed the good convergence of prediction with independently obtained structural data.

Keywords: Glycosylation effect; Database; Chemical shift; Polysaccharide; Structure; NMR; Computer; Prediction

1. Introduction

There has been several programs described for computer-assisted interpretation of ^{13}C NMR spectra, dealing either with oligosaccharides[1] or with regular polymers,[1–3] but all of them require knowledge of the repeating unit topology (often, only linear[4] or linear and single-branched[3,5] topologies are handled). In the present work we have created a computer predictor of the structure of regular glycopolymers built of residues linked by glycosidic, amidic and phospho-diester bridges. The input data are experimental ^{13}C NMR spectrum of the regular polymer and its monomeric composition (including absolute and anomeric configurations); the output is a list of possible structures with repeating unit topology, residue sequence and substitution pattern determined.

2. Experimental

All the chemical shifts were measured with Bruker NMR instruments (WM250, AM300, DRX500) at 318 K for D_2O solutions with acetone as internal standard (δ_C 31.45 ppm).

* Corresponding author. Tel.: +7-095-1359094; fax: +7-095-1355328.

E-mail address: tou@cacr.ioc.ac.ru (F.V. Toukach).

0008-6215/01/$ - see front matter
PII: S0008-6215(01)00214-2

The literature data on chemical shifts were considered as obtained under conditions described in 'Spectral database'.

The program utilized the algorithm described in 'Results and discussion'. The source code was written in Borland Pascal 8 and Turbo Assembler 4 and compiled into 16-bit executable binaries for the real and virtual modes of DOS/Windows-compatible operating systems. The object modules were linked together using Borland Tlink. The program was built as a single-thread console application. It requires the input task in a text file and outputs the results to another text file. All the databases were full-text. The format of databases and of all files that the program operates was documented in a user's manual supplied with distribution.

The program was tested on a PC under MS-DOS 5.0 + and in the DOS-mode of MS Windows 98/NT/2000. Depending on the complexity of the task, it required from 200 to 500 Kb of conventional memory. The amount of available memory did not affect the calculation performance. The program worked with any i80 × 86-compatible CPU higher than i386, but in the case of five or more residues per repeating unit, the execution time was acceptable only with CPUs of Intel Pentium class or higher (see Table 2).

The distribution of a non-commercial version (BIOPSEL v. 2.07) is available at http://nmr.ioc.ac.ru/Staff/Toukach/pv/ps.htm.

3. Results and discussion

Good convergence of predictions with independently obtained structural data was observed for about 70 polysaccharides and their derivatives containing up to two non-pyranose residues (alditols, furanoses, amino acids, phosphate groups etc.) per repeating unit of up to six residues. In 90% of cases, the correct structure was among the five most probable structures predicted; in 70% of these cases it was ranked as the most probable structure. The predictions obtained were used in structural studies of several medically important bacterial O-antigens.

The approach to the structure determination involved three steps:

1. Generation of all the possible structures of the repeating unit for the given monomeric composition (see 'Structure generation' below).
2. Evaluation of the ^{13}C NMR spectrum for each structure generated, based on the spectra of mono-, di- and trimeric fragments and the average substitution effects (see 'Theoretical spectrum calculation' below).
3. Search for the structures with the calculated spectrum closest to that observed for the polymer.

Structure generation.—Based on the entered monomeric composition, the program scanned (see below) all the possible structures of the chemical repeating unit, excluding repetitions and considering the type of linkage that each residue may form. Monomeric residues structural properties are stored in full-text database, initially containing about 100 residues (see Table 1).

This database included information on the following structural properties:

- orientation of protons in pyranose cycles (*ax*/*eq*)
- positions of carbons bearing nitrogen
- positions of deoxygenated carbons
- number of carbons in a cycle
- residue type (pyranose, furanose, alditol, non-carbohydrate)
- positions of carboxyl groups
- linkage type at C-1 (glycosidic, amidic, phospho-diester)

These properties can be inputted or updated using the formal language described in the user's manual supplied with the program.

The first step (the outmost cycle of structure generation) was working out all the possible topologies of the repeating unit (see Fig. 1). The set of topologies was pre-defined for each number of residues. The maximum number of residues per repeating unit is six[†]. The number of topologies varies from one for a linear homopolymer to 22 for the polymer with hexa-residue repeating unit. Some rarely oc-

[†] The commercial version that is expected in 2002 will operate repeating units of up to eight residues.

curring topologies can be excluded from the calculation by the special program key (widespread mode) to speed up the process and to make the interpretation easier.

The second step was generating so called sequences for each topology. The sequence is the combination of monomers considering the order of their connection to each other, but ignoring the substitution positions. Some of the sequences obtained may have no chemically possible substitution pattern and, thereby, they must be skipped. This relates to the following situations:

1. The sequence is identical to the sequence already generated before. This may occur if there are identical monomers in the monomeric composition.
2. The sequence contains residues having not enough free positions for the given substitution type (e.g., bisubstituted Rha2,3-$(NAc)_2$ or an amino acid substituting a residue without free carboxyl or amino groups).
3. The sequence may be obtained by a cyclic permutation of a sequence already generated before. This may occur for topologies tolerating cyclic shifts, i.e. all the linear topologies, –[x–]x–[x–]x–, –[x–]x–[x–]x–[x–]x–, –[x–x–]x–[x–x–]x–, –[x–]x–x–[x–]x–x– (x = any residue, side chains are in square brackets).
4. In widespread mode only: rarely occurring sequences (with more than two residues in side chains per repeating unit; with non-carbohydrate residues in a backbone; with alditols that are not 1-phosphorylated)

The third step was generating all chemically possible substitution patterns for each sequence, considering the type of linkage that each substituting residue forms at C-1. The total number of generated structures for the repeating unit depends on how many and what type of substitutable positions each residue has. The typical values are shown in Table 2. The widespread mode forces the program to analyze only widespread sequences, skipping rarely occurring ones with three or more residues in side chains or with non-sugar residues in the backbone or with non-1-phosphorylated alditols.

Theoretical spectrum calculation.—For each generated structure the theoretical ^{13}C NMR spectrum was calculated and its linear deviation from the experimental spectrum was stored. To calculate the theoretical spectrum, the program breaks the repeating unit into separate residues, then generates subspectra for each one with substitution effects added and combines subspectra into the whole spectrum again. The general scheme for a theoretical spectrum calculation is shown in Fig. 2.

Table 1
Monomeric residues structural properties (see 'Structure generation') database

Hexopyranoses	Glc*p*, Glc*p*NAc, Qui*p*, Qui*p*NAc, Qui*p*3NAc, Qui*p*4NAc, Man*p*, Man*p*NAc, Rha*p*, Rha*p*NAc, Rha*p*2,3$(NAc)_2$, Gal*p*, Gal*p*NAc, Fuc*p*, Fuc*p*NAc, Fuc*p*3NAc, Fuc*p*4NAc, All*p*, All*p*NAc, Alt*p*, Alt*p*NAc, Tal*p*, Tal*p*NAc, 6dTal*p* Gul*p*, Gul*p*NAc, Ido*p*, 6dIdo*p*
Pentapyranoses	Xyl, Sor, Tag, Lyx, Rib, 2dRib, Ara, Psy, Fru
Hexuronic acids	Glc*p*A, Glc*p*NAcA, Man*p*A, Man*p*NAcA, Gal*p*A, Gal*p*NAcA, All*p*A Alt*p*A Tal*p*A, Gul*p*A, Ido*p*A
Hexafuranoses	Gal*f*, Alt*f*, 6dAlt*f*, Tal*f*, 6dTal*f*, 6dIdo*f*
Pentafuranoses	Xyl*f*, Sor*f*, Tag*f*, Rib*f*, Ara*f*, Psy*f*, Fru*f*
Alditols	Gal*o*, Xyl*o*, Rib*o*, Rib*o*-1P, Ara*o*, Ery*o*, Thr*o*, Gr*o*, Gr*o*-1P
Amino acids linked via α-NH_2	Lys, Ala-Lys, Ala, Gly, Val, Leu, Ile, Ser, Thr, Arg, AspA, Asn, GluA, Gln, CysH, Cys, Met, Hys
Rest	KDO, Leg, Neu, Pse, 2-hydroxypriopionic acid, 2-aminoethyl phosphate, alanine-ethanolamine, ethyleneglycol phosphate, choline, 3-hydroxybutyrate

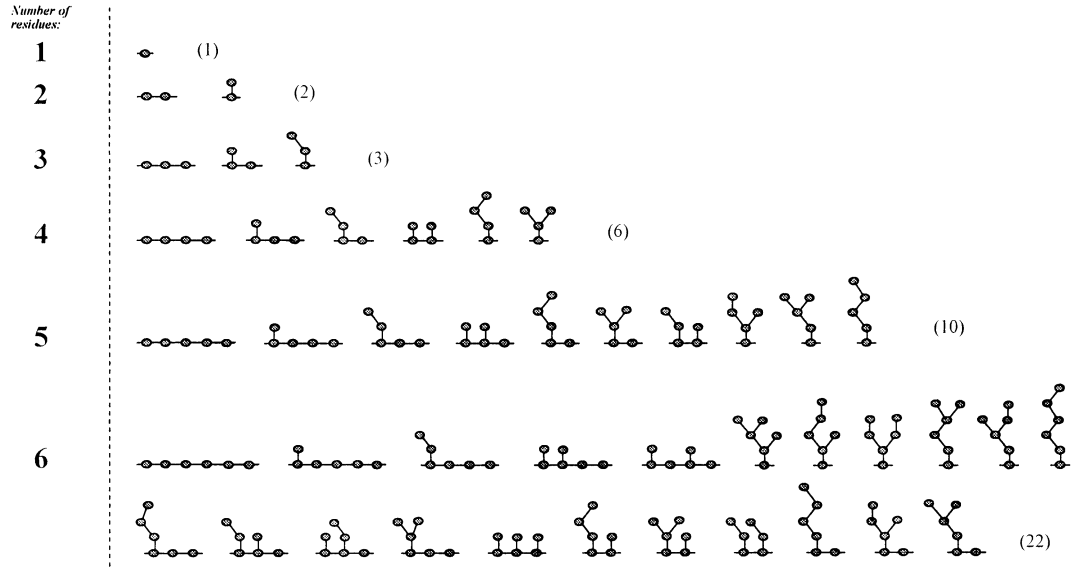

Fig. 1. Possible topology variants.

Table 2
The number of structures generated and calculation time

Number of residues in the repeating unit	Number of simplifications [a]	Number of analyzed structures	Number of possible sequences, not more than	Calculation time, (P-III 600 MHz) min:s
3 or less	0	704 or less	14	<0:0.01
4	0	21.6*K*	126	0:0.2
4+Lys, widespread mode	2 (except Lys)	5.7*K*	96	0:0.1
4+Lys	2 (except Lys)	18.6*K*	1104	0:0.3
5, widespread mode	2	150.8*K*	384	0:2.4
5	0	721.5*K*	1104	0:11.4
6, widespread mode	3	3.21*M*	3000	1:10.6
6	5	4.95*M*	15 240	1:49.1
6	3	14.24*M*	15 240	5:13.4
6	0	35.14*M*	15 240	12:35.8

$K = 1024$; $M = K^2 = 1\,048\,576$.

[a] The simplification is the inability of the residue to be substituted at a certain position. Any simplifications fasten the calculation, as well as presence of phosphate groups, identical residues, alditols and amide bonds, especially in widespread mode.

The most accurate (with accuracy better than ± 0.5 ppm for each signal) subspectrum is obtained if the exact structural fragment is found in the spectral database, otherwise the program applies average substitution effects to the spectrum of the terminal residue. The calculation of a subspectrum implied the following sequence of procedures:

1. The spectrum of the unsubstituted (terminal) residue is searched for in the spectral database.
2. In the case that all the residue's substituents are monosaccharides or phosphate groups, the same part of the spectral database is scanned for the chemical shifts of the residue substituted. If the residue is bisubstituted at neighboring positions, the theoretical calculation of glycosylation effects is not additive due to the spatially caused conformational changes in the trisaccharide unit.[6] In this case, the program tries to find the spectral data for similar fragments varying the type (–OH/–NH_2) and orientation (*ax*/*eq*) of the group at C-2 of both substituting residues. Each of two substituting residues may be of varied type, or orientation, or both, that forms $3 \times 3 = 9$ kinds of similar structural

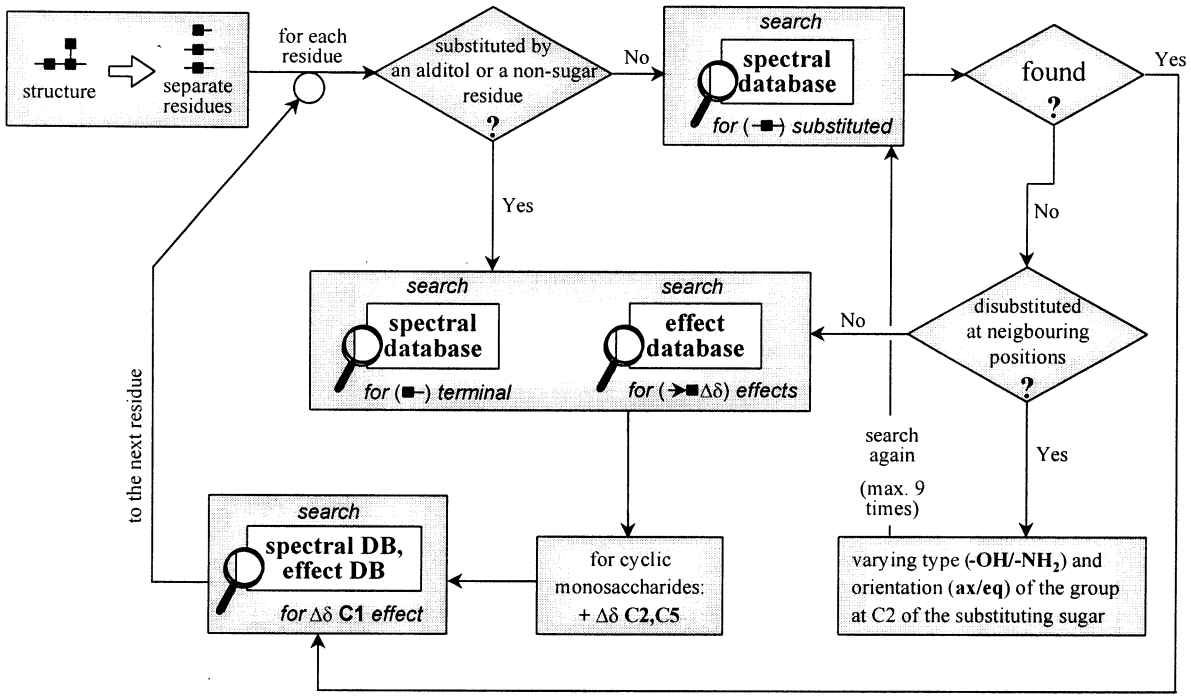

Fig. 2. Theoretical ^{13}C NMR spectrum generation

fragments. If none of these is described in the spectral database, the program considers the chemical shifts search failed and uses theoretical effects in additive approximation, i.e., it adds theoretical effects introduced by the first and second substituting residues independently.

3. If the search of chemical shifts fails or the residue is substituted by at least one alditol or non-carbohydrate residue, the program generates the subspectrum for the analyzed residue using data for the unsubstituted monomer and average substitution effects. The chemical shift of each signal within the subspectrum of the analyzed residue is added to a value read from the database dependent on the substituting residue and type of substitution (see 'Substitution effect database' below).
4. The effects for C-1, C-2 and C-5 were added to the subspectrum of the residue in the manner described in Ref. 1. The effects on C-2 and C-5 depend mainly on anomeric configuration of residues,[1] and effects on C-1 are taken from the spectral database.

The spectral database.—The spectral database was composed of literature data (Refs. 1,7–10,15 for the unsubstituted residues, a variety of available ^{13}C NMR data for bacterial and plant polysaccharides and derivatives for substituted residues)[‡] and data from model carbohydrates. Literature data were used if the ^{13}C NMR spectra were recorded under the following conditions: temperature range of 318 ± 10 K, TMS (δ_C 0.0) or acetone (δ_C 31.45) as a reference. Carbon chemical shifts do not vary much ($\Delta\delta_C < 1.0$ for carbons forming glycosidic bridges, $\Delta\delta_C < 0.3$ for other carbons) in the temperature range mentioned and, considering the accuracy of predictions, these deviations are acceptable. In the case of unavailable or suspicious references or of other references than acetone or TMS, chemical shifts scales were adjusted using characteristic carbohydrate signals (C-6 of terminal or 2- or 3-substituted Glc*p* at δ 61.9; C-6 of terminal or 2- or 3-substituted Rha*p* at δ 18.0; –NH–CO–CH_3 at δ 23.3[§]).

The prediction accuracy depended strongly on the completeness of the corresponding part of the spectral database. The database pro-

[‡] Each publication was the source for describing from 1 to 10 variants of substitution. As there are more than 2200 structural fragments described in the spectral database, all the publications can not be referenced within this document.

[§] Lower field signal if several NAc groups present.

Table 3

Configuration sugar type	Gluco	Galacto	Manno	Allo	Altro	Gulo	Talo	Ido
Sugar residues described in the spectral database [a]								
Basic pyranoses	Glc*p* **174** (α,β)	Gal*p* **505** (α,β)	Man*p* **128** (α,β)	All*p* **2** (α,β)	Alt*p* **3** (α,β)	Gul*p* **2** (α,β)	Tal*p* **2** (α,β)	Ido*p* **2** (α,β)
Acetamidic derivatives	Glc*p*NAc **203** (α,β) Glc*p*2,3(NAc)$_2$ **2** (α,β)	Gal*p*NAc **148** (α,β) Gal*p*2,3(NAc)$_2$ **2** (α,β)	Man*p*NAc **75** (α,β) Man*p*2,3(NAc)$_2$ **2** (α,β)					
6-Deoxy-pyranoses	Qui*p* **2** (α,β)	Fuc*p* **75** (α,β)	Rha*p* **526** (α,β)				6dTal*p* **6** (α,β)	6dIdo*p* **2** (α,β)
6-deoxy-pyranoses, acetamidic derivatives	Qui*p*NAc **51** (α,β) Qui*p*3NAc **3** (α,β) βQui*p*4NAc **1**	Fuc*p*NAc **66** (α,β) Fuc*p*3NAc **2** (α,β) αFuc*p*4NAc **2**	Rha*p*2,3(NAc)$_2$ **2** (α,β)					
Hexuronic acids	Glc*p*A **75** (α,β)	Gal*p*A **74** (α,β) Gal*fp*A **2** (α,β)	Man*p*A **2** (α,β) βMan*p*ANAc **3**		Alt*p*A **2** (α,β)			
Hexo-furanoses		βGal*f* **2**			Alt*f* **2** (α,β) 6dAlt*f* **2** (α,β)		Tal*f* **2** (α,β) 6dTal*f* **2** (α,β)	6dIdo*f* **2** (α,β)
Penta-pyranoses	Xyl*p* **2** (α,β) αSor*p* **1**	Ara*p* **2** (α,β)	Lyc*p* **2** (α,β) Tag*p* **2** (α,β)	Rib*p* **2** (α,β) Psy*p* **2** (α,β)	Fru*p* **2** (α,β)			
Penta-furanoses	Xyl*f* **2** (α,β) αSor*f* **1**	Ara*f* **4** (α,β)	βTag*f* **1**	Rib*f* **3** (α,β) Psy*f* **2** (α,β)	Fru*f* **2** (α,β)			
Non-sugar residues described in the spectral database [b]								
Alditols	Gr*o* (**9**), Rib*o* (**2**), Gal*o* (**2**), Gal*o*NAc (**1**), Glc*o* (**1**), Ara*o* (**1**), Xyl*o* (**1**), Ery*o* (**1**), Thr*o* (**1**)							
Amino acids	20 natural amino acids, 1 dataline for each (**20**)							
Phosphates	Gr*o*-1P (**14**), Rib*o*-1P (**3**), 2-aminoethanol-1-phosphate (**1**), ethylenglycol-1-phosphate (**1**)							
Sugar-derived acids	αKDO (**2**), βKDO (**1**), βLeg (**1**), αNeu (**1**), βNeu (**1**), αPse (**1**)							
Rest	alanino-lysine (**1**), alanino-ethanolamine (**1**), choline (**1**), 2-hydroxypriopionic acid (**1**), 3-oxy-butyrate (**1**)							

[a] The number of the substitution variants described is given below residue names.
[b] The number of the substitution variants described is in parentheses.

vided with the program distribution contains spectral characteristics for about 80 most common residues listed in Table 3. As *O*-acetyl groups are attached to the polymer chain, in many cases non-stoichiometrically, all the residues were described as non-O-acetylated and, therefore, de-O-acetylation has to be carried out before the computer analysis. The positions of *O*-acetyl groups were then revealed by comparison of the ^{13}C NMR spectrum of the native polymer to that of de-O-acetylated assigned with the help of the program.

Each entry in the spectral database implies the following information for each structural fragment described:

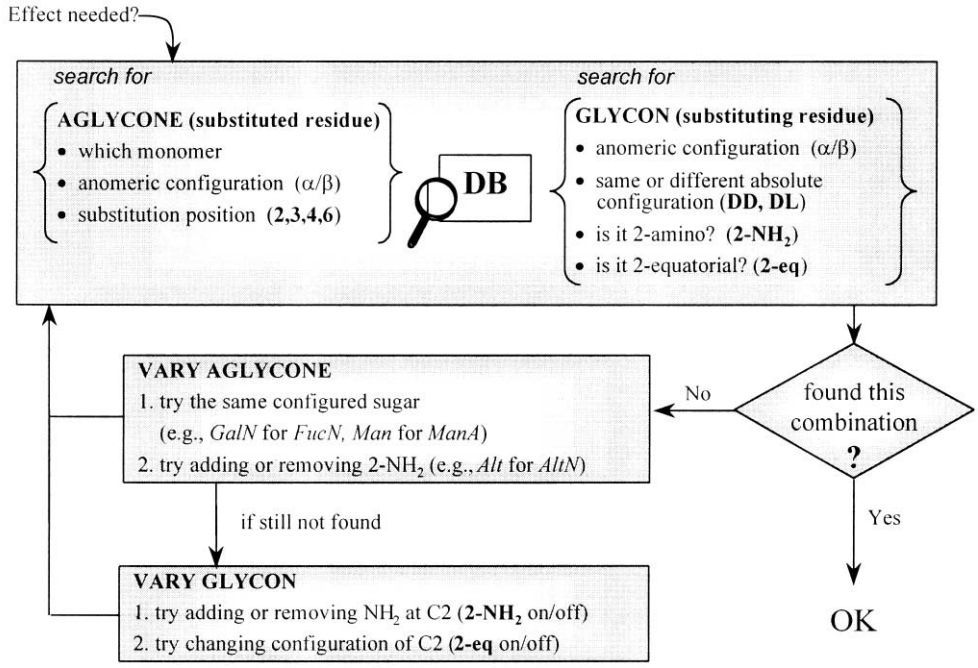

Fig. 3. Searching the effect database

1. Properties of the main residue in the fragment (monomer for the monomeric fragment, substituted residue for the dimeric fragment, residue at the branching point for the trimeric fragment):
   - which residue (the reference to the entry in the residues database, see Table 1)
   - its anomeric configuration, if it may have anomeric forms
   - the presence or absence of the phosphate group at C-1
2. Properties of the first substituting residue (for di- and trimeric fragments only):
   - the position of the main residue to which this residue is attached
   - anomeric configuration, if it may have anomeric forms
   - the presence or absence of the phosphate group at C-1
   - the type of substituent at C-2 (–OH/–NH_2)
   - the orientation of substituent at C-2 (*ax*/*eq*) (for sugar substituting residues only)
   - the combination of absolute configurations of the first substituting residue and the main residue (DD/DL)
3. Properties of the second substituting residue (for trimeric fragments only):
   - the same properties as for the first substituting residue.

The substitution effect database.—This database was filled in by literature data on glycosylation and phosphorylation effects,[1] calculated as the difference between chemical shifts of carbons in substituted and unsubstituted residues. The database contained information on about 100 variants of substitution. The way to find the certain effect is demonstrated in Fig. 3.

The effect database is full-text, with each line describing the following properties of substitution:

- which residue was substituted
- aglycone anomeric configuration (α/β; for substitution at C-2 only)
- substitution position (from C-2 to C-6)
- type of linkage (α-glycosidic/β-glycosidic/phospho-diester/amidic/substitution by alditol)
- are the absolute configurations of linked residues the same or different? (DD/DL)
- is the substituting residue 2-amino? (yes/no)
- the orientation of substituent at C-2 of the substituting residue (*ax*/*eq*; for sugar substituting residues only).

If the desired effect is not described in the database, the database engine tries adding or removing an amino group at C-2 of the substituted sugar, or searching for another sugar with the same configuration, e.g. GalNAc in-

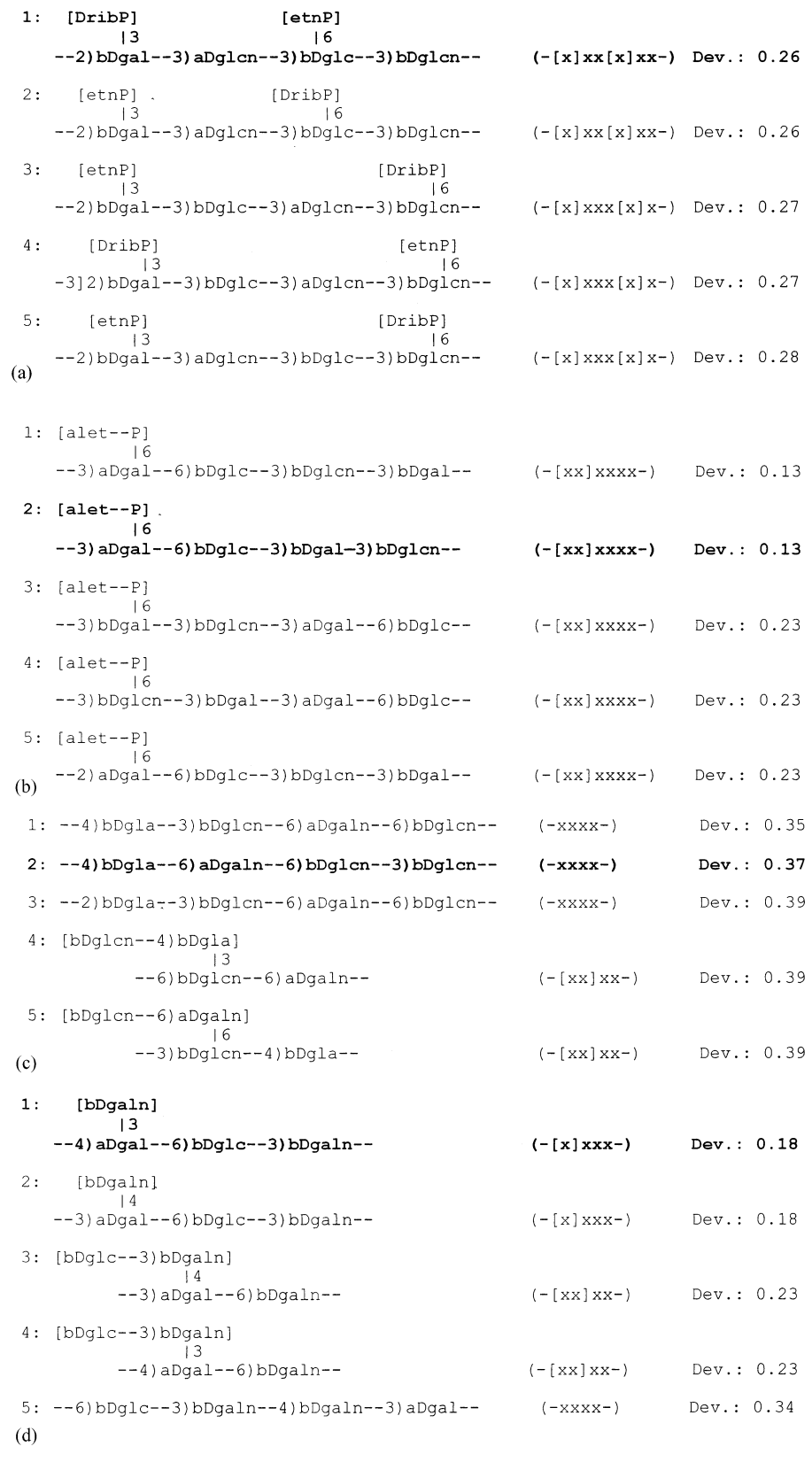

Fig. 4. (a) Program output for *P. mirabilis* O33; (b) *P. mirabilis* O3; (c) *P. mirabilis* O30; (d) *P. penneri* 1; (e) *P. penneri* 2; (f) *P. mirabilis* O16; (g) *P. mirabilis* O14; (h) *P. vulgaris* O44; (i) *P. penneri* 71; (j) *P. vulgaris* O17; (k) *E. coli* O28.

```
1:   [bDr23n]
       |3
  --4)aDglc--3)aLta6d--3)bDglcn--              (-[x]xxx-)    Dev.: 0.40

2:   [bDr23n]
       |4
  --3)aDglc--3)aLta6d--3)bDglcn--              (-[x]xxx-)    Dev.: 0.44

3:   [bDr23n]
       |3
  --4)aDglc--3)aLta6d--4)bDglcn--              (-[x]xxx-)    Dev.: 0.44

4:   [bDr23n--3)bDglcn]
                |4
           --3)aDglc--3)aLta6d--               (-[xx]xx-)    Dev.: 0.45

5:   [bDr23n]
       |4
  --3)aDglc--3)aLta6d--4)bDglcn--              (-[x]xxx-)    Dev.: 0.45
```

(e)

```
1:  [etnP]
       |6
  --3)aDglcn--4)DribP--6)bDgaln--4)aDgaln--    (-[x]xxxx-)   Dev.: 0.23

2:  [etnP]
       |6
  --4)aDglcn--4)DribP--6)bDgaln--4)aDgaln--    (-[x]xxxx-)   Dev.: 0.26

3: [etnP--6)bDgaln]
              |6
         --3)aDglcn--4)aDgaln--4)DribP--       (-[xx]xxx-)   Dev.: 0.27

4:  [etnP]
      |3
  --6)aDglcn--4)aDgaln--4)DribP--6)bDgaln--    (-[x]xxxx-)   Dev.: 0.27

5:  [etnP]
       |3
  --6)aDglcn--6)bDgaln--4)aDgaln--4)DribP--    (-[x]xxxx-)   Dev.: 0.28
```

(f)

```
1: [alet--P]
       |6
    --4)aDgal--3)bDgaln--3)aDgal--4)bDgaln--   (-[xx]xxxx-)   Dev.: 0.20

2: [alet--P]
       |6
    --4)aDgal--4)bDgaln--3)aDgal--3)bDgaln--   (-[xx]xxxx-)   Dev.: 0.20

3: [alet--P]
       |6
    --4)aDgal--3)bDgaln--4)aDgal--4)bDgaln--   (-[xx]xxxx-)   Dev.: 0.23

4: [alet--P]
       |6
    --4)aDgal--4)bDgaln--4)aDgal--3)bDgaln--   (-[xx]xxxx-)   Dev.: 0.23

5: [alet--P]
       |6
    --4)aDgal--3)bDgaln--2)aDgal--4)bDgaln--   (-[xx]xxxx-)   Dev.: 0.24
```

(g)

```
1:  [Lala]
      |6
 --4)bDgla--3)bDgaln--4)bDglc--3)aDgal--4)bDgaln--   (-[x]xxxxx-) Dev.: 0.32

2: [Lala]
      |6
 --4)bDgla--3)aDgal--4)bDgaln--4)bDglc--3)bDgaln--   (-[x]xxxxx-) Dev.: 0.32

3: [Lala]
      |6
 --4)bDgla--4)bDglc--3)bDgaln--3)aDgal--4)bDgaln--   (-[x]xxxxx-) Dev.: 0.33

4: [Lala]
      |6
 --4)bDgla--3)bDgaln--3)aDgal--4)bDgaln--4)bDglc--   (-[x]xxxxx-) Dev.: 0.33

5: [Lala]
      |6
 --4)bDgla--3)bDgaln--4)bDglc--2)aDgal--4)bDgaln--   (-[x]xxxxx-) Dev.: 0.34
```

(h)

Fig. 4. (*Continued*)

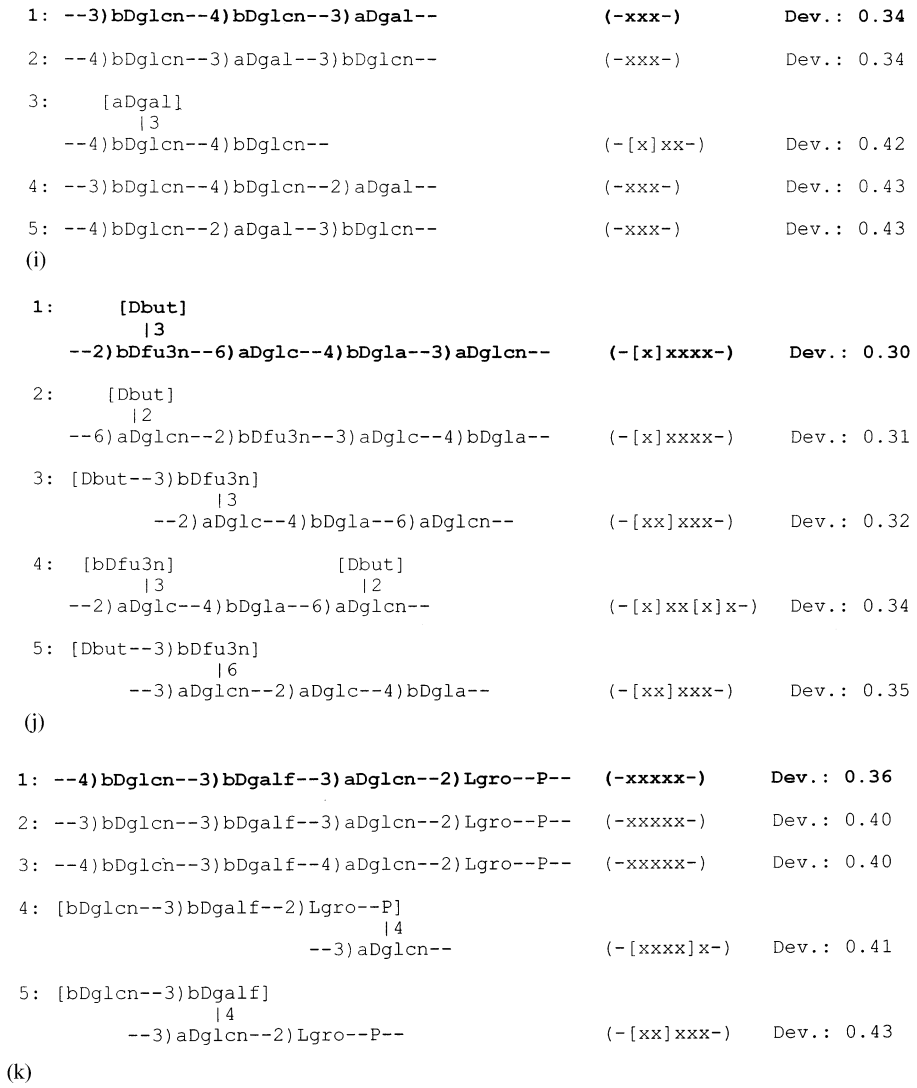

Fig. 4. (*Continued*)

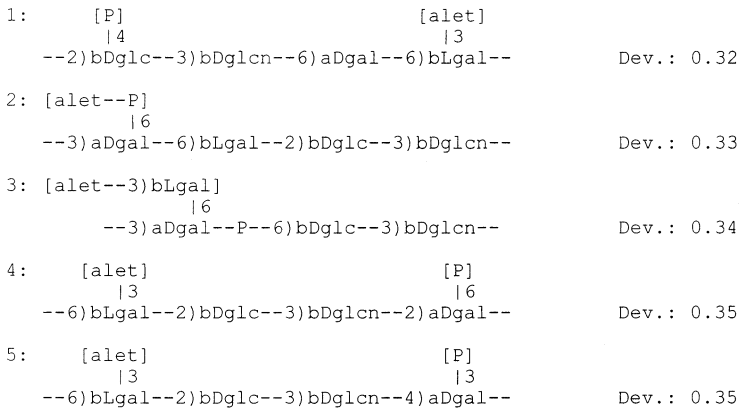

Fig. 5. Program output for *P. mirabilis* O3 with improperly specified absolute configuration of βGal*p* residue.

stead of FucNAc. If these effects were not described either, the database engine tried to vary the type and orientation of the group at C-2 of the substituting residue.

The prediction of the structure of polymers built of widespread carbohydrate residues was most accurate because glycosylation effects for three widespread sugar configurations (glc,

Table 4
Prediction accuracy

Prediction accuracy	Number of structures that were predicted with the described accuracy, in percentage of total number of structures used for verification (%)
The proper structure was ranked as most probable structure	60
The proper structure was among five most probable structures predicted (its rank was from 2 to 5)	20
The proper structure was not among five most probable structures predicted but its rank became from 2 to 5 after updating databases with data for specific residues or linkage types	15
The proper structure could not be predicted at all (its rank was 20 or more)	5

gal, man) were represented in the database most completely, i.e. there were effects for all the possible types of substitution for these sugar configurations. In 90% cases, if there were not more than one non-carbohydrate or rarely occurring residue per repeating unit, the proper structure was found among five most probable structures predicted.

Structure ranking.—All the generated structures were ranked linearly by the value of deviation (see below) between the calculated and experimental spectrum. The user-defined number of structures that possessed lower deviation values was selected from all the possible structures to form the output list.

The number of signals in calculated spectrum was determined by monomeric composition only and due to this it stayed constant within the structure generation process. The procedure of comparison of the calculated spectrum for each structure to the experimental spectrum sorted the theoretical spectrum in chemical shift ascending order, then compared chemical shifts by pairs and averaged the difference to obtain the deviation value. Signals with multiple integral intensities were treated as several signals of unit integral intensity. Accordingly, all the signals in the experimental spectrum were considered of equal integral intensity and so signals with double or triple intensity had to be entered two or more times.

Verification.—The created program was tested on bacterial polysaccharides, mainly of the medically important genus *Proteus*. Running on a personal computer (P-III 600 MHz, 100 MHz system bus), the program made predictions in 70% cases in less than 0.5 min for pentasaccharide repeating units and in less than 15 min for hexasaccharide units (see Table 2 for details). The ability of the program to predict structures was verified on about 60 glycopolymers from bacterial strains. The average accuracy of predictions is summarized in Table 4. Examples of program output for several glycopolymers of *Proteus* bacteria, including those that have not been studied before, are given in Table 5.

Conclusions.—The potential of the created program to predict the structure of regular glycopolymers was demonstrated on a variety of polymers with independently determined structure. The strong point of the program is that the only input data are the experimental ^{13}C NMR chemical shifts, monomeric composition, anomeric[¶] and absolute configurations of residues, and the output structures are selected from all possible ones for the repeating unit, including all linear and branched topologies** and all possible substitution patterns. The databases for the structural properties of monomeric residues currently contain data on monosaccharides, sugar phosphates, alditols, amino acids and other residues and can be easily updated by user, as well as databases for chemical shifts and substitution effects.

[¶] The commercial version that is expected in 2002 will not require anomeric configurations as obligatory input.

** Except trisubstituted branching points.

The latter databases are presently being expanded on the basis of literature data for di- and trimeric fragments.

The future aim is to exclude input of anomeric and absolute configurations and extend the number of residues per repeating unit up to eight (currently six in the non-commercial version), and number of carbons per residue up to 12 (currently 6–9 in the non-commercial version).

Table 5
Verification

Bacterial strain	Program input	Proper structure and program output	Ref.	Additional experiments
P.m. O33 (*P.m.* D52)	DribP bDgal aDglcn etnP bDglc bDglcn *Wide-spread =OFF*	**Rbo-*P*** **1** ↓ **3** **→2)-β-D-Gal*p*-(1→3)-α-D-Glc*p*NAc-(1→3)-β-D-Glc*p*-(1→3)-β-D-Glc*p*NAc-(1→** Program output is on *Fig. 4a.*	[11]	{^{1}H-^{31}P} HMQC + NOESY (ROESY) (1 *vs.* 4)
P.m. O3	bDgal bDglcn alet P aDgal bDglc *Wide-spread =ON*	**AlaEtn-*P*** **1** ↓ **6** **→3)-β-D-Gal*p*-(1→3)-β-D-Glc*p*NAc-(1→3)-α-D-Gal*p*-(1→6)-β-D-Glc*p*-(1→,** **(*AlaEtn* = *2-[R-1-carboxyethylamino]-ethanol)*** Program output is on *Fig. 4b.*	[12]	NOESY (ROESY) (1 *vs.* 2) (3), (4), (5) posses much greater deviation
P.m. O30	bDgla aDgaln bDglcn bDglcn *Wide-spread =OFF*	**→4)-β-D-Glc*p*A-(1→6)-α-D-Gal*p*NAc-(1→6)-β-D-Glc*p*NAc-(1→3)-β-D-Glc*p*NAc-(1→** Program output is on *Fig. 4c.*	[13]	HMQC-TOCSY or methylation analysis (1,2 *vs.* 3,4,5) + NOESY (ROESY) (1 *vs.* 2)
P.p. 1	bDglc bDgaln bDgaln aDgal *Wide-spread =OFF*	**β-D-Gal*p*NAc** **1** ↓ **3** **→6)-β-D-Glc*p*-(1→3)-β-D-Gal*p*NAc-(1→4)-α-D-Gal*p*-(1→** Program output is on *Fig. 4d.*	[14]	NOESY (ROESY)
P.p. 2	bDr23n aDglc aLta6d bDglcn *Wide-spread =ON*	**β-L-Rha*p*2NAc3NAc** **1** ↓ **3** **→4)-α-D-Glc*p*-(1→3)-α-L-6dTal*p*-(1→3)-β-D-Glc*p*NAc-(1→** Program output is on *Fig. 4e.*	[15]	HMQC-TOCSY or methylation analysis (1,2,4 *vs.* 3,5) + NOESY (ROESY) (3 *vs.* 2,4)
P.m. O16	bDgaln aDgaln etnP aDglcn DribP *Wide-spread =ON*	***P*EtN** | **6** **→4)-α-D-Gal*p*NAc-(1→3)-α-D-Glc*p*NAc-(1→4)-Rb*o*-(1-*P*-6)-β-D-Gal*p*NAc-(1→** Program output is on *Fig. 4f.*	[16]	{^{1}H-^{31}P} HMQC

Table 5 (*Continued*)

P.m. O14[a]	bDgaln aDgal bDgaln alet P aDgal *Wide-spread =ON*	**AlaEtn-*P*** **1** ↓ **6** **→4)-β-D-Gal*p*NAc-(1→3)-α-D-Gal*p*-(1→3)-β-D-Gal*p*NAc-(1→4)-α-D-Gal*p*-(1→,** *(AlaEtn = 2-[R-1-carboxyethylamino]-ethanol)* Program output is on *Fig. 4g.*	[17]	NOESY (ROESY) or HMBC
P.v. O44	Lala bDgla bDgaln bDglc aDgal bDgaln *Wide-spread =ON*	**→3)-α-D-Gal*p*** **L-Ala** **1** **2** ↓ | **4** **6** **β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*-(1→** Program output is on *Fig. 4h.*	[18]	NOESY (ROESY) or HMBC
P.p. 71	bDglcn bDglcn aDgal *Wide-spread =OFF*	**→3)-β-D-Glc*p*NAc-(1→4)-β-D-Glc*p*NAc-(1→3)-β-D-Gal*p*-(1→** Program output is on *Fig. 4i.*	[19]	NOESY (ROESY) or HMBC (1 *vs.* 2) (3), (4), (5) posses much greater deviation
P.v. O17[b]	aDglc aDglcn bDfu3n bDgla Dbut *Wide-spread =ON*	**→3)-α-D-Glc*p*NAc-(1→2)-β-D-Fuc*p*3NBut-(1→6)-α-D-Glc*p*-(1→4)-β-D-Glc*p*A-(1→,** *But=-CO-CH$_2$-CH(OH)-CH$_3$* Program output is on *Fig. 4j.*	[20]	HMQC-TOCSY (Fuc3N C-3 displacement) (1, 3, 5) *vs.* (2, 4) + methylation analysis or HMQC-TOCSY (1 *vs.* 3,5)
E.coli O28	Dgro P bDglcn bDgalf aDglcn *Wide-spread =ON*	**→2)-Gr*o*(1-*P*-4)-β-D-Glc*p*NAc-(1→3)-β-D-Gal*f*(1→3)-α-D-Glc*p*NAc-(1→** Program output is on *Fig. 4k.*	[21]	HMQC-TOCSY or methylation analysis

[a] De-O-acetylated.
[b] De-O-phosphorylated.

Bacterial strains (column 1) includes the following species abbreviations: *P.v.* = *Proteus vulgaris*, *P.p.* = *Proteus penneri*, *P.m.* = *Proteus mirabilis*, *E.c.* = *Eschericia coli*. *E. coli* O28 glycopolymer was included to verify the ability of the program to work with furanose-containing glycopolymers.

Program input column (column 2) includes monomeric composition and widespread mode either ON or OFF. The experimental spectrum is omitted. The first character in each line is the anomeric configuration, the second is the absolute configuration, rest form the residue name. All the residues were of D configuration in all the reported polymers used for verifications. Changing absolute configuration of any input residue leads to the significant increase of deviations for the structures generated, e.g., specifying (b-**L**-gal b-D-glcn D-alet P a-D-gal b-D-glc) as monomeric composition instead of (b-**D**-gal b-D-glcn D-alet P a-D-gal b-D-glc) leads to the list of five best structures for the polysaccharide of *P. mirabilis* O3 shown in Fig. 5, as compared to the primary (all monosaccharides have D configuration) list shown in Fig. 4(b). As it can be seen from these two lists of structures, the deviation of the best-fitting structure increased from 0.13 to 0.32 after changing the absolute configuration of one of residues to the opposite value. Thus, the program determines absolute configurations of monosaccharides basing on the thorough differences in glycosylation effects in DD and DL pairs of monomers.

Program output (column 3, Fig. 4) includes the list of five (may be from 1 to 100) structures most closely fitting the experimental spectrum. The proper structure is given in bold. Each line contains also schematic representation of topology (e.g. –[x]xxx–) and the deviation of the calculated ^{13}C NMR spectrum from the experimental.

References (column4) are to publications with these structures determined independently.

Additional experiments (column 5) column lists the additional NMR and chemical methods that may be applied to easily select the proper structure from the list generated. If a particular experiment is needed to distinguish two certain structures, their numbers are presented in parentheses, e.g. (1 vs. 4).

Acknowledgements

This work was supported by Russian Federation for Basic Research, grants N96-04-50460, N96-15-97380 and N99-04-48279.

References

1. Lipkind, G. M.; Shashkov, A. S.; Knirel, Y. A.; Vinogradov, E. V.; Kochetkov, N. K. *Carbohydr. Res.* **1988**, *175*, 59–75.
2. Cumming, D. A.; Hellerqvist, C. G.; Touster, O. *Carbohydr. Res.* **1988**, *179*, 369–380.
3. Jansson, P.-E.; Kenne, L.; Widmalm, G. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 508–516.
4. Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1987**, *168*, 67–77.
5. Jansson, P.-E.; Kenne, L.; Widmalm, G. *Carbohydr. Res.* **1989**, *188*, 169–191.
6. Bradbury, J. H.; Genkins, G. A. *Carbohydr. Res.* **1984**, *126*, 125–156.
7. Bock, K.; Pedersen, C. *Adv. Carbohydr. Chem. Biochem.* **1983**, *41*, 27–66.
8. L'vov, V. L.; Tochtamysheva, N. V.; Shashkov, A. S.; Dmitriev, B. A.; Capek, K. *Carbohydr. Res.* **1983**, *112*, 233–239.
9. Knirel, Y. A.; Vinogradov, E. V.; Shashkov, A. S.; Dmitriev, B. A.; Kochetkov, N. K.; Stanislavsky, E. S.; Mashilova, G. M. *Eur. J. Biochem.* **1987**, *163*, 627–637.
10. Knirel, Y. A.; Paramonov, N. A.; Shashkov, A. S.; Kochetkov, N. K.; Yarullin, R. G.; Farber, S. M.; Efremenko, V. I. *Carbohydr. Res.* **1992**, *233*, 185–193.
11. Zych, K.; Sidorczyk, Z.; Arbatsky, N. P.; Toukach, F. V.; Shashkov, A. S.; Knirel, Y. A. *Eur. J. Biochem.* **2001**, *268*, 1–7.
12. Vinogradov, E. V.; Kaca, W.; Shashkov, A. S.; Kraewska-Pietrasik, D.; Knirel, Y. A.; Kochetkov, N. K. *Eur. J. Biochem.* **1990**, *188*, 645–651.
13. Shashkov, A. S.; Toukach, F. V.; Paramonov, N. A.; Ziolkowski, A.; Senchenkova, S. N.; Kaca, W.; Knirel, Y. A. *FEBS Lett.* **1996**, *386*, 247–251.
14. Authors' unpublished data
15. Arbatsky, N. P.; Shashkov, A. S.; Toukach, F. V.; Moll, H.; Zych, K.; Knirel, Y. A.; Zahringer, U.; Sidorczyk, Z. *Eur. J. Biochem.* **1999**, *261*, 392–397.
16. Toukach, F. V.; Arbatsky, N. P.; Shashkov, A. S.; Knirel, Y. A.; Zych, K.; Sidorchyk, Z. *Carbohydrate Res.* **2001**, *331*, 213–218.
17. Perepelov, A. V.; Ujazda, E.; Senchenkova, S. N.; Shashkov, A. S.; Kaca, W.; Knirel, Y. A. *Eur. J. Biochem.* **1999**, *261*, 347–353.
18. Authors' unpublished data.
19. Zych, K.; Kocharova, N. A.; Kowalczyk, M.; Toukach, F. V.; Kaminska, D.; Shashkov, A. S.; Knirel, Y. A.; Sidorczyk, Z. *Eur. J. Biochem.* **2000**, *267*, 808–814.
20. Authors' unpublished data.
21. Rundlof, T.; Weintraub, A.; Widmalm, G. *Carbohydr. Res.* **1996**, *291*, 127–139.