Supporting Information for Publication

RESTLESS: Automated Translation of Glycan Sequences from Residue-Based Notation to SMILES and Atomic Coordinates

Ivan Yu. Chernyshov,[†] Philip V. Toukach*[‡]*

[†]All-Russia Research Institute of Agricultural Biotechnology, Russian Academy of Sciences,
Timiryazevskaya st., 42, Moscow 127550, Russia

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

E-mail: ivan-chernyshoff@yandex.ru or netbox@toukach.ru

1. Description of REStLESS algorithm

Figure S1 illustrates the translation from the CSDB Linear code to SMILES.

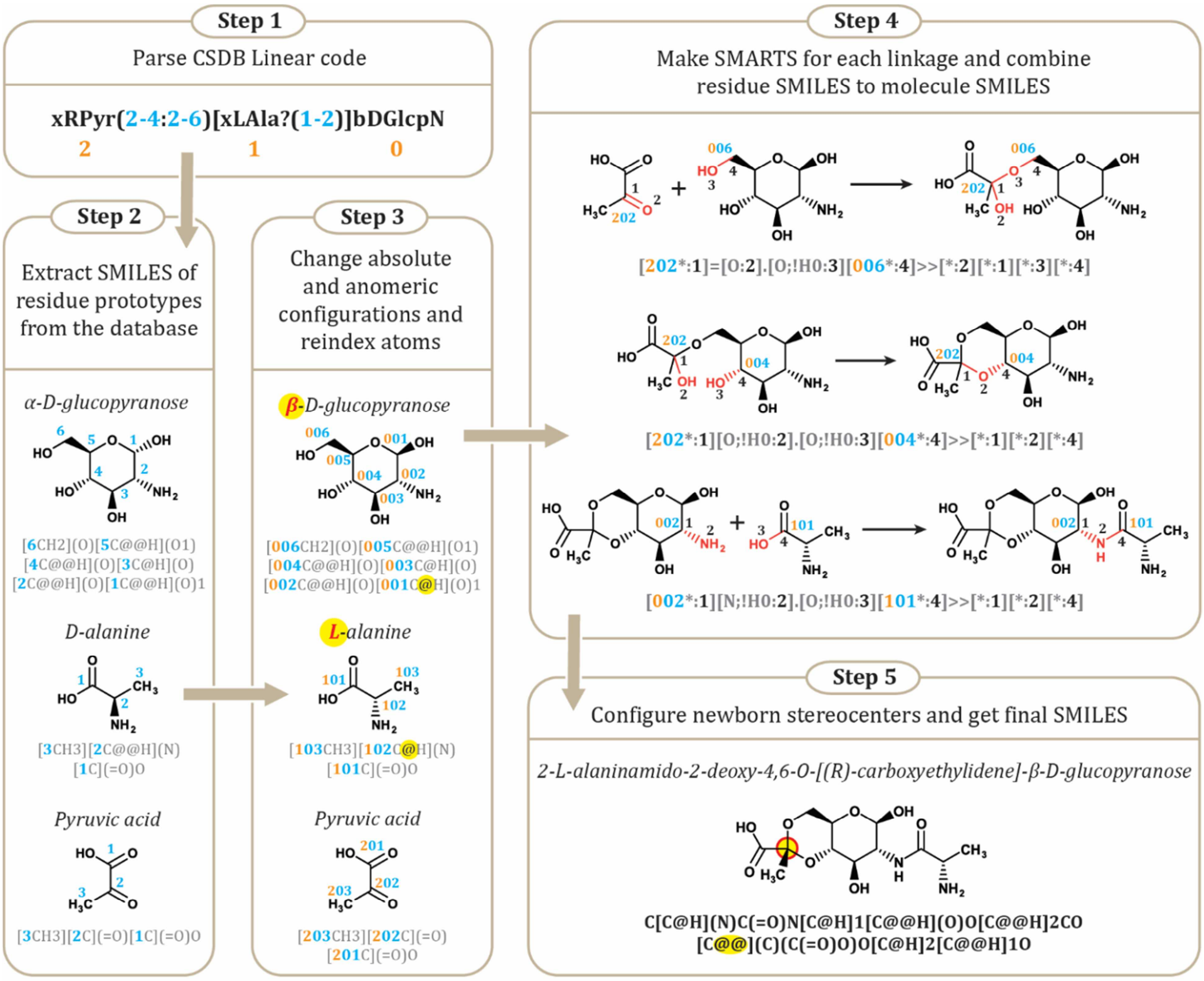
Step 1
Parse CSDB Linear code
xRPyr(2-4:2-6)[xLAla?(1-2)]bDGlcpN
2
1
0
Step 2
Extract SMILES of residue prototypes from the database
α-D-glucopyranose
[6CH2](O)[5C@@H](O1)
[4C@@H](O)[3C@H](O)
[2C@@H](O)[1C@@H](O)1
D-alanine
[3CH3][2C@@H](N)
[1C](=O)O
Pyruvic acid
[3CH3][2C](=O)[1C](=O)O
Step 3
Change absolute and anomeric configurations and reindex atoms
β-D-glucopyranose
[006CH2](O)[005C@@H](O1)
[004C@@H](O)[003C@H](O)
[002C@@H](O)[001C@H](O)1
L-alanine
[103CH3][102C@H](N)
[101C](=O)O
Pyruvic acid
[203CH3][202C](=O)
[201C](=O)O
Step 4
Make SMARTS for each linkage and combine residue SMILES to molecule SMILES
[202*:1]=[O:2].[O;!H0:3][006*:4]>>[*:2][*:1][*:3][*:4]
[202*:1][O;!H0:2].[O;!H0:3][004*:4]>>[*:1][*:2][*:4]
[002*:1][N;!H0:2].[O;!H0:3][101*:4]>>[*:1][*:2][*:4]
Step 5
Configure newborn stereocenters and get final SMILES
2-L-alaninamido-2-deoxy-4,6-O-[(R)-carboxyethylidene]-β-D-glucopyranose
C[C@H](N)C(=O)N[C@H]1[C@@H](O)O[C@@H]2CO
[C@@](C)(C(=O)O)O[C@H]2[C@@H]1O

Figure S1. Algorithm of translation from CSDB Linear to SMILES exemplified on a hypothetic 2-L-alaninamido-2-deoxy-4,6-O-[(R)-carboxyethylidene]-β-D-glucopyranose. Orange, cyan and black digits stand for residue indices in the parsed structure, atom numbering within residues, and atom mapping of reactions, respectively. Captions under molecules and reactions are isomeric SMILES strings and reaction SMARTS expressions, respectively.

SMILES strings were preliminary deposited in a database for each combination of the base name, ring size and modifiers, which fully define the molecular connectivity, giving a cache of 941 SMILES-encoded residues in a free (isolated) form. According to the Carbohydrate Structure Database (CSDB) content analysis, this list of supported residues and their configurations (see a monomer namespace subdatabase [1] for details) covers virtually all carbohydrate and non-carbohydrate constituents present in natural glycans, glycopolymers and glycoconjugates [2]. SMILES strings for combinations of absolute and anomeric configurations other than stored in the residue cache, are obtained by inversion of stereochemical configuration of all atoms or only the anomeric atom, respectively. The SMILES strings of residues are concatenated into the SMILES code of a target molecule using the RDKit [3] implementation of SMARTS reaction expressions [4]. Carbon atoms within the residues are enumerated using isotopic specification in order to link specific positions of the residues. Isotope numbers are represented as $100 \times N_{\mathrm{res}} + N_{\mathrm{atom}}$, where N_{res} is the index of the residue in a parsed structure, and N_{atom} is the carbon number in the residue. This numbering scheme allows generation of structures containing up to 645 residues with up to 100 carbons in each residue, which is far enough for natural carbohydrates. SMARTS reaction expressions are prepared depending on the type of the linked atoms. In a few cases, bonding leads to formation of a new stereocenter (e.g. in glycopyruvates). If the configuration of such centers is specified in the CSDB Linear code, it is configured during postprocessing.

2. Details of generation of atomic coordinates

The algorithm of generation of atomic coordinates is shown in Figure S2.

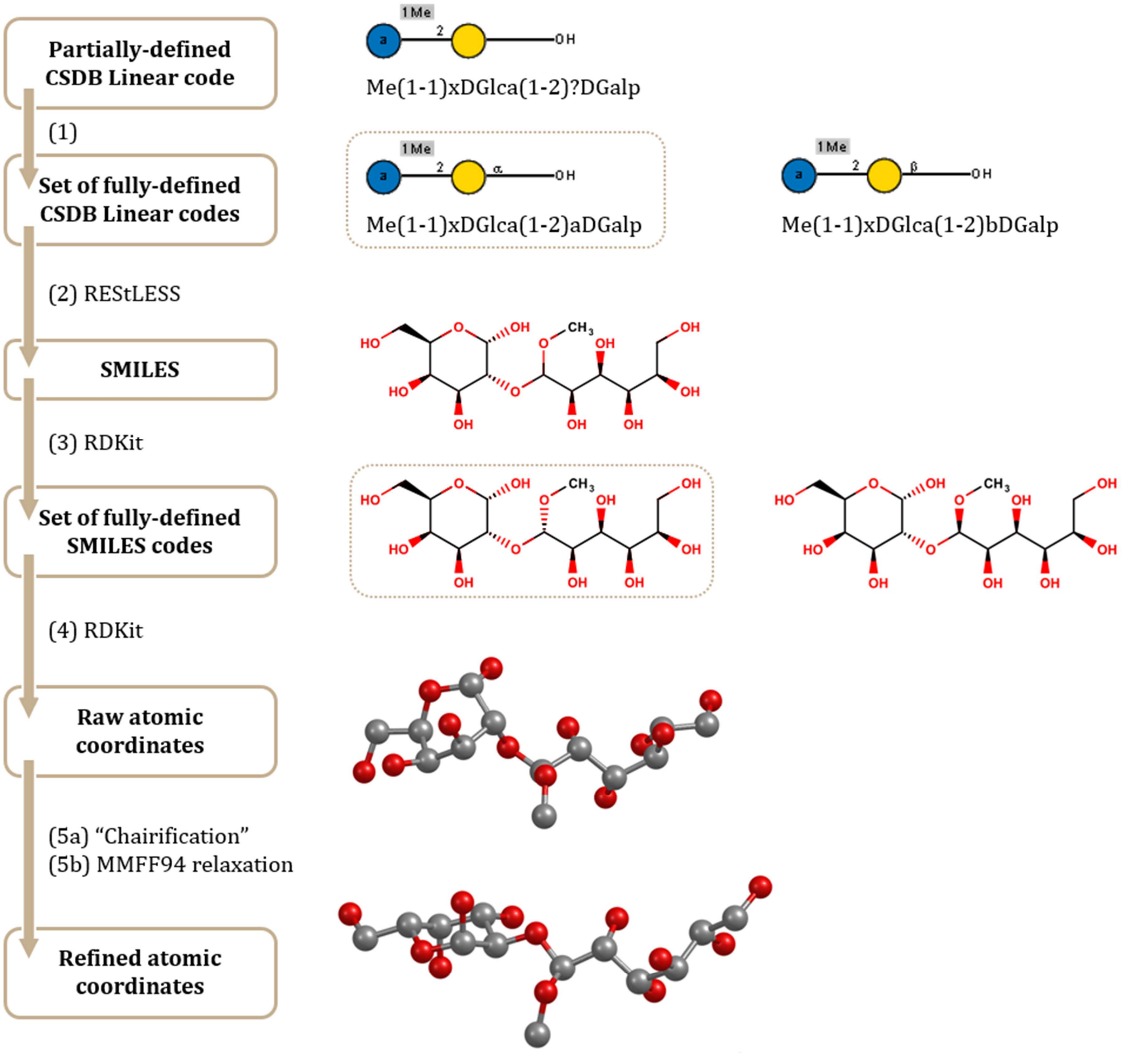

Figure S2. Generation of atomic coordinates exemplified on a hypothetic 1-O-methyl 1-O-D-galactopyranosyl D-glucose (open form). Dotted frames highlight the object used in the next step. Hydrogen atoms in 3D models are omitted for clarity.

The SMILES strings obtained from a carbohydrate or derivative structure can have undefined configurations of stereocenters due to unknown configurations of some atoms in some residues (e.g. anomeric ones). For such structures, a set of fully-defined SMILES strings required for generation of atomic coordinates is derived in three steps: the initial CSDB Linear code is transformed to a set of fully-defined CSDB Linear codes (Fig. S2, step 1), each fully-defined CSDB Linear code is translated to SMILES (Fig. S2, step 2), and each of the obtained SMILES strings is transformed to a set of fully-defined SMILES codes (Fig. S2, step 3). Fully-defined CSDB Linear code almost always corresponds to fully-defined SMILES, so the third step is required only in rare cases, such as chiral acetals or residues with stereo uncertainties encoded in the name (e.g. fatty acids with undetermined stereo configurations of distal branches).

Atomic coordinates for fully-defined SMILES codes are generated by RDKit (Fig. S2, step 4). However, we found out that if a molecule contained several saturated six-membered rings, such as pyranoses, they are often erroneously simulated to adopt twist, boat or even inverted chair conformation (e.g. $^1C_4$ for D-glucopyranose). To overcome this problem, the torsion angles of each pyranose ring were adjusted to model either $^1C_4$ or $^4C_1$ conformation (Fig. S2, step 5a). The high temperature molecular dynamics simulations [5] were used to identify the preferred conformation for each of 381 pyranoses deposited in CSDB residue subdatabase. The MM3 force field, as implemented in the TINKER suite [6], was used to calculate 1-ns trajectories at 1000 K. MM3 has been reported as an appropriate force field for modeling of carbohydrates [7]. The choice of the preferred conformation followed counting the number of 1-ps steps during which a pyranose ring adopted the $^1C_4$ or $^4C_1$ conformation. The chosen conformations of pyranoses were stored in a dedicated database.

After this “chairification”, molecules were relaxed by MMFF94 optimization (Fig. S2, step 5b). At this step, MMFF94 was used instead of MM3 as more general force field, which better suits non-carbohydrates residues with diverse atomic types populated in CSDB structures (there are no MM3 parameters for a number of atom combinations present in CSDB structures).

The designed algorithm worked well with processing of molecules containing up to 200−250 non-hydrogen atoms. However, generation of bigger structures might exceed a timeout of 60 sec introduced to save server resources during bulk operations on multiple requests. We overcame this problem by caching of the atomic coordinates at the first user request and by pre-generation of 37571 MOL-files corresponding to 19946 structures stored in CSDB.

3. Support of structural features

Table S1 contains structural features of natural glycans with indicated support by the REStLESS translator. If an input CSDB Linear code describes a repeating unit of a regular polymer, the start and the end of the repeating fragment are represented by dummy atoms with zero atomic numbers (Figure S3). The CSDB Linear notation allows superclasses and aliases if certain residues in a structure are underdetermined or unsupported by a monomer subdatabase. Such residues are displayed as dummy atoms with an assigned isotopic number in the SMILES code. Not every code can be translated into a single SMILES string. This may occur due to an unspecified absolute configuration, ring size or bond positions. In this case, our algorithm produces all possible structures, for each of which a SMILES string is generated. If a residue contains only one stereo center and its absolute configuration is undefined, the corresponding atom is set as non-configured in SMILES. The same is done for an anomeric atom if the anomeric configuration is not known.

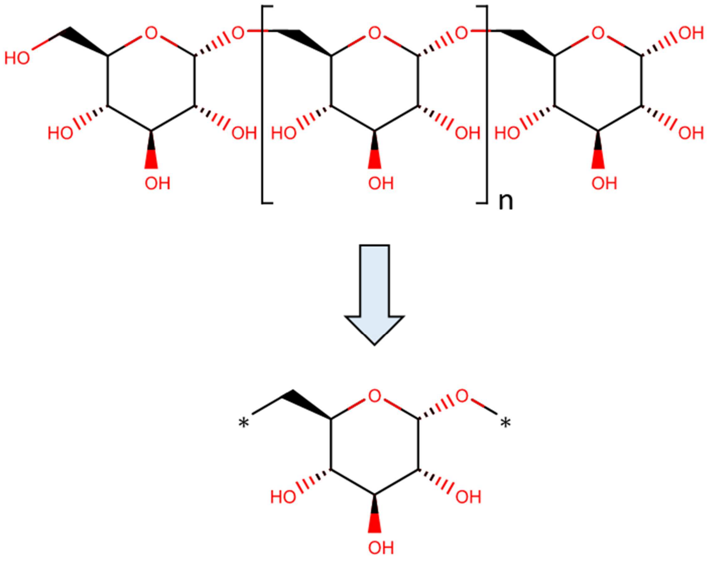

Figure S3. Representation of polymer repeating unit bounds with dummy atoms exemplified on α-1,6-glucan.

Table S1. Support of structural features.

Feature	Supported in CSDB Linear [a]	Transla-table to SMILES [b]	CSDB Linear examples	Corresponding SMILES (for the first CSDB Linear example) [c]	Comments
Residue level					
Monosaccharides	+	+	aDGlcpN, aXKdop, aXLDmanHepp	N[C@H]1[C@@H](O)O[C@H](CO)[C@@H](O)[C@@H]1O	234 residue prototypes [d]
Pyranoses, furanoses and open chain sugars	+	+	aDGlcp, aDGlcf, aDGlca	OC[C@H]1O[C@H](O)[C@H](O)[C@@H](O)[C@@H]1O	
Onic acids, alditols and inositols	+	+	xDGro, xDGlcN-ol, xXmyoIno myoIno = myo-inositol	OC[C@@H](O)CO	63 residue prototypes [d]
Phosphates and sulfates	+	+	xXEtN(1-P-P-5)[P-4)]aXKdop, aDGlcp(1-P-5)xDRib-ol EtN = ethanolamine	NCCOP(=O)(O)OP(=O)(O)O[C@@H]1[C@H](OP(=O)(O)O)C[C@](O)(C(=O)O)O[C@@H]1[C@H](O)CO	terminal or inline
Fatty acid residues	+	+	lXPam, lX3HOLau Pam = palmitic acid, 3HOLau = 3-hydroxylauric acid	CCCCCCCCCCCCCCCC(=O)O	167 residue prototypes [d] (including 25 sphingoids)
Amino acid residues	+	+	xLLys, xXPmN2, xXSRCetLys SRCetLys = S,R-carbo-xyethyl-lysine, PmN2 = diaminopimelic acid	NCCCC[C@H](N)C(=O)O	42 residue prototypes [d]
Other non-carbohydrate residues	+	+	xSPyr, xXCho, xXSuc Pyr = pyruvic acid, Cho = choline, Suc = succinic acid	C[C@](=O)C(=O)O	70 residue prototypes [d] (including 9 nucleotides)
Atypical residues (free text aliases and aglycons)	+	+/–	aDGlcp(1-3)Subst // Subst = enoxolone	**[1*]**O[C@H]1O[C@H](CO)[C@@H](O)[C@H](O)[C@H]1O	represented as isotopically labeled dummy atoms in SMILES
Residue superclasses	+	+/–	LIP(1-3)xDGro(1-P-2)HEX	**[1*]**OP(=O)(O)OC[C@H](O)CO**[2*]**	28 superclasses; represented as isotopically-labeled dummy atoms in SMILES
Linkage level					
Residue modifications	+	+	Ac(1-2)[Me(1-3)]aDGlcp	CO[C@@H]1[C@@H](OC(C)=O)[C@@H](O)O[C@H](CO)[C@H]1O	alkylation, acetylation, etc.

Dual linkages	+	+	xSPyr(2-4:2-6)aDGalp	C[C@@]1(C(=O)O)OC[C@H]2O[C@H](O)[C@H](O)[C@@H](O)[C@H]2O1	4,6-O-pyruvates, 1,1-linked monosaccharide diacetates
Carbon-carbon and carbon-nitrogen linkages	+	+/–	Me(1C-3)aDGlcp	C[C@@]1(O)[C@@H](O)[C@@H](O)O[C@H](CO)[C@H]1O	C-glycosyl compounds, N-glycans; carbon-carbon bonds with undefined bond positions are unsupported
Non-stoichiometrical linkages	+	+/–	-4)[30%Ac(1-3),xXEtN(1-%P-6)]aDGlcp(1-	[*]O[C@H]1O[C@H](COP(=O)(O)OCCN)[C@@H]([*])[C@H](OC(C)=O)[C@H]1O	30% of glucose residues are acetylated, and an unknown part of glucose is phosphorylated; SMILES is generated for the structure with fully stoichiometric linkages
Ester and amide bonds	+	+	Ac(1-1)xLLys(2-6)aDGalpA	CC(=O)OC(=O)[C@H](CCCCN)NC(=O)[C@H]1O[C@H](O)[C@H](O)[C@@H](O)[C@H]1O	
Topology level					
Oligomeric structures	+	+/–	bDGlcp(1-2)aDFruf	OC[C@H]1O[C@@H](O[C@@]2(CO)O[C@H](CO)[C@@H](O)[C@@H]2O)[C@H](O)[C@@H](O)[C@@H]1O	
Polymeric structures	+	+/–	-9)[Ac(1-5)]aXNeup(2-	**[*]**C[C@@H](O)[C@@H](O)[C@@H]1O[C@@](O**[*]**)(C(=O)O)C[C@H](O)[C@H]1NC(C)=O	the bounds of the repeating unit are denoted as dummy atoms
Cyclic polymers	+	–	CYCLO -4)bDGlcp(1-		supported externally as “CSDB molecule type”
Nested repeating units	–	–			
Repeating parts in oligomers	–	–			
Biological repeats	+	–	BIOL -4)aLRha(1-3)[Ac(1-2)]aDGlcpN(1-		supported externally as “CSDB molecule type”
Structural ambiguities					
Undefined anomeric configurations	+	+	?DGlcp	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	
Undefined absolute configurations	+	+	a?Rhap	C[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O ; C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O)[C@H]1O	undefined chiral center is produced for residues with only one chiral atom,; otherwise several structures (all combinations of chirality) are generated

R/S-configuration of acetal carbons acquiring chirality on bond formation	–	+/–	Ac(1-1)[Me(1-1)]xDGlca	COC(OC(C)=O)[C@H](O)[C@@H](O)[C@H](O)[C@H](O)CO	CSDB Linear does not support specification of such chiral centers; however, at the stage of producing atomic coordinates all possible stereomers are generated
Undefined ring size	+	+	bDGal?	OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@H]1O ; OC[C@@H](O)[C@@H]1O[C@@H](O)[C@H](O)[C@H]1O	several isomeric structures are generated
Undefined bond positions	+	+	aDGlcp(1-?)aLRhap	C[C@@H]1O[C@@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H](O)[C@H](O)[C@H]1O ; C[C@@H]1O[C@@H](O)[C@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H](O)[C@H]1O ; C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H]1O ; C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O)[C@H]1O[C@H]1O[C@H](CO)[C@@H](O)[C@H](O)[C@H]1O	all chemically possible structures are generated
Alternative residues or branches	+	+	<Ac(1-2)|Me(1-3)>aDGlcp, -3)<<aDGlcp(1-4)|aDGalp(1-4)>>aLRhap(1-	CC(=O)O[C@H]1[C@@H](O)O[C@H](CO)[C@@H](O)[C@@H]1O ; CO[C@@H]1[C@@H](O)[C@@H](O)O[C@H](CO)[C@H]1O ; CO[C@@H]1[C@@H](OC(C)=O)[C@@H](O)O[C@H](CO)[C@H]1O	OR or XOR logic for any number of alternatives in CSDB Linear; multiple SMILES are generated
Alternative or undefined attachment of branches at node level	–	–			
Residue composition only	–	–			

[a] "+" and "–" signs stand for "supported" and "unsupported", respectively.

[b] "+", "+/–"and "–" signs stand for "translatable", "translatable with limitations" and "untranslatable", respectively.

[c] Fragments of SMILES strings discussed in the comments are shown in bold.

[d] Residue prototypes are residues without N-linked acetyl groups which can further form N-acetylated structures (e.g., the residue prototype GlcN gives two structural fragments, GlcN and Ac(1-2)GlcN). The number of residues is usually greater than the number of prototypes.

4. RESTLESS API

Glycan structures can be translated from the CSDB Linear notation to SMILES in the unmanned mode by using an automated programming interface (API) at http://csdb.glycoscience.ru/database/core/convert_api.php. The structure in the CSDB Linear language should be passed as the HTTP POST or HTTP GET parameter named *'csdb'*. If you use GET, the parameters should be URL-encoded. The destination format should be specified in the *'format'* parameter as *'smiles'*. The plain text output contains: the destination format (line 1), one or more obtained SMILES strings (one line per structure, lines 2+), a blank line, error messages if present, a blank line, and a copy of input CSDB code, for example:

Input:

```
csdb=aDGlc?

format=smiles
```

Output:

```
SMILES:
STRUCTURE 0: OC[C@H]1O[C@H](O)[C@H](O)[C@@H](O)[C@@H]1O
STRUCTURE 1: OC[C@@H](O)[C@H]1O[C@H](O)[C@H](O)[C@H]1O

Errors: none

Input: aDGlc?
```

5. Programming technologies used

The tool engine and web interface were implemented using MySQL 5, PHP 5, Python 3 and were tested in modern versions of Mozilla Firefox, Google Chrome and Microsoft Internet Explorer. Structural formulas are visualized as raster or vector images by the CSDB engine, which uses the Open Babel toolbox [8], fed by SMILES strings output by the translator.

6. Abbreviations

API: Application Programming Interface;

CSDB: Carbohydrate Structure DataBase;

HELM: Hierarchical Editing Language for Macromolecules;

REStLESS: REsidues as SMILES and LinkagEs as SMARTS;

SCSR: Self-Contained Sequence Representation;

SMARTS: SMILES Arbitrary Target Specification;

SMILES: Simplified Molecular-Input Line-Entry System;

WURCS: Web3 Unique Representation of Carbohydrate Structures.

7. References

1. http://csdb.glycoscience.ru/ database/core/residues.php
2. Toukach, Ph. V. and Egorova, K. S. (2016) Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Res*, **44**, D1229–D1236.
3. http://www.rdkit.org
4. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
5. Frank, M., Lütteke, T. and von der Lieth, C. W. (2007) GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res*, **35**, D287–D290.
6. https://dasher.wustl.edu/tinker/
7. Toukach, Ph. V. and Ananikov, V. P. (2013) Recent advances in computational predictions of NMR parameters for the structure elucidation of carbohydrates: methods and limitations. *Chem Soc Rev*, **42**, 8376–8415.
8. O'Boyle, N. M., et al. (2011) Open Babel: An open chemical toolbox. *J Cheminform*, **3**, 33.