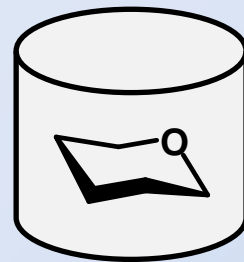




**Ф.В. Тоукач**

# **Углеводные базы данных**

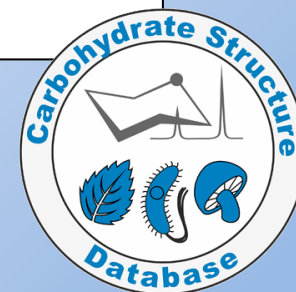


<http://toukach.ru/rus/glyco-db.htm>

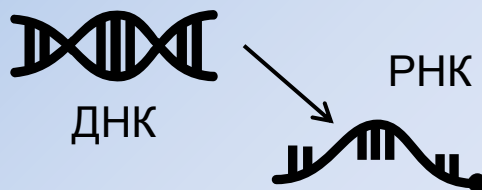
# Углеводные базы данных

2

- Что такое углеводы и что такое базы данных?
  - Что мы от них ожидаем?
  - Почему мы этого не получаем?
- Какой должна быть идеальная база данных?
- Кто и что уже сделал?

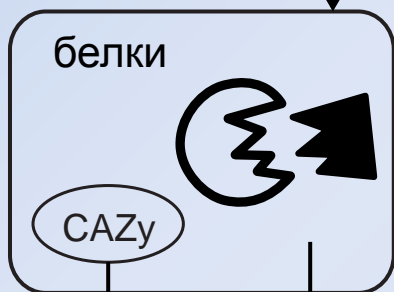


# Углеводы в клетках

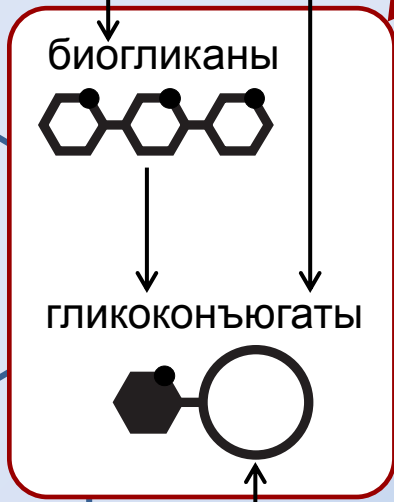
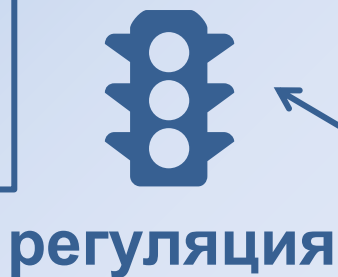


*самый химически-разнообразный класс биомолекул*

узнавание, адгезия, клеточные контакты  
стабилизация конформации белков  
барьерная функция, локальный pH  
антикоагуляция



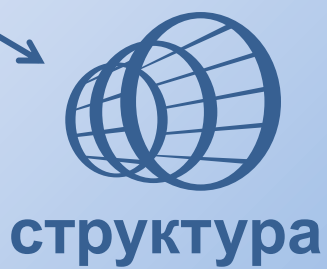
иммунитет и мимикрия (контакт «свой-чужой»)  
защита от протеолиза (контакт «свой-свой»)  
биоактивность  
механическая защита



биосинтез АТФ  
энергетический резерв

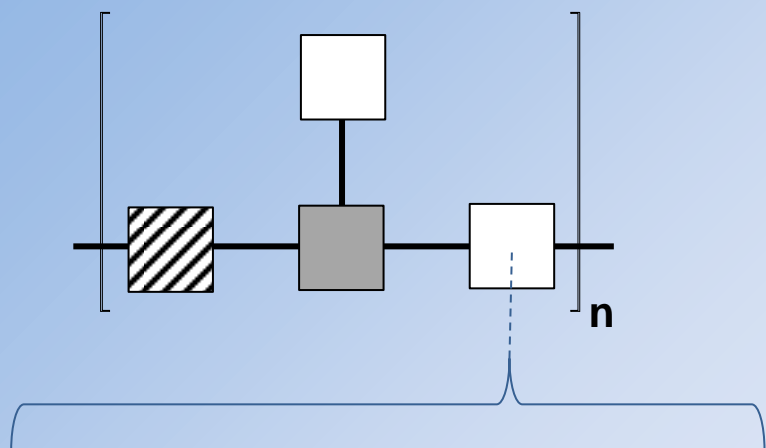


протеины, липиды, др. агликоны



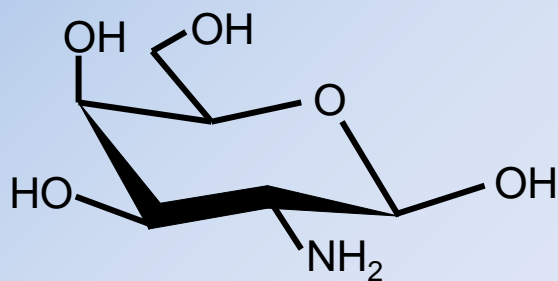
нуклеиновые к-ты  
GPI-якоря  
клеточная стенка

# Структура углеводов

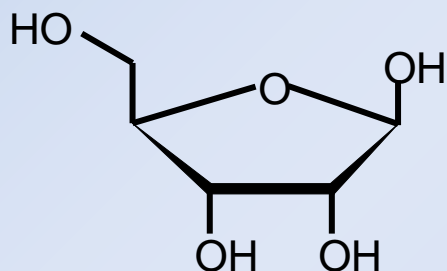


## Полная структура

- мономерный состав, в т.ч. неуглеводный
- топология и последовательность
- позиции замещения
- стехиометрия боковых цепей
- число и границы звеньев



пример альдо-пиранозы ( $\beta$ -D-GalpN)



пример альдо-фуранозы ( $\beta$ -D-Ribf)

## Структура остатка

- размер скелета (4-10)
- стереохимия всех центров (=мономер)
- размер цикла (*p/f/a*, *aldo/keto*)
- аномерная форма ( $\alpha/\beta$ )
- абсолютная конфигурация (D/L)
- модификации (-NH<sub>2</sub>, -COOH, deoxy)



# Актуальные темы гликомики

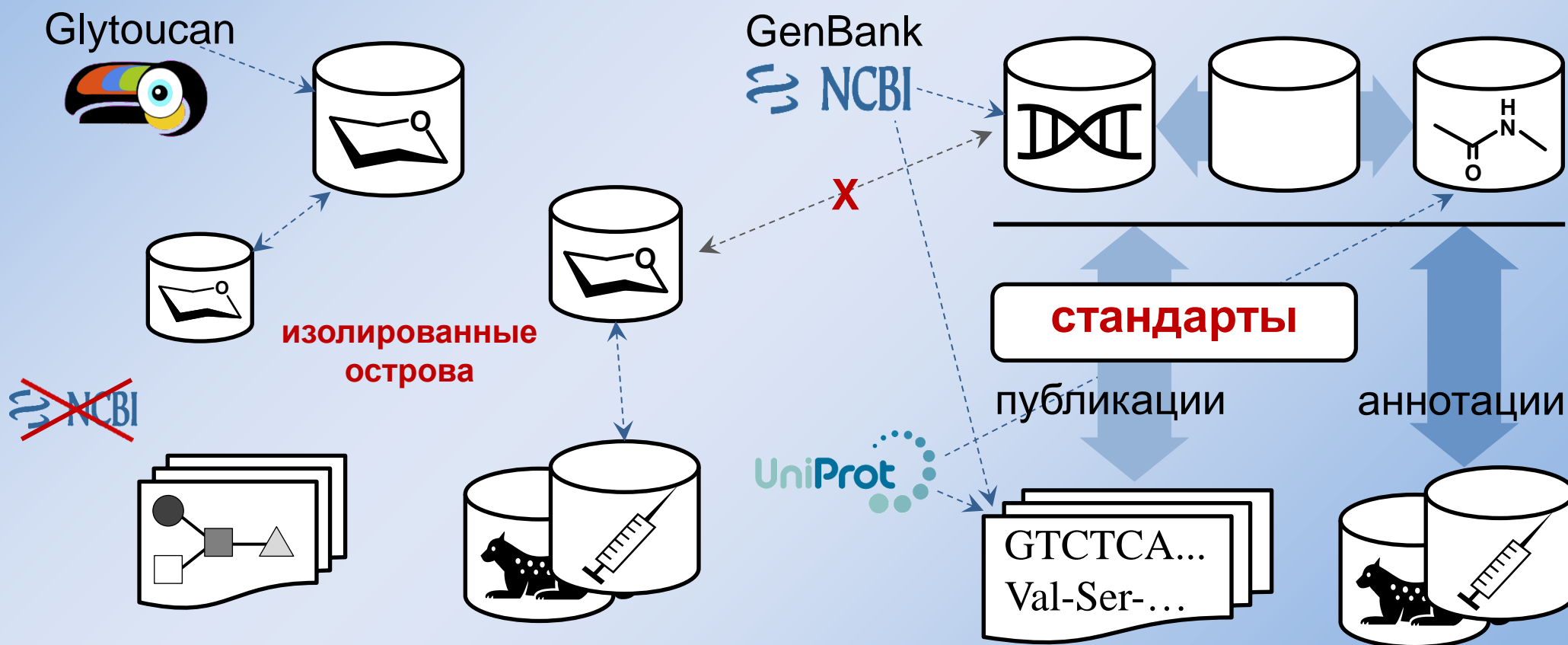
- Структура, изомерия, конформация углеводов в клетках
- Таксономия и классификация микроорганизмов
- Гликоэпитопы и иммуноспецифичность организмов
- Объяснение взаимодействия антиген-антитело
- Гликосодержащие вакцины и лекарства
- Корреляция свойств организма с его углеводами
- Биосинтез и круговорот углеводов

# Гликомика vs. геномика, протеомика

6

по сравнению с другими *-omics*:

- сходный объем информации (>100 000 известных структур)
- бóльшая химическая вариативность
- меньшее использование IT (базы данных, сервисы)
- меньшая стандартизация



# Зачем нужна база данных?

## База данных: легкий доступ к знаниям

- Какие природные структуры похожи на заданные? Какие их фрагменты специфичны для заданных биологических видов? В каких местах гликанов склонны находиться те или иные структурные единицы?
- Где они опубликованы, в привязке к каким таксонам и болезням? Какие ферменты их синтезируют и с какой достоверностью это показано? На какие гликоэпитопы реагируют антитела?

## Платформа: моделирование и предсказание

- Какие конформации характерны для гликозидной связи в заданном трисахариде? Какие сигналы ЯМР характерны для его необычных компонентов? С какой достоверностью можно об этом судить? Какие еще структуры дают похожий набор сигналов?
- Какие бактериальные виды близки в контексте структуры O-антигенов? Как соотносятся их кластеры с классическим деревом жизни и серогруппами?

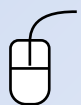
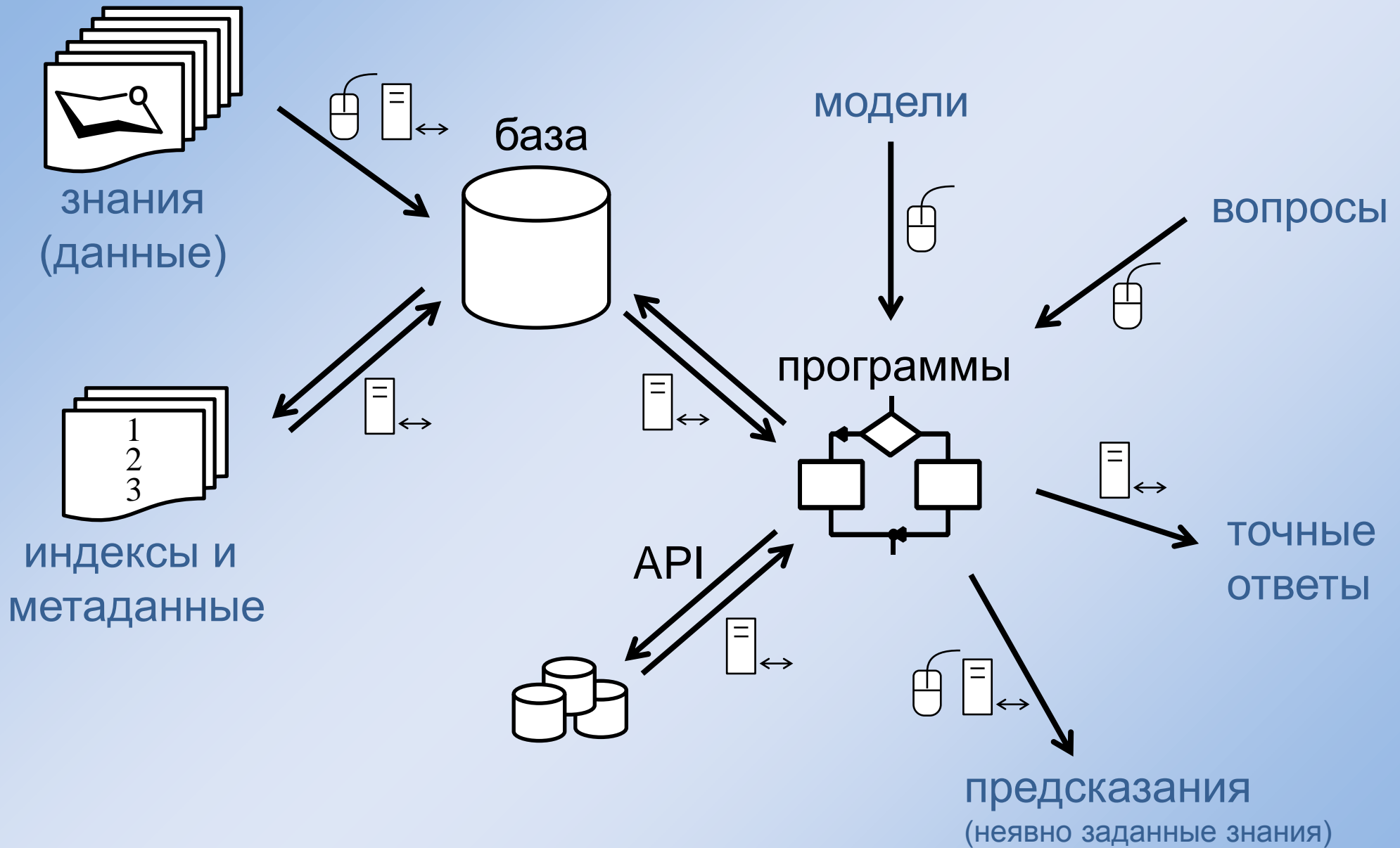
## Методология и автоматизация обработки накопленных знаний

- Идентификация и кодирование молекул в базах данных и программах
- Визуализация молекул в публикациях и программах
- Распознавание контекста и аннотирование публикаций
- Взаимодействие проектов, унификация интерфейсов

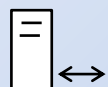
# Проблемы в гликоинформатике

- Вариативность и гетерогенность объектов
- Неоднозначное описание структуры
- Сложности с вводом и визуализацией больших структур
- Отсутствие стандартов
- Изолированность проектов
- Неполнота и низкое качество данных в базах
- Ресурсоемкие алгоритмы
- Нехватка системного видения у разработчиков и пользователей  
(нет общепризнанных сервисов, инициативы несовместимы друг с другом)

# Что такое база данных?



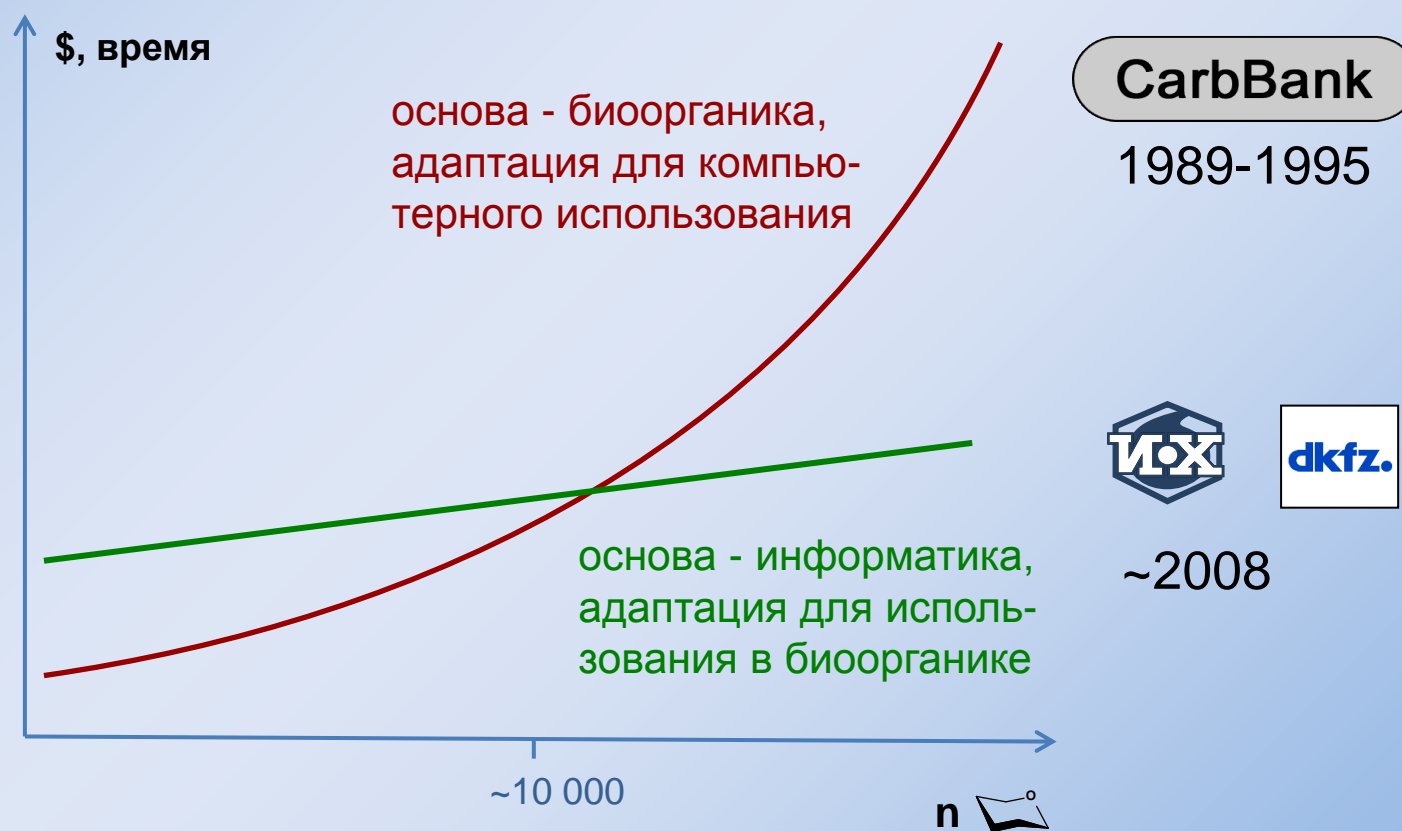
с участием человека



автоматически

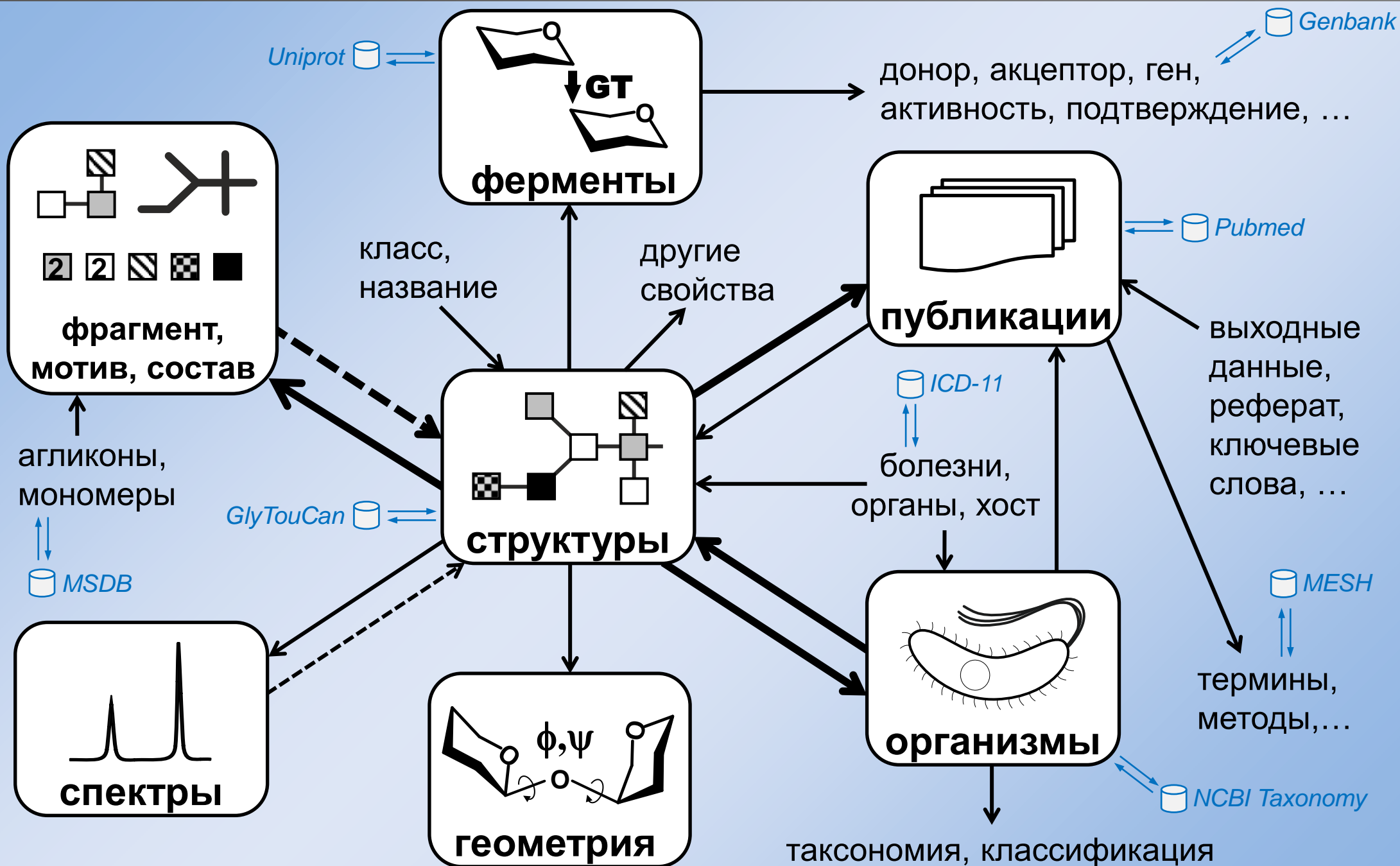
# Организация разработки

- База данных и платформа должны быть построены по правилам информатики.
- Правила были конкретизированы и адаптированы для углеводов.



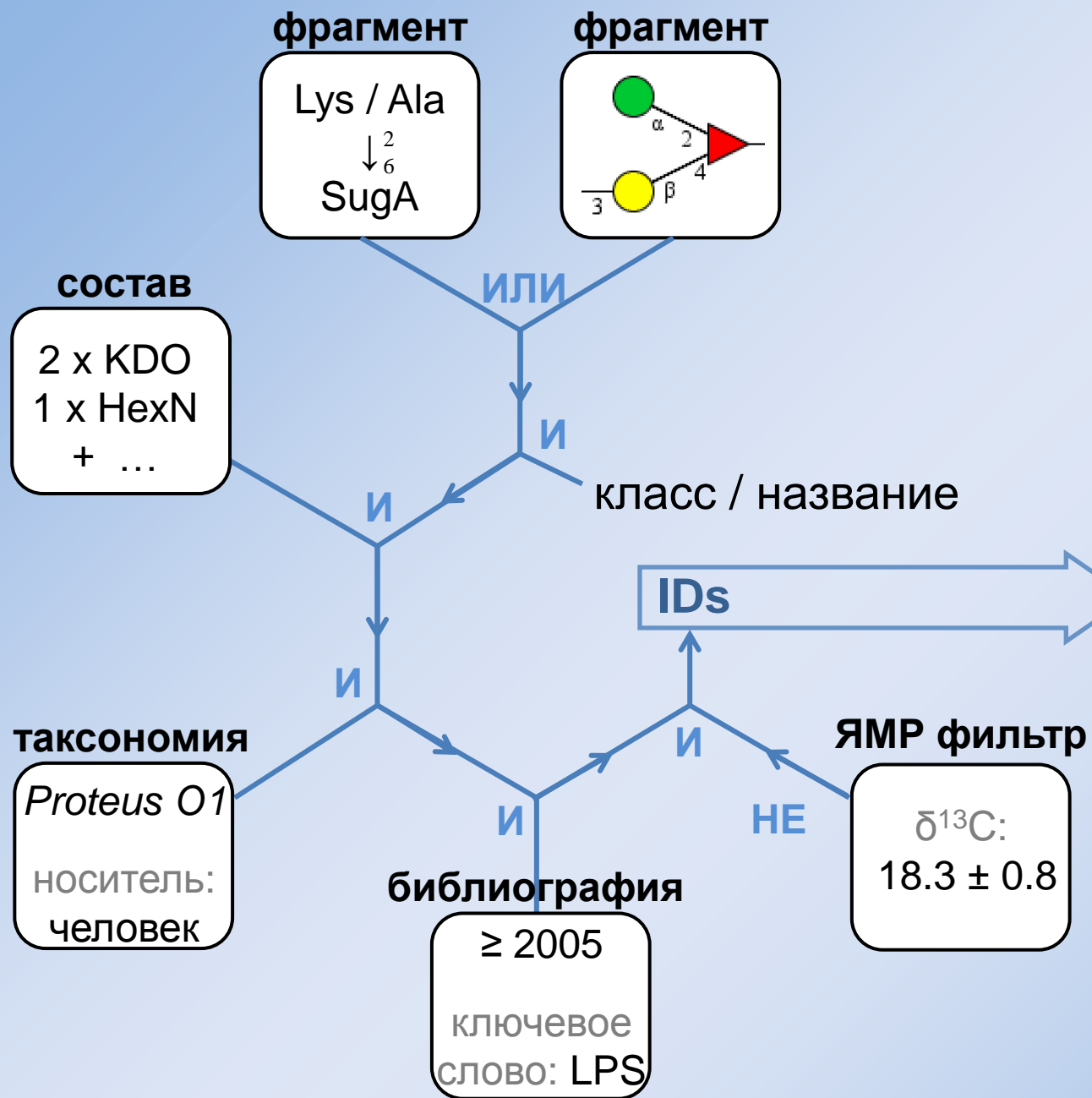


# Типичные запросы





# CSDb: составной запрос



Данные сгруппированы по соединениям, публикациям, организмам и т.д.

There are 2 chemically distinct structures. Please, select:  
-4)[60%Me(1-3)aDRhap(1-3)]aLFucp(1-4)?DXylp(1-4)]aLFucp(1-3)]aLFucp(1-4)xDXyla(1-

Found 5 structures. Displayed structures from ...  
Expand all compounds Show all as text (Sweet-13)

1. Compound ID: 10502

Structure type: polymer chemical repeating unit  
Compound class: O-polysaccharide, O-antigen

Structural formula & atomic coordinates  
Sweet-II 3D model

The structure is contained in the following publications:

- Article ID: 4266  
Boйко AS, Dmitrenko AS, Fedonenko YF  
**O-polysaccharide of the lipopolysaccharide of *Azospirillum brasilense* strain Jm6B2 (D-Acofrioise)** - Carbohydrate Research

Two types of neutral O-polysaccharide were isolated from the aqueous phenol-water extraction from the azygous cells of *Azospirillum brasilense* strain Jm6B2. The major O-polysaccharide was characterized by 1D and 2D (1)H and (13)C NMR spectroscopy. The presence of D-rhamnose is indicated by italics.

Lipopolysaccharide, structural, O-antigen, physiology, *Azospirillum brasilense*, D-Acofrioise

NCBI PubMed ID: 22575749  
Publication DOI: 10.1016/j.carres.2012.04.006  
Journal NLM ID: 0043535  
Publisher: Elsevier  
Correspondence: room308@ibppm.sgu.ru  
Institutions: Institute of Biochemistry and Physiology of Plants and Microorganisms, Russian Academy of Sciences, 410055, Moscow, Russia  
Methods: 1H NMR, 13C NMR, NMR-2D, methylation, chemical analysis, GLC, Smith degradation

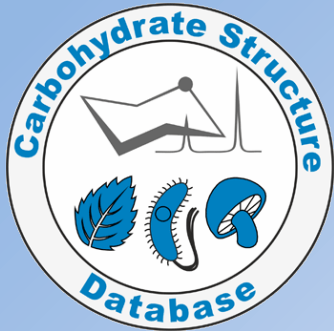
*Azospirillum brasilense* Jm6B2  
CSDb ID 2831

- Article ID: 4833  
Fedonenko YF, Dmitrenko AS, Boйко AS  
**The review of the polysaccharide structure of the genus *Azospirillum***

The spectrum also has 2 signals at unknown positions (not plotted).

# Carbohydrate Structure Database

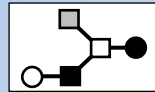
13



CSDB

## База данных природных углеводов

## Платформа для сервисов гликоинформатики



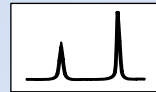
33K

первичные  
структуры



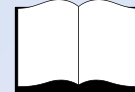
16K

таксономия



19K

спектры  
ЯМР



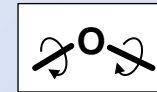
13K

библио-  
графия



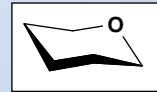
2K

гликозил-  
трансферазы



3K

геометрия



3K

мономеры

- Структурная база
  - прокариоты, грибы, растения, простейшие
  - проверяемый контент, полное покрытие
  - ежегодные обновления
- База гликозилтрансфераз
- Точная ЯМР-симуляция с отнесением
- Предсказание структуры по спектрам
- Конформационные расчеты сахаридов
- Углеводные деревья жизни
- Углеводная нотация CSDB Linear
- Symbol Nomenclature for Glycans
- Онтология GlycoRDF и другие стандарты
- Редакторы, конвертеры, стат. анализаторы, ...

<http://csdb.glycoscience.ru>



# Структурные базы

**CarbBank** 23 полная до 1996 ORIG архитектура, % ошибок

**GlycomeDB** 216 мета-репозиторий неполные аннотации нет агликонов

**GLYTOUCAN**

**CFG glycan** млекопитающие, > 6

SweetDB, SugaBase **SCIENTES.DE** 27 / 21

**GLYCAN** 11

**Eurocarb DB** архитектура только модель

BCSDB PFCSDB **Carbohydrate Structure Database** ORIG 15 / 6 (бактерии, археи) 16 / 6 (грибы, растения)

полная по прокариотам и грибам курируемая

**GlycoSuite** ORIG млекопитающие+... 10 / 1 полная до 2005

**JCGGDB** > 70 коллекция баз слабо аннотированы

**UniCarbKB** 4 / 1 курируемая

**nibrT** 0.7 O- & N- ORIG

**Glycoconjugate Data Bank** 44

**EcoDAB** 0.2 *E. coli* ORIG

**GlyGen** 34 (человек, мышь, крыса, вирусы)

**GlycoBase** 0.3 животные ORIG



# Специальные базы



углеводные гены человека

**GGdb**

~0.2

MS<sup>2,3,4</sup> N- и O-гликанов



**GMDB**

~0.2

химические реакции

**GlycoNAVI**

~3.0 (4.4 str)

конъюгаты & агликаны



**GlycosideDB**

>70

N-гликопротеины  
*C. elegans* + мышь



**GlycoProtDB**

~2.5

методики синтеза  
и анализа

**GlycoPOD**  
GlycoScience Protocol Online Database

~0.2 (0.5 sub)

глико-эпитопы  
и антитела



**GlycoEpitope**

~0.2 (0.6 ABs)

GlyTOUcan,  
репозиторий идентификаторов



~216

адгезия к патогенам



**SugarBindDB**  
Pathogen Sugar-Binding Database

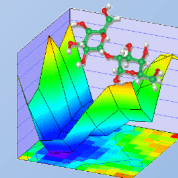
~0.9

O-glycBase,  
O- и C-гликопротеины

CENTER FOR  
RIBIOLOGICAL  
SEQUENCE  
ANALYSIS  
CBS

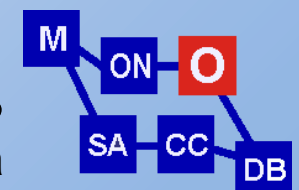
~0.2

GlycoMaps,  
расчетные конформационные карты



~2.6

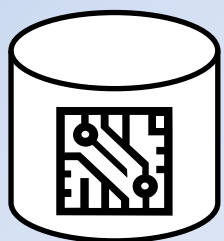
MSDB  
моносахариды и номенклатура



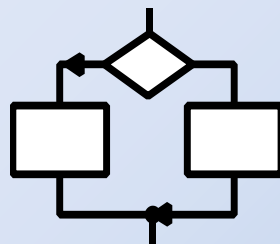
~0.8

# Критерии оценки

- **Функциональность** (типы данных и индексов, обработка запросов)
- **Полнота покрытия** (+ выбранный класс)
- **Качество данных** (% ошибок, прозрачность)
- **Интеграция** (поддержка форматов, импорт-экспорт, API, RDF)
- **Интерфейс** (простота, стабильность, производительность)



внутренняя  
архитектура

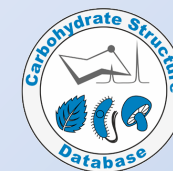


управляющие  
программы



наполнение  
данными

# Архитектура



- Реляционная база данных
- Индексация + стандартные индексы (DOI, TaxID, ICD-11, PMID, ...)
- Структуры, таксономия, библиография – разные типы записей
- Человекочитаемый дамп (организация процесса наполнения)
- Контролируемые словари терминов (мономеров; MSDB)
- ~~Free text~~ ↗
- Таблица связности

## минимум

структура,  
таксономия,  
библиография,  
ссылки на другие БД

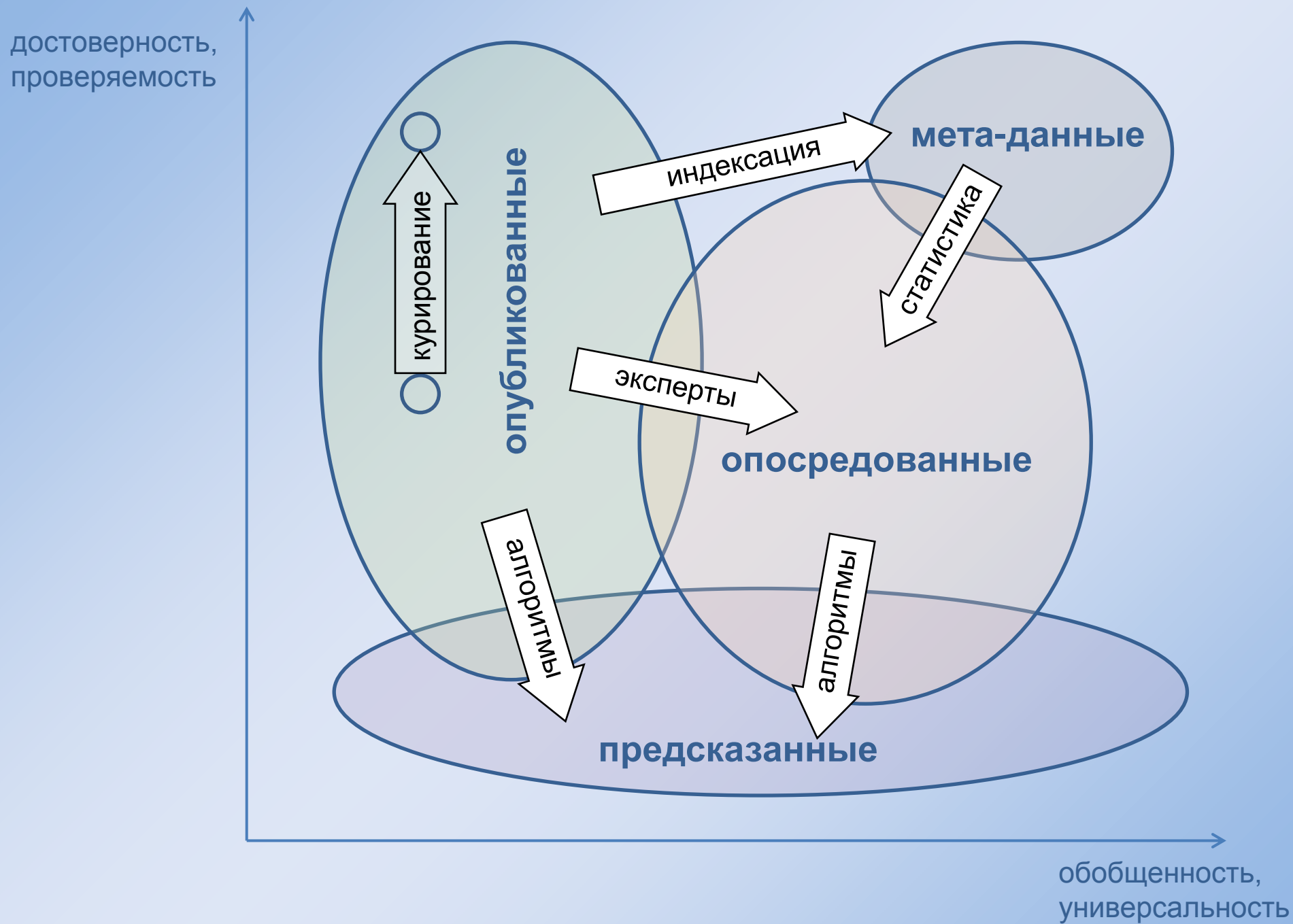
## дополнительно

тривиальные названия,  
спектры ЯМР, МС  
условия съемки спектров,  
биоактивность,  
гены, ферменты,  
конформация

болезни  
органы, ткани  
генотип, стадия

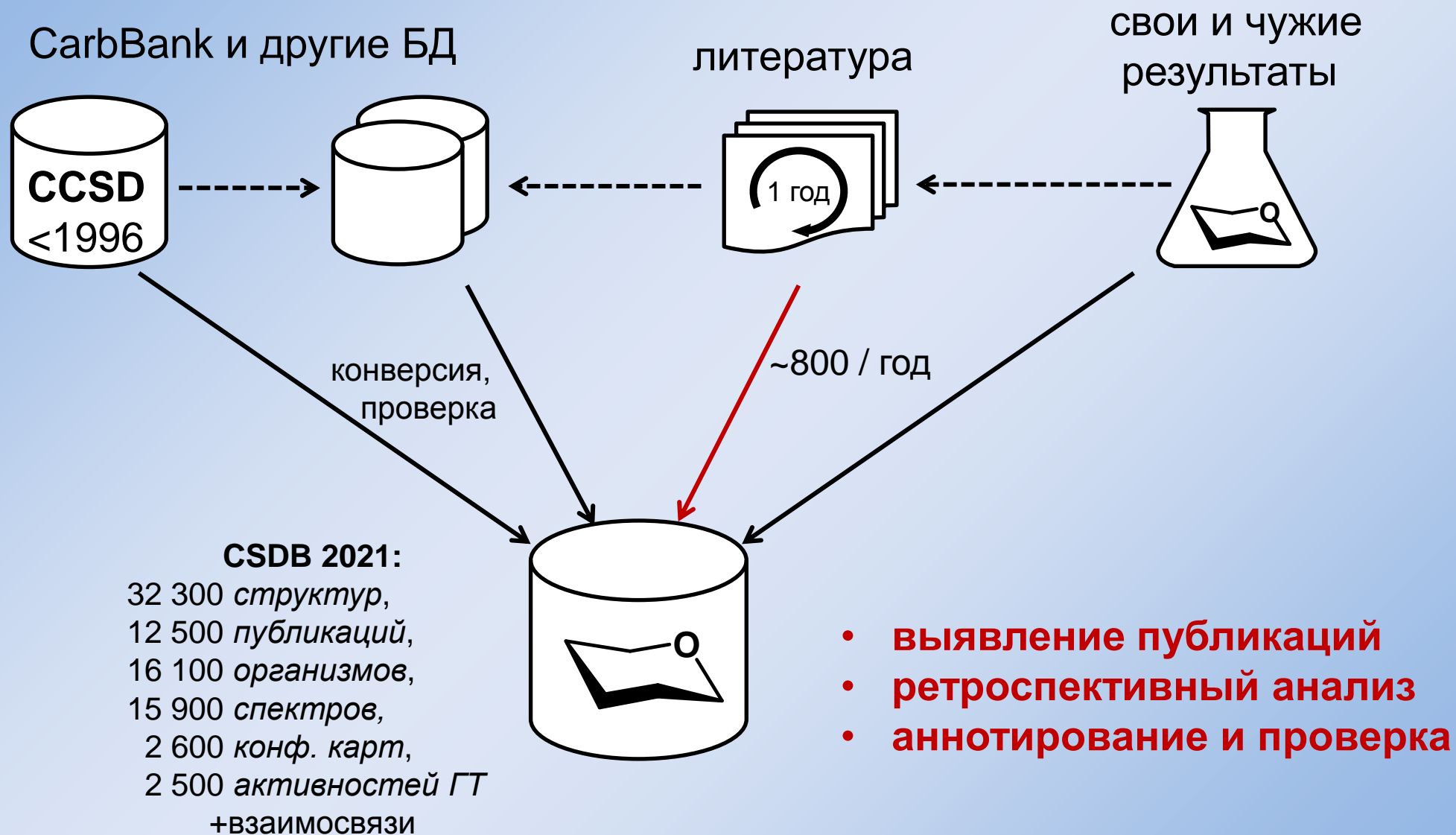
ключевые слова  
рефераты  
институты  
методы

# Уровни данных





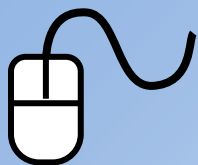
# Источники данных



**полное покрытие по прокариотам и грибам:**

отрицательный результат поиска = значимая научная информация

# Качество данных



операторские

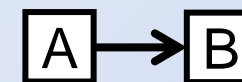


в базах

найдено &  
исправлено



в статьях



в программах

## ошибки, противоречия

### исправляемые

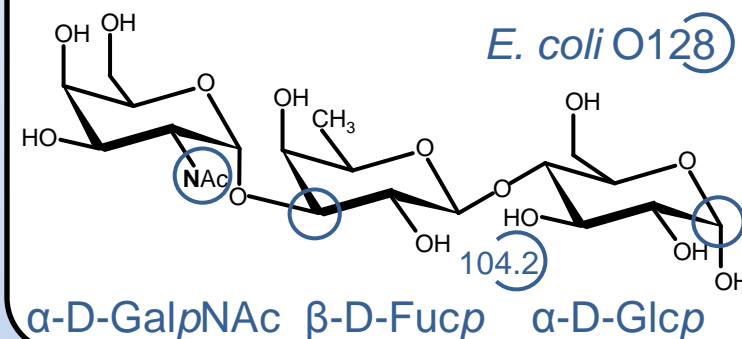
2dGlc → araHex,  
α-Rib-ol → Rib-ol,  
D-Kdo → Kdo,  
1-methyl → 1-Me,  
n.m.r. → NMR,  
taxid 583 → Proteus,  
...

### выявляемые

Glc(1-2)Glc**N**,  
anhydro-Kdo,  
**D**-manHep,  
Gal**p**5N,  
Ac(1-2)[Glc(1-2)]Gal,  
*Escherichia sapiens*,  
*Dev Food Sci* 2012,  
#Ac : 23 м.д., **65** м.д.  
D-G**cl**, ...

### невывявляемые

*E. coli* O127:  
aDGalp**N**(1-4)bDFucp(1-4)**b**DGlc p  
Glc C1 10**3.2** ppm



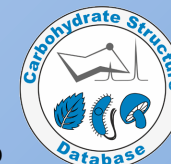
CarbBank



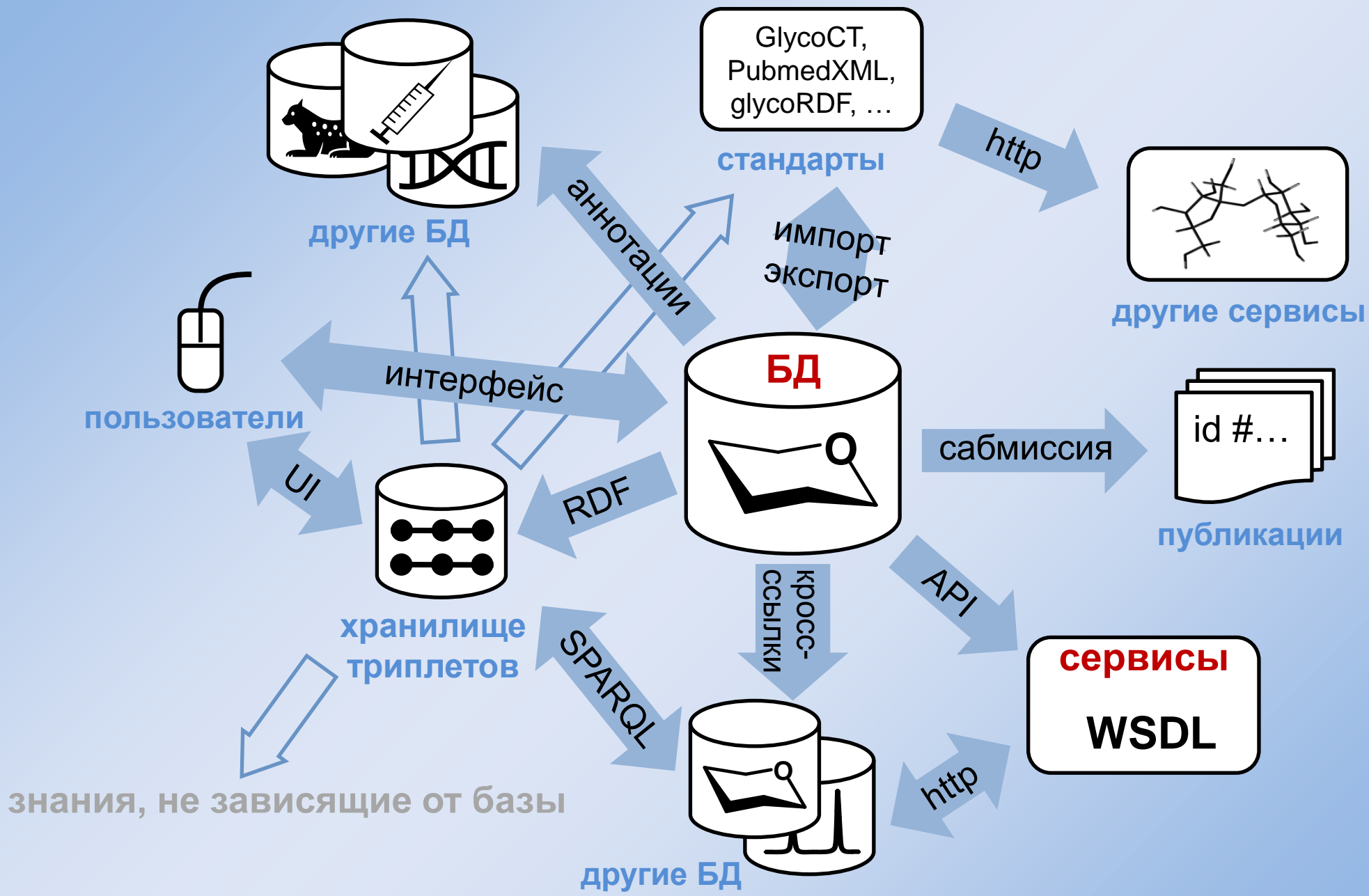
>50% (неправильные, отсутствующие, ложно присутствующие структуры, штаммы, аннотации)

↳ другие БД

😊 <10%



# Идеальная интеграция



# Интерфейс

SNFG,  
WURCS,  
GlycoCT,  
SMILES,  
MOL,  
PDB,  
Glyde II,  
LinUCS,  
Sweet-DB  
GLYCAM,  
GlycoRDF,  
DCI XML, ...

DOI,  
NCBI Pubmed,  
NCBI Taxonomy,  
Uniprot,  
NCBI Genbank,  
PubChem,  
ImmuneEpitopeDB,  
Glytoucan,  
ICD-11

Конверсия данных ↔ другие форматы

Автоматические web-сервисы (WSDL)

Импорт, экспорт

Документация, HELP

Ссылки на записи в других проектах

Дружественность, быстрое действие

Ввод и вывод структур

веб-помощник,  
граф. редактор,  
библиотека  
структур,  
CSDB Linear,  
GlycoCT

Topology: 4 residues (branched: ([A->B->],[C->]D) )

Structure: bDGalpNAc (1-3) aDG1cp (1-4) [bDManp (1-3) ] ?LFucp

Residue A: bDGalpNAc (1-

Residue B: aDG1cp (1-

wizard

Popular Small sugars Hexoses Higher sugars Alditols Aliphatic a

Novice Expert Insert Replace Oligo Poly Ac Am Cm Cho Fo M

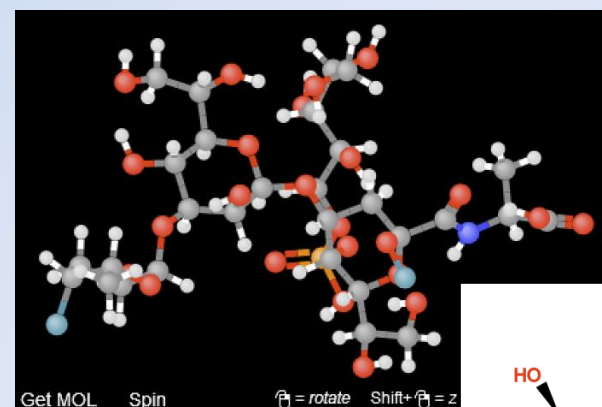
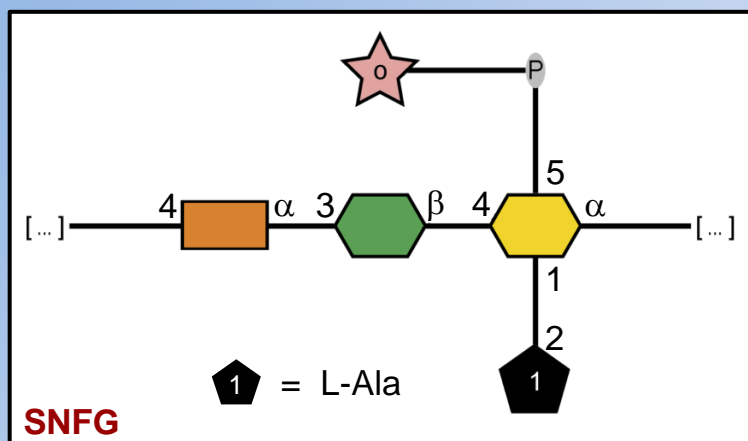
editor

bDManp (1-3) [Ac (1-2) bDGalpN (1-?) aDG1cp (1-4) ] ?LFucp

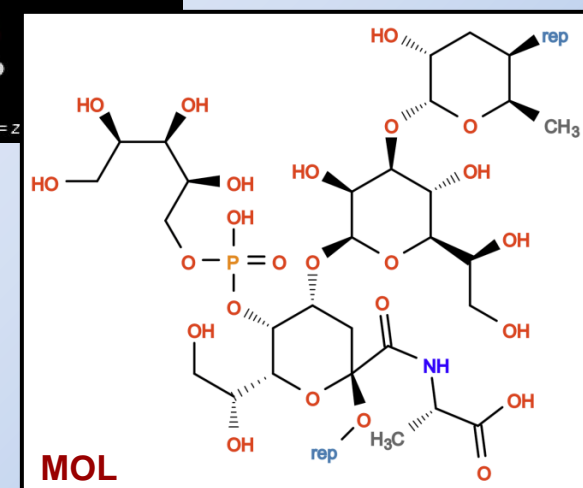
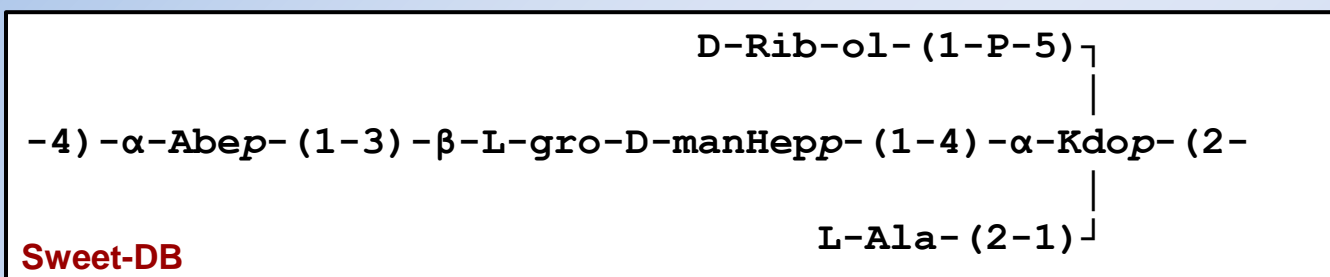
нотация CSDB Linear

# Вывод структур

Визуализация в человекочитаемых форматах:



**MOL (3D)**



↔ Экспорт в машиночитаемых форматах:

[\*]O[C@]1(C(=O)N[C@@H](C)C(=O)O)C[C@@H](O[C@@H]2O[C@H]([C@@H](O)CO)[C@@H](O)[C@H](O[C@H]3O[C@H](C)[C@H]([\*])C[C@H]3O)[C@@H]2O)[C@@H](OP(=O)(O)OC[C@H](O)[C@H](O)[C@H](O)CO)[C@@H]([C@H](O)CO)O1

**SMILES**

2.0/5,5,5/[Aad1122h-2a\_2-6][h222h][a11221h-1b\_1-5][a2d12m-1a\_1-5][A1m\_2\*N]/1-2-3-4-5/a1-e2\_a2-d4~\_a4-c1\_a5-b1\*OPO\*/3O/3=O\_c3-d1

**WURCS**

-4) aXAbep (1-3) bXLDmanHepp (1-4) [xDRib-ol (1-P-5) , xLAla? (2-1) ] aXKdop (2-

**CSDB Linear**



# CSDb/SNFG editor

online редактор

Chemical repeating unit; n=10

-3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]?DGlc(1-7)[Ac(1-2)]bDGalfN(1- // Subst = chrysin = SMILES O=c2cc(c1cccc1)oc3c{7}c(0)c{5}c(0)c23

- режимы «новичка» и «эксперта»
- все поли- и олигомерные топологии
- 250+ моносахаридов и 350+ других остатков
- SMILES для атипичных компонентов
- все типы связей (включая хелатные и C-C)
- поддержка неопределенностей, повторов, вариативности, суперклассов

Previews Refresh

Hi-res image

RES  
1r:r1  
REP  
REP1:5o(3+1)2d=-1--1  
RES  
2b:b-dgal-HEX-1:4  
3s:n-acetyl

«ЖИВОЙ»  
ЭКСПОРТ

Subst-(7-3)-D-Rib-ol-(1--P--4)--+  
-3)-α-L-Fucp-(1-6)-D-Glc(1-7)-b-D-GalfNac-(1- Subst = chrysin = SMILES O=c2cc(c1cccc1)oc3c{7}c(0)c{5}c(0)c23

There are 3 chemically distinct structures. Please, select:

1. -3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]?DGlc(1-3)[Ac(1-2)]bDGalfN(1- //
2. -3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]?DGlc(1-5)[Ac(1-2)]bDGalfN(1- //
3. -3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]?DGlc(1-6)[Ac(1-2)]bDGalfN(1- //

There are 2 sterically distinct structures. Please, select:

1. -3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]aDGlc(1-6)[Ac(1-2)]bDGalfN(1- //
2. -3)αLFucp(1-6)[Subst(7-3)αDRib-ol(1-P-4)]bDGlc(1-6)[Ac(1-2)]bDGalfN(1- //

Files: MOL, PDB, Glycam

SMILES code:  
[\*][C@H]1O[C@H]([C@H](O)COC2O[C@H](CO[C@H]3O[C@H](C)[C@H](O)[C@H]([C@H]([C@H]3O)[C@H](OP(=O)(O)OC[C@H](O)[C@H](OC3CC(O)C4C(=O)CC(-C5CCCC5)OC4C3)[C@H](O)CO)[C@H](O)[C@H]2O)[C@H](O)[C@H]1NC(C)=O

- работа с геометрией в браузере
- экспорт на углеводных и химических языках
- экспорт атомных координат
- визуализация
- цвета SNFG

HEATM	29	B2	PDA	1	1.473	7.165	4.614	1.00
HEATM	30	O1	GLC	2	2.550	2.920	-1.808	1.00
HEATM	31	O5	GLC	2	1.561	3.154	-2.751	1.00
HEATM	32	O5	GLC	2	2.051	3.270	-4.096	1.00
HEATM	33	O6	GLC	2	0.256	7.723	-4.232	1.00

# Редакторы структур

редактор	мономеры	неопределенности	повторы	импорт, экспорт	хим. модель
GlycanBuilder™	70+30	конф., связи, классы, топология	-	++	-
Sugar Sketcher	70+20	конф., связи, классы	+	+	-
Polys builder	100+0	конф., классы	-	+	-
GlycoGlyph	80+10+user	конф., связи, классы	++	+	-
DrawRings	70+0	конф., классы	-	+	-
CSDB editor	300+300+user	конф., связи, классы, альтернативы	++	++	+

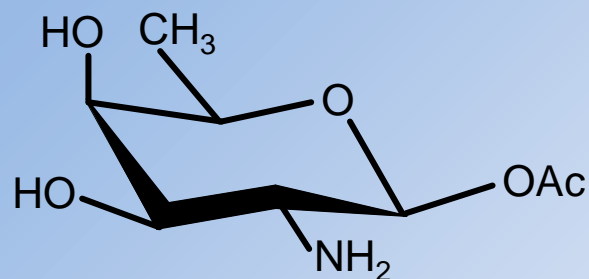
Углеводные,  
SNFG

## Химические

программа	плагин	мономеры	результат	оптимизация
Hyperchem	Sugar Builder	30+26+user	3D, можно задать $\varphi, \psi, \omega$	MM+, Amber, OPLS: MM, MD, ...
PCModel	Template: sugars	24+28	3D	MM3, MMFF, OPLS: MM, MD, ...
ACDLabs Chemsketch	Template: sugars	36+35+...	2D: stereo, chair, Haworth, Fisher	MM2: MM
ChemDraw / Chem3D	Template: hexoses	16+24+...	2D: chair, Fisher	MM2, MMFF: MM, MD



# Неоднозначность номенклатуры



bDFucpN(1-1)Ac (CSDB) ←

однозначно соотнесено со структурой, но при этом понятно людям

D-FucpN- $\beta$ 1OAc

beta-fucosamine acetate

1-acetoxy-beta-D-fucopyranosamine

2-deoxy-2-amino- $\beta$ -D-fucopyranosyl acetate (IUPAC)

$\beta$ -D-fucosamine acetic ester

$\beta$ -6-deoxy-D-galactosamine acetate

b-dgal-HEX|1:5|2-amino|1-acetate (GlycoCT)

$\beta$ -D-фукозамин-1-О-ацетат (на естественных языках)

(2S,3R,4R,5R,6R)-3-amino-4,5-dihydroxy-6-methyltetrahydro-2H-pyran-2-yl acetate (IUPAC)

N[C@H]([C@H]([C@H]([C@@H](C)O1)O)O)[C@@H]1OC(C)=O (SMILES)

1S/C8H15NO5/c1-3-6(11)7(12)5(9)8(13-3)14-4(2)10/h3,5-8,11-12H,9H2,1-2H3/t3-,5-,6+,7-,8+/m1/s1 (InChI)

# Почему не аналог MOL?

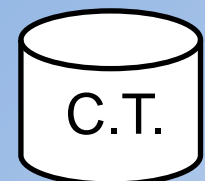
- Очень сложный перевод поатомного описания в семантическое (в структуру типа  $\alpha$ -D-Galp-(1-3)- $\beta$ -D-Glcp)
- Сложный перевод семантического описания в поатомное => трудоемкость аннотирования статей
- Нельзя описать структуры с неопределенностями
- Данные визуально не сопоставлены со знаниями
- Нечеловекочитаемый → трудно курировать → ошибки в данных
- Координаты атомов не являются первичными данными но неполный MOL (без 3D) может быть воспринят как 3D MOL
- Громоздко для хранения и передачи в сети (и не передается как параметр в URL)



52.0606	<del>6.3910</del>	<del>-0.1606</del>	<del>O</del>	0	0	0
51.8591	<del>8.6986</del>	<del>-0.1875</del>	<del>N</del>	0	0	0
52.9844	<del>9.0584</del>	<del>0.7259</del>	<del>C</del>	0	0	1
53.8550	<del>8.9929</del>	<del>0.0662</del>	<del>H</del>	0	0	0
52.9684	<del>10.5530</del>	<del>1.3121</del>	<del>C</del>	0	0	2
52.2705	<del>11.0903</del>	<del>0.6993</del>	<del>H</del>	0	0	0
1117	1	0	0	0	0	
1118	1	0	0	0	0	
1119	1	0	0	0	0	

**атомы, координаты, связность**

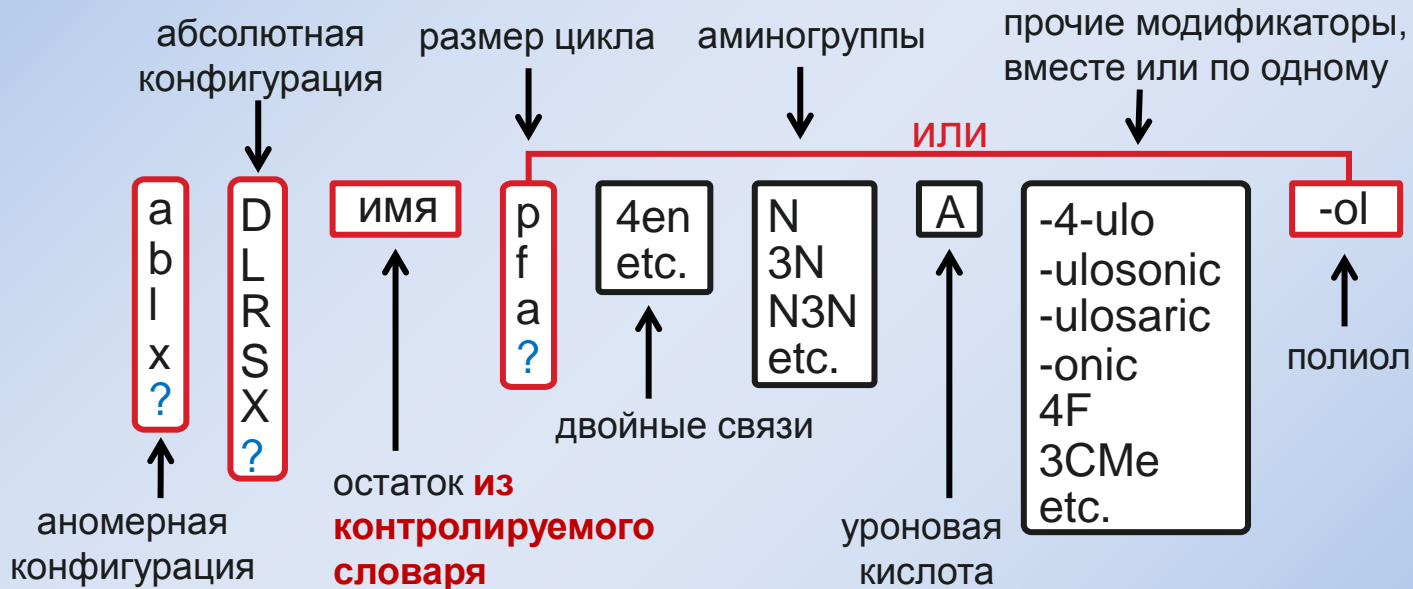
# Особенности структур и язык



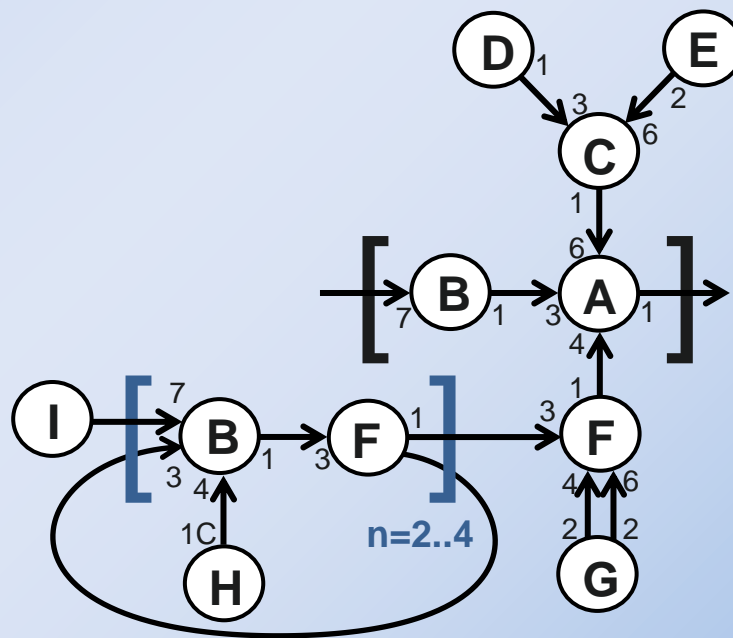
углеводный язык  
**CSDB Linear**

аннотаторы,  
операторы,  
другие языки  
(IUPAC, GlycoCT, WURCS, ...)

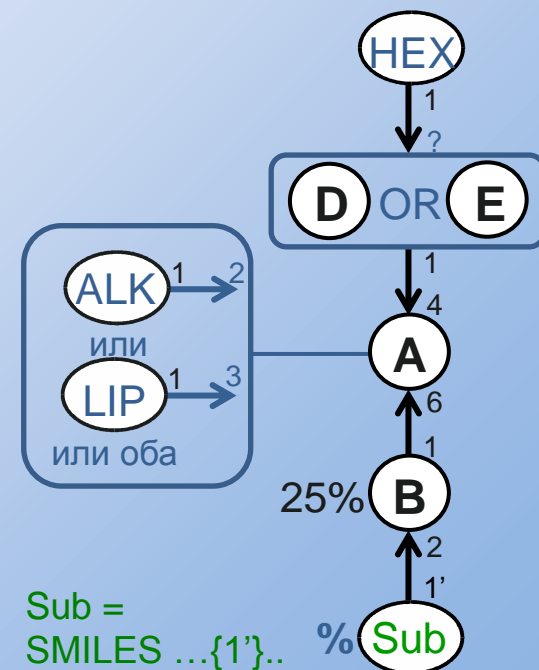
**МОНОМЕРЫ**



**ТОПОЛОГИЯ**



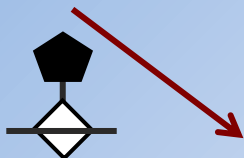
**НЕТОЧНОСТИ**



# Уровни абстракции

-3) [xLAla (2-6) ,Ac (1-2) ]bDGalpNA (1-

*точный фрагмент и его связи*

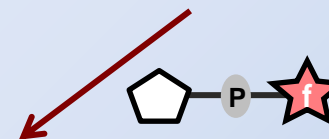


L-Ala (2-6) +

-5) α-D-Ribf(1-3) β-D-GalpNAcA (1-2) D-Gro (1-P-

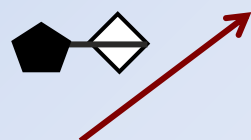
xDGro (1-P-5) aDRibf

*не указана топология и места присоединения*



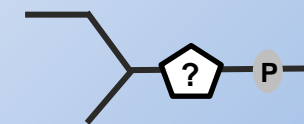
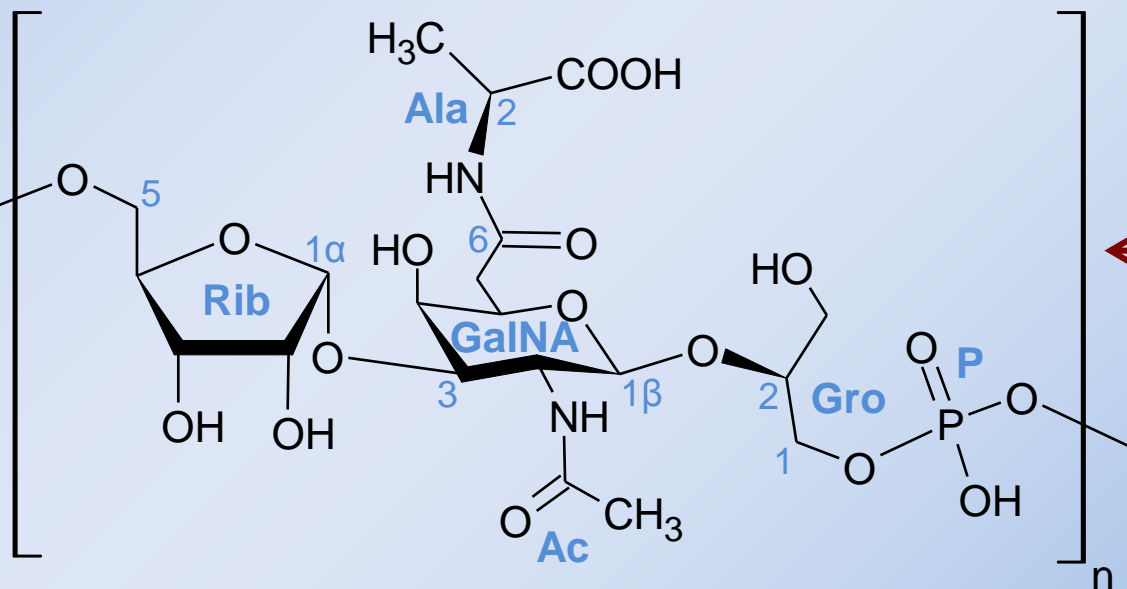
-?) ?Dhex (1-

*указаны только класс и 1-связь*



x?Ala (2-?) ?DGal?NA

*не указаны конфигурации, позиции замещения, размер цикла, N-ацетилирование*



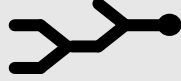






















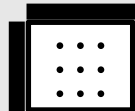
















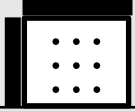








*приблизительный мотив*



HEX, xDRib?, PEP

*частичный состав*

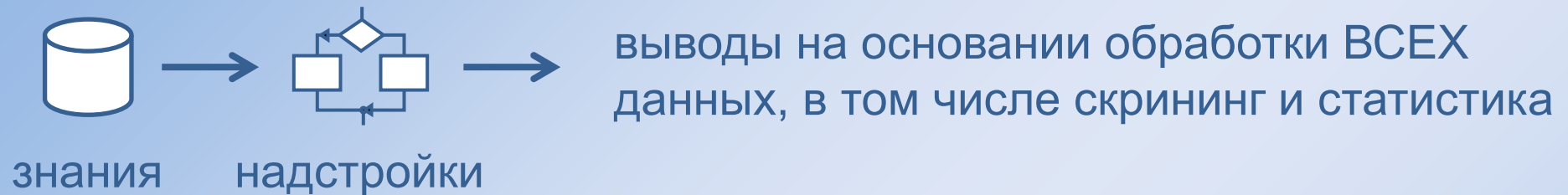
# Сравнение языков

	<i>подход</i>	полнота	однозначность	контроль	парсинг	неточные структуры
IUPAC	 					
IUPAC extended (SweetDB, Carbbank)	pseudo-graphics 					
Glyde I	 XML 				 URL	
WURCS (JCGGDB, ChEBI, PDB)					 URL	
GlycoCT (Glycome-DB)						
LinearCode (CFG)					 URL	
LinUCS (GlycoSCIENCES)					 URL	
KCF (KEGG)						
CSDB linear (CSDB)					 URL	

 СОВМЕСТИМОСТЬ

хуже     лучше

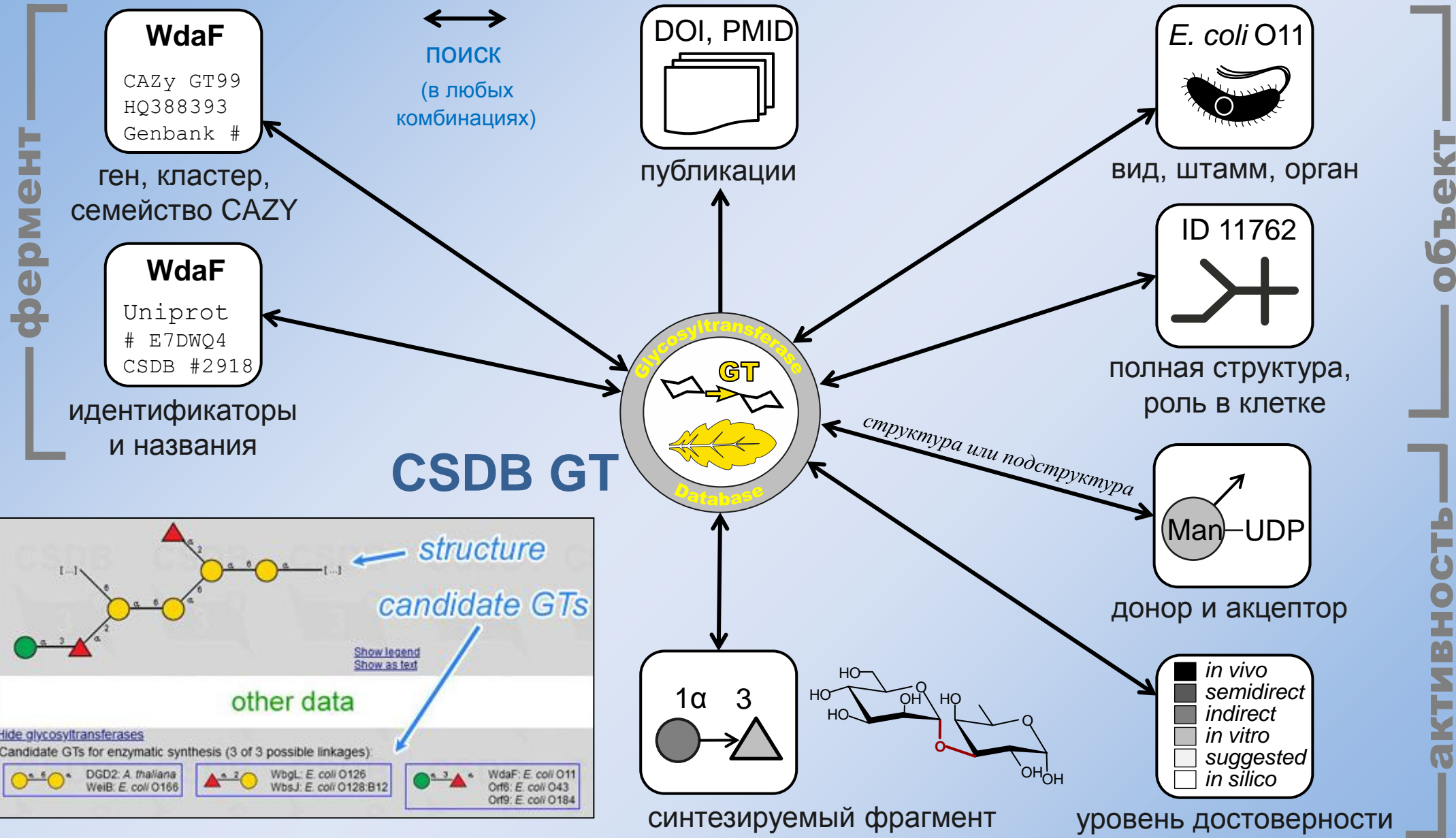
# CSDB: надстройки



- Анализ путей биосинтеза (база гликозилтрансфераз)
- Конформационные карты олигосахаридов
- Предсказание и отнесение спектров ЯМР  $^{13}\text{C}$ ,  $^1\text{H}$ , 2D
- Предсказание структуры по спектрам и другим данным
- Кластеризация таксонов на основании их гликомов
- Распределение фрагментов по таксонам и положению в структурах
- Классификация мономеров



# Гликозилтрансферазы



Egorova K, Toukach Ph **CSDB\_GT: a new curated database on glycosyltransferases** *Glycobiology* 2017, 27:285-290

Egorova K, Toukach Ph **Expanding CSDB\_GT glycosyltransferase database with *Escherichia coli*** *Glycobiology* 2019, 29:285-287

Egorova K, Smirnova NS, Toukach Ph **CSDB\_GT, a curated glycosyltransferase database with close-to-full coverage on three most studied non-animal species** *Glycobiology* 2021, 31:524-529



# Гликоферментные базы



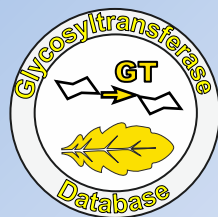
300K+ ферментов CAZy, подтверждено ~4%,  
*все организмы*  
 (полноценное использование – через соавторство)



гликозилтрансферазы, подтверждено 7%  
*E. coli* (полная)



гликозилтрансферазы и др. CAZy  
*O. sativa*



гликозилтрансферазы, подтверждено 80%  
*E. coli, A. thaliana, S. cerevisiae, A. baumannii* (ESKAPE – в процессе)  
 (полная по подтвержденным)



биосинтетическая информация среди прочей  
*все организмы*



биосинтетическая информация среди прочей  
*A. thaliana*



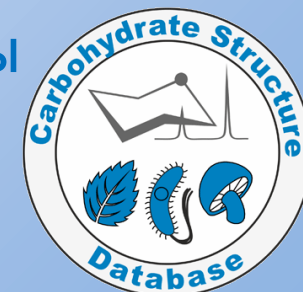
биосинтетическая информация, нет поиска  
*H. sapiens, M. musculus, R. norvegicus, ...*



углеводные гены,  
*H. sapiens, C. elegans*

# Доступ к 3D-структурам

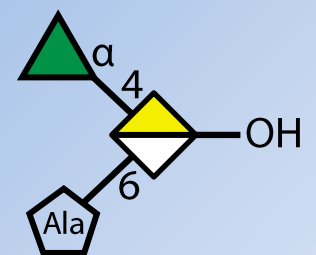
- Большинство баз = 3D-структуры гликанов **млекопитающих**  
(как часть гликопротеинов)
  - <5% статей содержат 3D-данные
  - Структуры симулированы в **разных** условиях
  - **Гликополимеры** полностью отсутствуют
  - Каждая симуляция превращается в отдельное исследование
- Нам нужен инструмент «из коробки», для не-информатиков
- → стандартизованные модели структур в растворе
  - → автоматическая генерация, обширные пред-расчеты
  - → экспорт в атомарные форматы



# Анализ конформаций

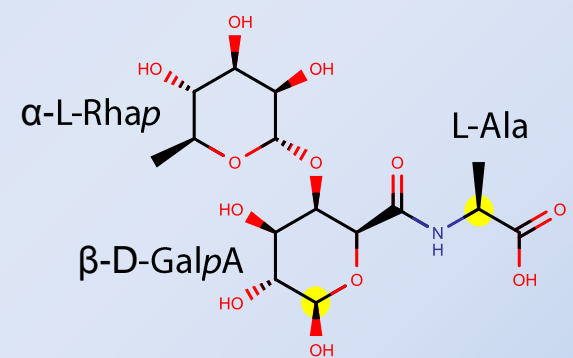
aLRhap(1-4)[x?Ala?(2-6)]?DGalpA

структуры,  
в т.ч. неполные

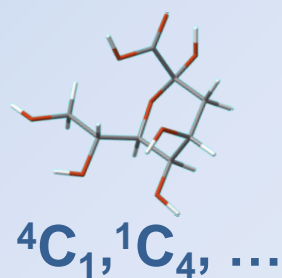


другие варианты  
(α-GalA, D-Ala, и т. д.)

SMILES

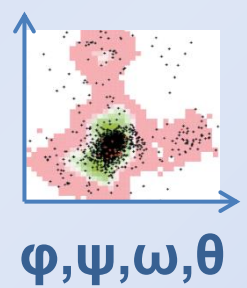
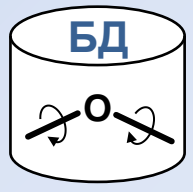


выгодные  
конформеры  
~1000 остатков

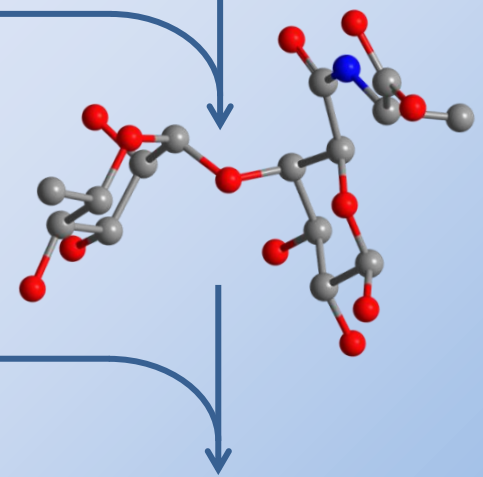


«креслификация»

заселенные  
состояния  
мостиков



выбор минимумов  
ММ-релаксация



мол. динамика  
300К, 100нс, H<sub>2</sub>O

конформеры  
+ их энергии



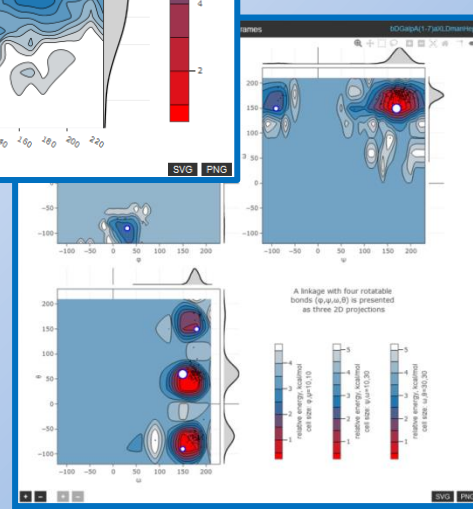
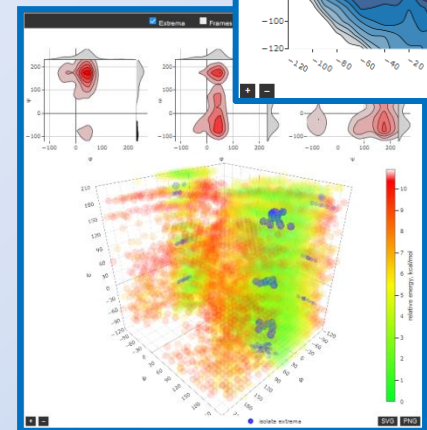
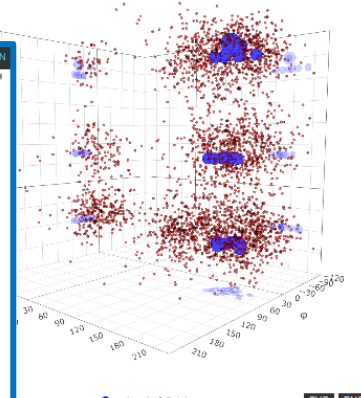
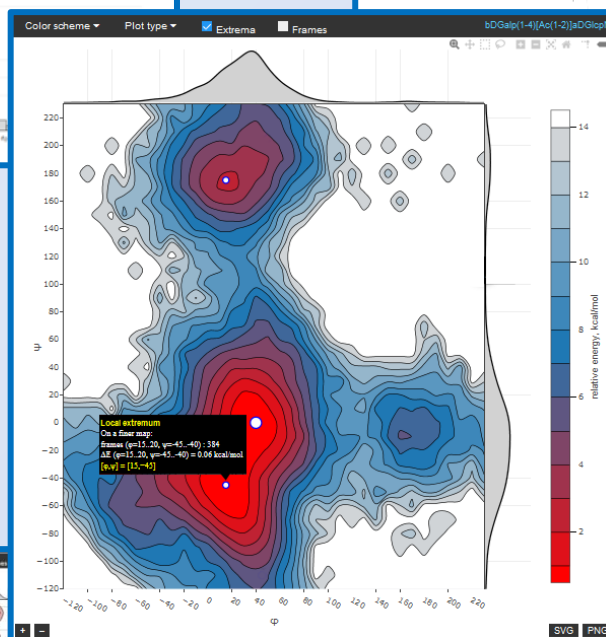
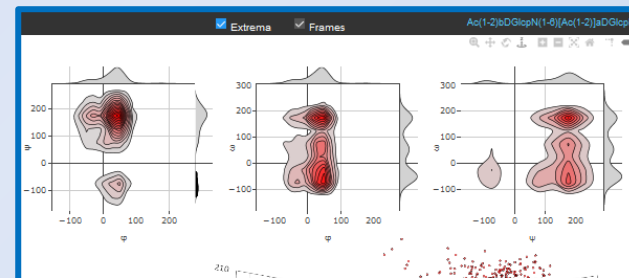
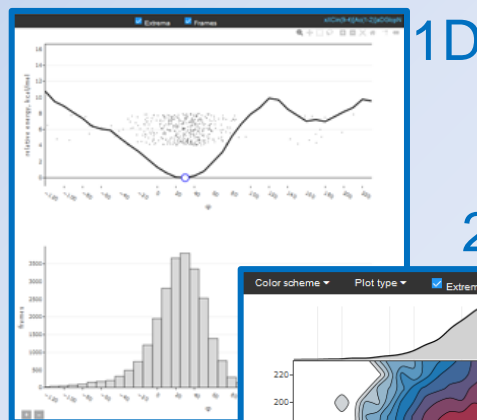
# Конформационный модуль CSDB

~2700 ди- и трисахаридов:  
100-нс MD в явной H<sub>2</sub>O, 300K

- поиск ID, структур, параметров расчета

CSDB conformation data search  
73 conformation maps have been found.

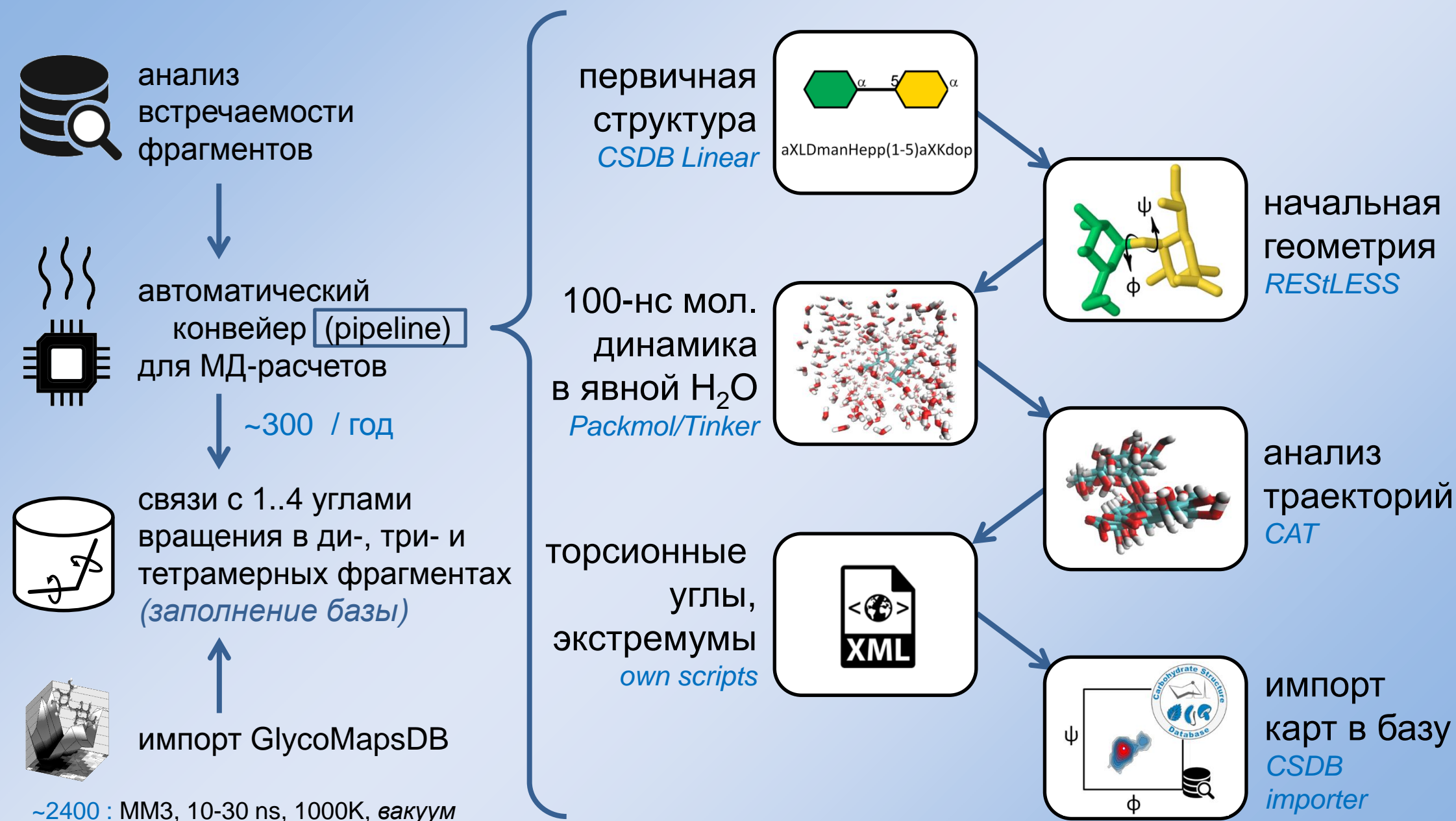
Model structure	Conformation map	Energy minima	Details
		$\varphi=40, \psi=0 \Delta E=0.00 \text{ Kcal/mol}$ $\varphi=20, \psi=40 \Delta E=0.53 \text{ Kcal/mol}$	Force field: MM3-2000 Solvent model: Tip3P MD temperature: 300 MD duration: 100 ns Frames: 50K MD summary file: <a href="#">download</a>
		$\varphi=40, \psi=175, \omega=60 \Delta E=0.00 \text{ Kcal/mol}$ $\varphi=30, \psi=165, \omega=60 \Delta E=0.00 \text{ Kcal/mol}$ $\varphi=30, \psi=165, \omega=180 \Delta E=0.10 \text{ Kcal/mol}$ $\varphi=40, \psi=185, \omega=60 \Delta E=0.32 \text{ Kcal/mol}$ $\varphi=25, \psi=150, \omega=180 \Delta E=0.44 \text{ Kcal/mol}$ $\varphi=40, \psi=185, \omega=165 \Delta E=0.71 \text{ Kcal/mol}$ $\varphi=40, \psi=195, \omega=180 \Delta E=0.71 \text{ Kcal/mol}$ $\varphi=55, \psi=185, \omega=60 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=40, \psi=185, \omega=45 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=40, \psi=150, \omega=180 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=30, \psi=180, \omega=45 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=15, \psi=170, \omega=165 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=50, \psi=175, \omega=165 \Delta E=0.86 \text{ Kcal/mol}$ $\varphi=25, \psi=155, \omega=60 \Delta E=1.01 \text{ Kcal/mol}$ $\varphi=40, \psi=210, \omega=60 \Delta E=1.01 \text{ Kcal/mol}$	Force field: MM3-1996 Solvent model: None MD temperature: 1000 MD duration: 30 ns Frames: 30K MD summary file: <a href="#">download</a>



- работа с картами энергий и заселенности
- проекции, экстремумы, экспорт
- до 4 углов / связь
- ИНТЕРАКТИВНОСТЬ (прокрутка, масштаб, вращение, контроль плотности, цвет, слои)

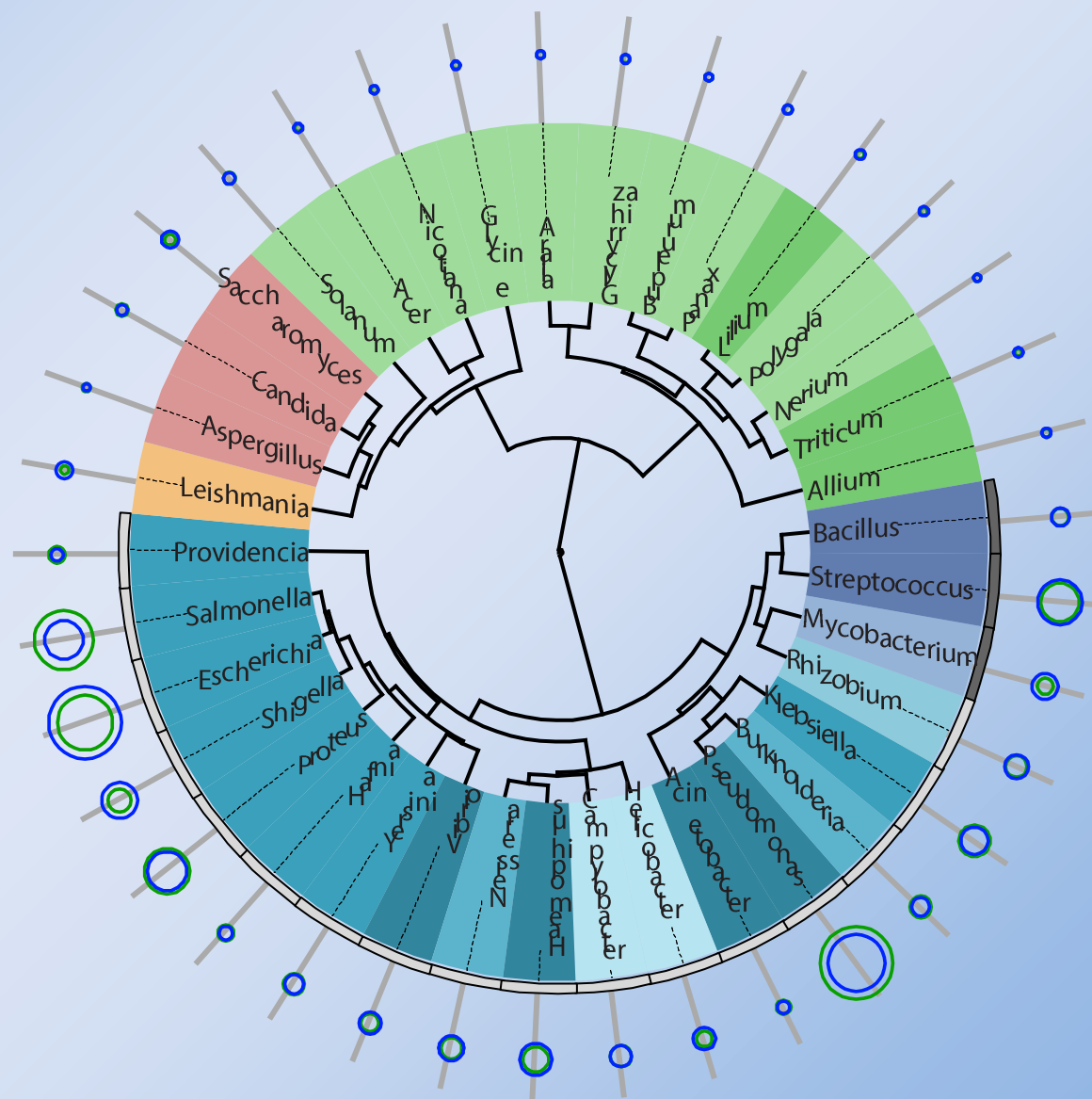
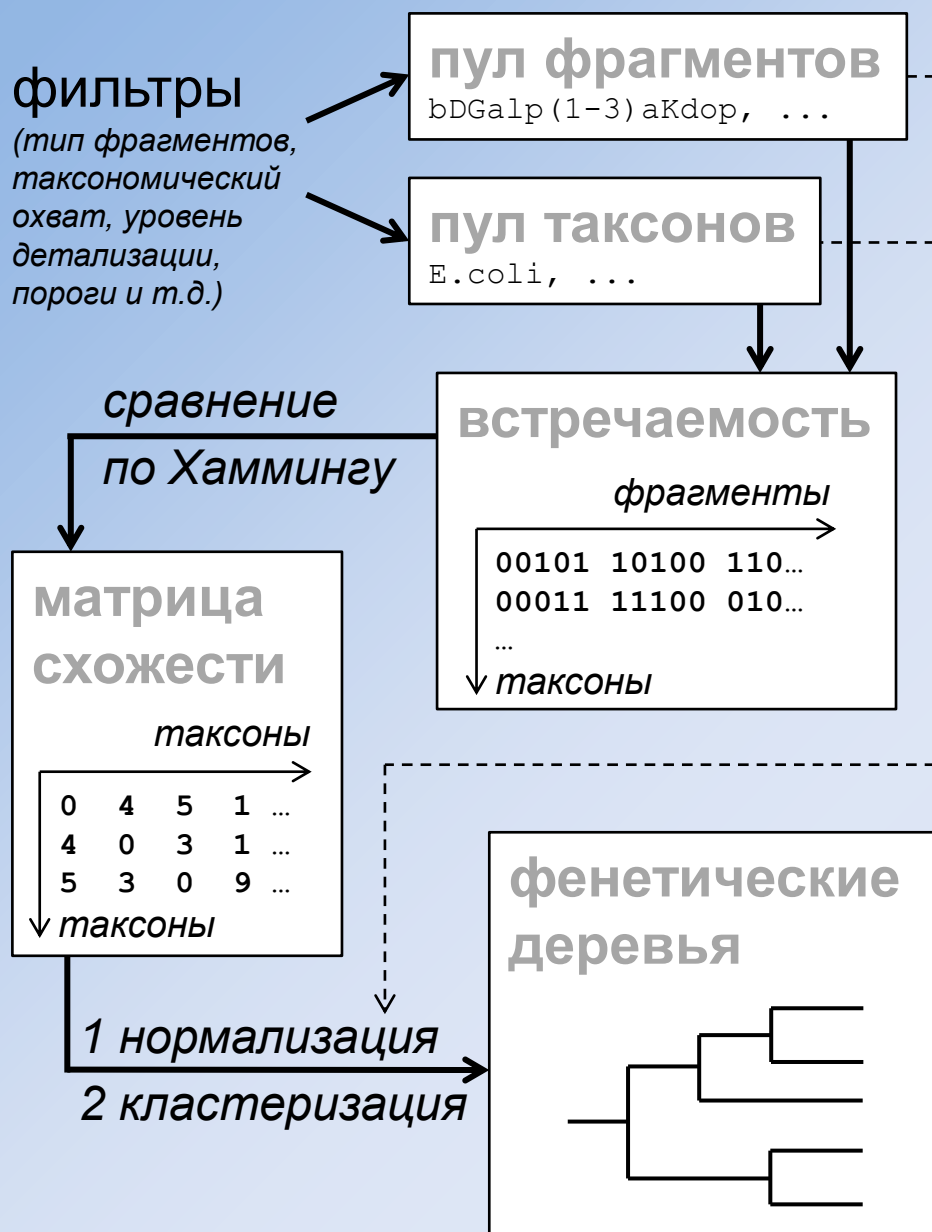


# Заполнение базы





# Кластеризация таксонов



ассоциированные структуры и организмы

бактерии: Грам+ Грам-

ЯМР – основной метод анализа первичной структуры в гликобиологии

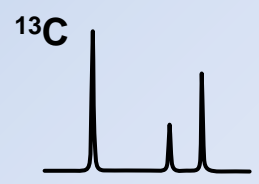


- Моделирование спектров ( $^{13}\text{C}$  ~ 0.7 ppm,  $^1\text{H}$  ~ 0.06 ppm)
- Помощь эксперту в установлении структуры
- Отнесение сигналов и проверка гипотез
- Перебор структур и сравнение симуляции с экспериментом
- Верификация расчетной геометрии молекул

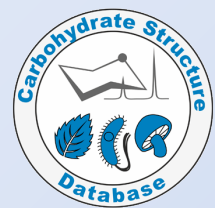
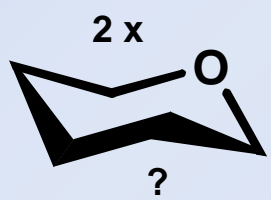
# Итератор структур



неотнесенный спектр ЯМР

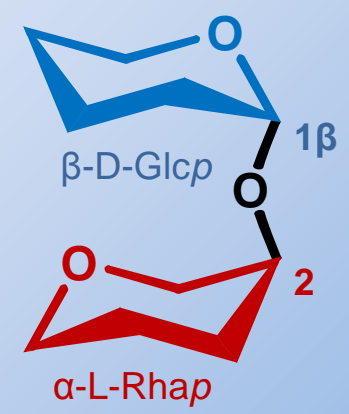


известные данные о структуре



алгоритм GRASS

Generation, Ranking and Assignment of Saccharide Structures



полная структура (оставшиеся неизвестные)

мономеры, конфигурации, модификации, позиции замещения, последовательность

# Дальнейшее развитие

41

● сделано в CSDB    ● предстоит сделать    ● близко к завершению

- **Стандартный человекочитаемый язык** (SNFG, CSDB Linear, ...)
- **Расширение онтологий** (интеграция через GlycoRDF и GlycoCoO)
- **Кросс-проектные сервисы**  
(ввод-вывод структур, конформационные расчеты, предсказание свойств)
- **Стандартные индексы**  
(Glytoucan ID, MSDB, PMID, DOI, TaxID, ICD-11, PDB id, Genbank, ...)
- **Стандартные протоколы** (API, WSDL, SPARQL, REST, ...)
- **Идеологическая замена CarbBank** ~~\$~~
- **Требование включать ID в публикации** (Glytoucan ID?)  
(кто будет чистить базы от неопубликованных/ошибочных данных?)

# Как воспользоваться?

42

S. Scherbinina, Ph. Toukach **3D structures of carbohydrates and where to find them** (2020) *Int J Mol Sci* **21**, ID 7702. doi: 10.3390/ijms21207702

Abrahams J. et al. **Recent advances in glycoinformatic platforms for glycomics and glycoproteomics** (2020) *Curr Opin Struct Biol* **62**, 59-69. doi: 10.1016/j.sbi.2019.11.009

K.F. Aoki-Kinoshita **A practical guide to using glycomic databases** (2017) Springer. doi: 10.1007/978-4-431-56454-6

Ph. Toukach, K. Egorova **Carbohydrate Structure Database merged from bacterial, plant and fungal parts** (2016) *Nucl Acid Res* **44**, D1229–36. doi: 10.1093/nar/gkv840



<http://glytoucan.org>



[http://jcggdb.jp/index\\_en.html](http://jcggdb.jp/index_en.html)

T. Lütke **The use of glyco-informatics in glycochemistry** (2012) *Beilstein J Org Chem* **8**, 915-929. doi: 10.3762/bjoc.8.104



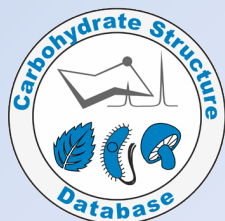
<http://glycosciences.de>



<http://www.genome.jp/kegg/glycan/>



<http://www.unicarbkb.org/>



<http://csdb.glycoscience.ru>



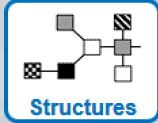
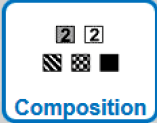


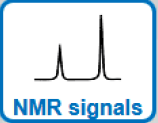
<http://toukach.ru/rus/glyco-db.htm>





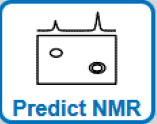
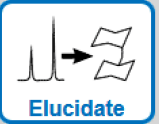
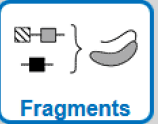
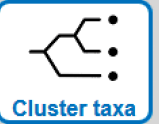


# Сайт в Интернете

**Database search**

 Structures
  Composition
  Organisms
  Publications
  NMR signals

Additional operations are available from the [left menu](#). If you don't see it [click here](#)

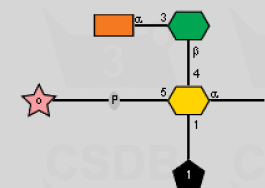
**Useful tools**

 Predict NMR
  Elucidate
  Fragments
  Cluster taxa
  GT activities
  Examples

## NMR spectrum simulation

Please, select how to input a structure:

- [Input using Structure Wizard](#)
- [Select from library](#)
- [Draw in Glycan Builder](#)
- [Convert from GlycoCT](#)
- [Use expert form \(field below\)](#)



Structure in CSDB encoding:

`aXAbep(1-3)bXLDmanHepp(1-4)[xDRib-ol(1-P-5),xLAla?(2-1)]aXKdop`

(this field is editable) [Help on structure encoding](#)

Nucleus: 1H/13C (2D)  More parameters...

Solvent: Water (H or D)  Coverage

**Carbohydrate Structure Database**

Prokaryotes » Plants » Fungi

7005 publications (1941-2017):  
18923 compounds from  
8859 organisms  
last update: 2017 Jun 2

**Search**

- CSDB IDs
- (Sub)structure
- Composition
- Taxonomy
- Bibliography
- NMR signals

**Help**

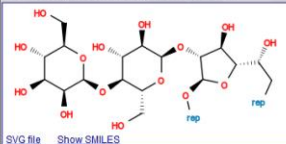
**Extras**

- NMR simulation
- Elucidation from NMR
- Monomer namespace
- Fragment abundance
- Coverage stats
- Taxon clustering
- Submit record
- Translate structure
- Feedback

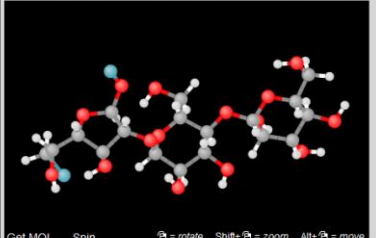
**Maintenance**

Related record ID(s): 101  
NCBI Taxonomy refs (TaxIDs): 64489

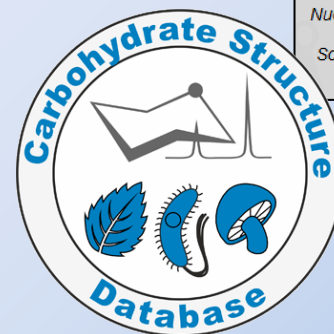
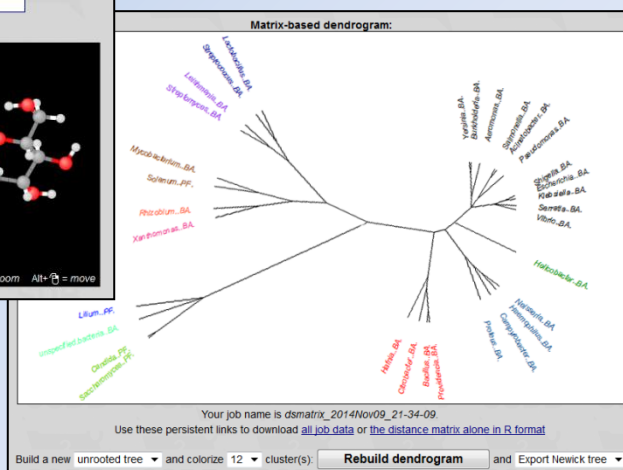
There is only one chemically distinct structure:



SVG file Show SMILES



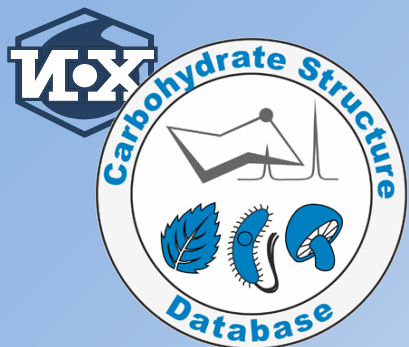
Get MOL Spin rotate Shift+ zoom Alt+ move



<http://csdb.glycoscience.ru>

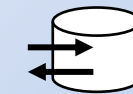
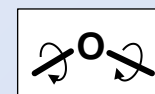
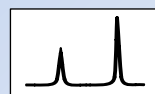
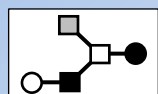
- свободный доступ
- подробная документация
- примеры решения задач

# Участники и спонсоры



База данных природных углеводов

Платформа для сервисов гликоинформатики



программирование

аннотирование и проверка данных

общая поддержка и сбор данных


интеграция, онтология

конформационный анализ

идеи, язык, архитектура, интерфейс,  
программирование, координация


партнеры

  Роман Капаев, Андрей Бочков, Иван Чернышов, ...

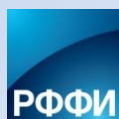
 Ксения Егорова, Надежда Калининчук, Кирилл Казанцев, ...

 Юрий Книрель

  Рене Ранцингер, Кийоко Аоки-Киношита, Томас Люттеке, ...

 Виктор Стройлов, Софья Щербинина, ...

 Филипп Тоукач



Российский Фонд  
Фундаментальных  
Исследований

2005-2020 (x4)



Международный  
Научно-Технический  
Центр

2004-2005



Комиссия по  
грантам при  
президенте РФ

2005-2007



Немецкий Центр  
Исследования  
Рака

2007-2011 (x4)



Фонд Содействия  
Отечественной  
Науке

2008-2009



Российский  
Научный  
Фонд

2018-2022 (x2)

# **Дополнения**

(демонстрируются в рамках ответов на вопросы)

# Примеры вопросов к CSDB

- Как введение аминогруппы влияет на химические сдвиги ЯМР в лактозном фрагменте?
- Какие структуры О-антигенов, содержащие галактуроновую кислоту и еще как минимум одну гексозу, были опубликованы после 2005-го года?
- Какие гликозиды, выделенные из растений рода паслёна, содержат агликон соланидин?
- Какие углеводы, кроме октозосодержащих, имеют сигнал ЯМР  $^{13}\text{C}$  около 34 м.д. ?
- Какие бактериальные структуры, опубликованные А.С. Шашковым или Ю.А. Книрелем, содержат хиновоз-4-амин, амидированный любой N-ацетилированной аминокислотой?
- Гомополимеры каких нонапираноз встречаются в бактериях?
- Каков ожидаемый спектр ЯМР  $^{13}\text{C}$  3-О- $\alpha$ -абеквозил-6-деокси- $\beta$ -D-манногептопиранозил-(D-рибитол-1)-фосфата в воде и на основании каких источников предсказаны химические сдвиги, для которых указана наименьшая достоверность?
- Какова наиболее вероятная последовательность остатков бациллозамина, глюкуроновой кислоты и лизина в олигомере с указанным экспериментальным спектром ЯМР  $^{13}\text{C}$ ?
- Какие моносахариды склонны занимать концевые позиции в гликанах Аспергилла дымящего и Аспергилла кодзи?
- Какие димерные фрагменты (включая дисахариды) гликанов высших растений специфичны для рода люпинов?
- Для скольких гликанов протеобактерий опубликованы спектры ЯМР?



# Resource Description Framework

47

**RDF – модель данных в виде триплетов *объект-предикат-субъект*.**

- 😊 допускает распределенные запросы с минимальным знанием форматов баз
- ☹️ требует репозитория триплетов и согласованной онтологии

**Задача:** найти белок-носитель для произвольного гликана из JCGGDB.

**Проблема:** JCGGDB не связана ссылками с белковыми базами.

**Преамбула:**

Записи в JCGGDB имеют ссылки на идентификаторы в GlycomeDB.

Как GlycomeDB, так и UniCarbKB могут экспортировать структуры в формате GlycoCT.

Записи в UniCarbKB имеют ссылки в белковую базу UniProt.

**Решение** (*9-строчный скрипт на SPARQL*):

Сопоставить идентификаторы JCGGDB и UniCarbKB, используя GlycomeDB, и получить идентификаторы UniProt из UniCarbKB для каждого идентификатора JCGGDB.

**Требуется:**

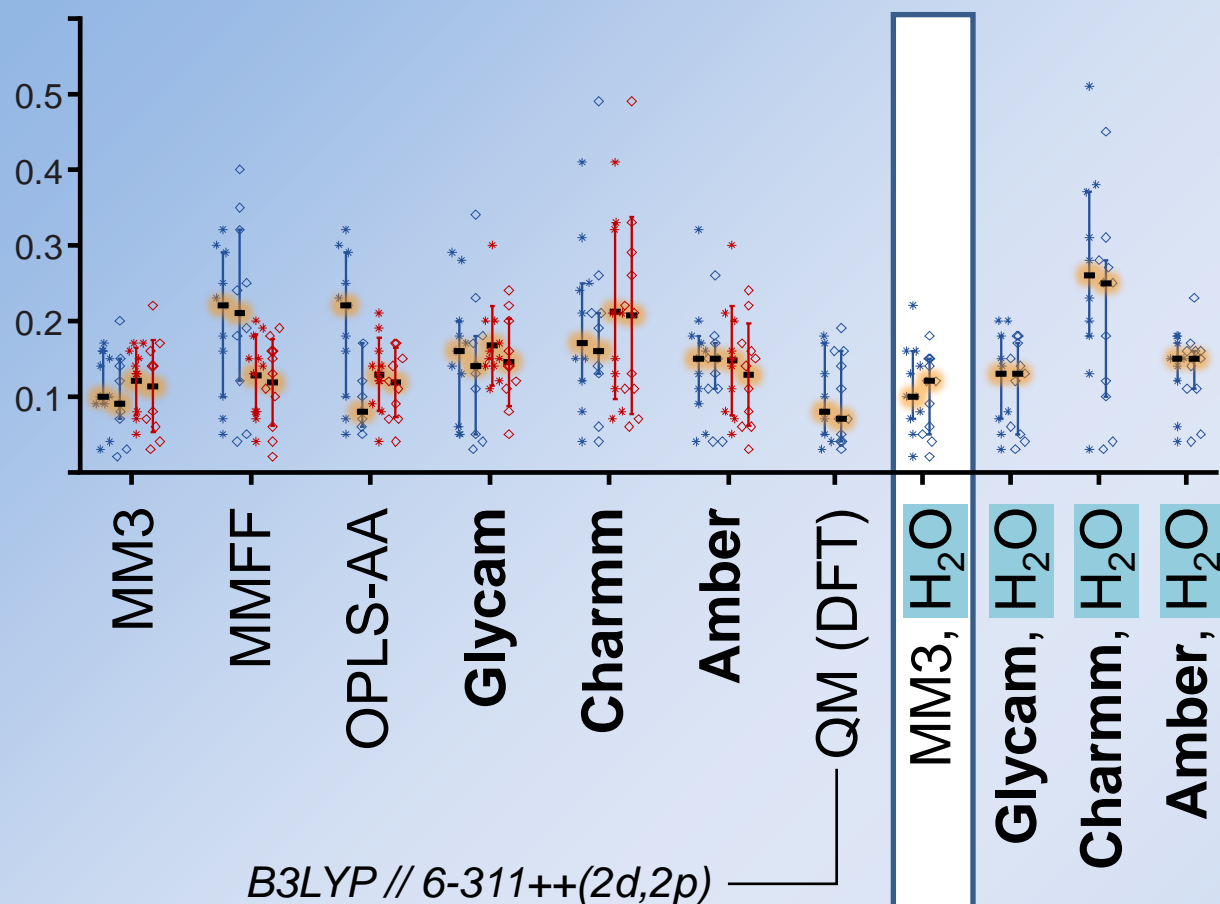
стандартная онтология → экспорт данных в RDF → репозиторий триплетов → интерфейс SPARQL

**GlycoRDF – первая формальная углеводная онтология (OWL)**



# Метод расчета

NOE, RMSD



выборка:

11 дисахаридов разных классов

\* все H-H NOEs, 300K

◇ трансгликозидные NOEs, 300K

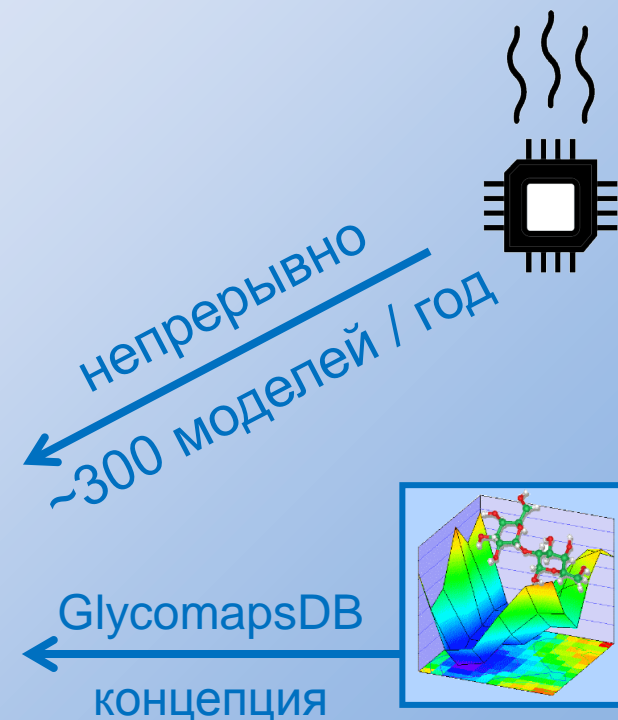
\* все H-H NOEs, 1000K

◇ трансгликозидные NOEs, 1000K

B3LYP // 6-311++(2d,2p)

~200 моделей: MM3, 100+ ns, 300K, явная H<sub>2</sub>O

~2400 моделей: MM3, 10-30 ns, 1000K, вакуум



# Ввод и вывод структуры

CSDB/SNFG structure editor

Popular Small sugars Hexoses Higher sugars Alditols Aliphatic acids Other acids Superclasses

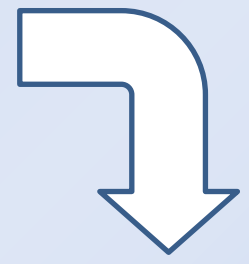
Glc GlcNAc GlcA QuiNAc Gal GalNAc GalA Fuc FucNAc Man Rha LDmanH Ara Ara4N Xyl Fru P Kdo Neu5Ac Gro Ala

Novice Expert Insert Replace Oligo Poly Ac Am Cm Cho Fo Me Et Pr EtN Allyl Bz P S Pyr NH2

search residues search modifications

Chemical repeating unit; n=10

-3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]?DGlcp(1-?)[Ac(1-2)]bDGalfN(1- // Subst = chrysin = SMILES O=c2cc(

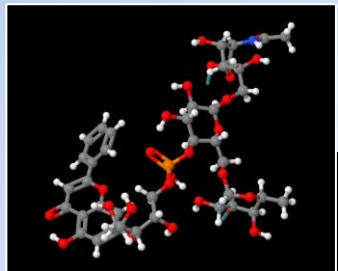


Previews Refresh

RES  
1r:r1  
REP  
REP1:6o(3+1)2d=-1--1  
RES  
2b:b-dgal-HEX-1:4  
3b:x-dglc-HEX-1:5

Subst-(7-3)-D-Rib-ol-(1--P--4)--+  
|  
-3)-a-L-Fucp-(1-6)-D-Glcp-(1-?)-b-D-GalfNAC-(1-

Subst = chrysin = SMILES O=c2cc(c1ccccc1)oc3c{7}c(0)cc(0)c23



REMARK 300 USING GENERATED NAME Fda FOR RESIDUE SUB:(Ac,2)bdGalFN  
REMARK 300 USING GENERATED NAME Fda FOR RESIDUE xSubst

AUTHOR GENERATED BY OPENSABEL 2.4.1, RESIDUE ASSIGNMENT BY CARBOHYDRATE STRUCT

CONFND	CSDB linear =									
1	R	Fda	A	1	0.156	1.111	2.494	1.00	0.00	XX
2	O1	Fda	A	1	1.377	1.487	3.097	1.00	0.00	O
3	C1	Fda	A	1	1.643	2.889	2.961	1.00	0.00	O
4	O4	Fda	A	1	1.951	3.280	1.601	1.00	0.00	O
5	O4	Fda	A	1	3.374	3.123	1.402	1.00	0.00	C
6	O5	Fda	A	1	3.933	4.392	3.736	1.00	0.00	O
7	O5	Fda	A	1	3.890	5.497	1.642	1.00	0.00	O
8	O6	Fda	A	1	3.185	4.772	-0.562	1.00	0.00	C
9	O6	Fda	A	1	3.501	3.900	-1.665	1.00	0.00	O
10	O3	Fda	A	1	4.005	2.825	2.771	1.00	0.00	C
11	O3	Fda	A	1	4.252	1.419	2.915	1.00	0.00	O
12	O2	Fda	A	1	2.912	3.229	3.759	1.00	0.00	C
13	H2	Fda	A	1	3.019	4.689	4.105	1.00	0.00	O
14	H1	Fda	A	1	0.780	3.467	3.305	1.00	0.00	H
					2.260	0.749	1.00	0.00	H	
					4.211	0.481	1.00	0.00	H	
					6.274	1.168	1.00	0.00	H	
					5.771	-0.887	1.00	0.00	H	
					4.869	-0.431	1.00	0.00	H	
					3.336	2.931	1.00	0.00	H	
					1.146	2.209	1.00	0.00	H	
					2.647	4.689	1.00	0.00	H	
					5.266	3.368	1.00	0.00	H	
					5.202	5.040	1.00	0.00	H	
					6.697	5.164	1.00	0.00	H	
					4.543	5.689	1.00	0.00	H	
					6.973	6.220	1.00	0.00	H	
					7.056	4.765	1.00	0.00	H	
					7.165	4.614	1.00	0.00	H	
					2.820	-1.808	1.00	0.00	C	
					3.154	-2.751	1.00	0.00	O	
					3.270	-4.096	1.00	0.00	C	
					3.738	-4.973	1.00	0.00	C	

There are 3 chemically distinct structures. Please, select:

1. -3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]?DGlcp(1-3)[Ac(1-2)]bDGalfN(1- // Subst
2. -3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]?DGlcp(1-5)[Ac(1-2)]bDGalfN(1- // Subst
3. -3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]?DGlcp(1-6)[Ac(1-2)]bDGalfN(1- // Subst

SMILES  
[\*]O[C@@H]1O[C@@H]([C@@H](O)COC2O[C@H](CO[C@@H]3O[C@@H](C)[C@@H](O)[C@@H]([\*])[C@@H]3O)[C@@H](OP(=O)(O)OC[C@H](O)[C@@H](Oc3cc(O)c4c(=O)cc(-c5ccccc5)oc4c3)[C@H](O)CO)[C@H](O)[C@@H]2O)[C@@H](O)[C@@H]1NC(C)=O

There are 2 sterically distinct structures. Please, select:

1. -3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]aDGlcp(1-6)[Ac(1-2)]bDGalfN(1- // Subst
2. -3)aLFucp(1-6)[Subst(7-3)xDRib-ol(1-P-4)]bDGlcp(1-6)[Ac(1-2)]bDGalfN(1- // Subst

Files: MOL, PDB, Glycam

3D Shift+Z = zoom Shift+X = pan Alt+R = rotate Ctrl+M = menu

### Structure wizard

Topology: 3 residues (linear: A->B->C) (A)→(B)→(C)

Structure:

Residue (A):

()

[aLFucp](#)  
substitutes  of Residue B

is terminal

add substitution  
 add substituent Acetylated at   
 add substitution  
 add substitution  
 add substitution

Residue (B):

()

[DRib-ol](#)  
substitutes  of Residue C

add substitution  
 add substitution  
 add substitution  
 add substitution

Residue (C):

()

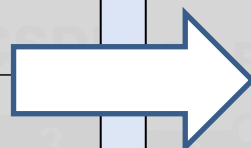
[a?6dTal?](#)  
 has aglycon:

add substitution  
 add substitution  
 add substitution  
 add substitution

**Structure in CSDB encoding:**

[Return the structure to the search page and close this window](#)

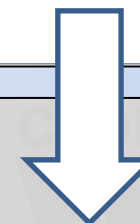
[Home](#) [Help](#)



### Glycan Builder

File Edit Structure View Help

Linkage    Chirality  Ring



### Search for (sub)structure

Please, select how to input structure:

- [Input using Structure Wizard](#)
- [Select from library](#)
- [Draw in Glycan Builder](#)
- [Convert from GlycoCT](#)
- [Copy from the previous query \(aLFucp3N\)](#)
- [Use expert form \(field below\)](#)

**Structural fragment in CSDB encoding:**

*(this field is editable)* [Help on structure encoding](#)

Only those containing text:   in aglycons, aliases or linear code  in trivial names

**Search scope:**

Search the whole database  
 Search in the result of the previous query (logical AND)  
 Combine with the result of the previous query (logical OR)  
 Negate search (find results NOT matching current query)

Treat search term as a   
Search for molecule types:   
 Search for structures with published NMR data only  
 Restrict compound class:   
 Restrict taxonomical domain:

Previous results: 122 structures: [<ID list>](#)

& display  records per page.

[Predict NMR](#) [Sweet 3D model](#) [Home](#) [Help](#) [HELP !!!](#)



# Поиск по таксономии

Found **12** organisms. Displayed organisms from **1** to **12**

[Expand all organisms](#)   [Show all as text \(SweetDB notation\)](#)

---

**1.** (Organism ID: 1005)

***Acetobacter xylinum***  
 (Ancestor NCBI TaxID 28448, [species name lookup](#))

Later renamed to: [Komagataeibacter xylinus](#)  
 Taxonomic group: bacteria  
 Phylum: Proteobacteria

**The following compound(s) are assigned to this organism:**

- Compound ID: 1717

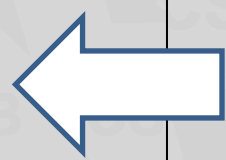
[Show legend](#)  
[Show as text](#)

Carbohydrate Research 2004, "Synergistic interactions between the genetically modified bacterial polysaccharide P2 and carob or konjac mannan"  
[CSDB ID 9262](#) (all data & tools)

- Compound ID: 1720

[Show legend](#)  
[Show as text](#)

Carbohydrate Research 2004, "Synergistic interactions between the genetically modified bacterial polysaccharide P2 and carob or konjac mannan"  
[CSDB ID 9414](#) (all data & tools)



## Search for organism

**Display domains:**  bacteria  archaea  protista  algae  fungi  plants  animals

---

**Genus:**

**Species:**

**Strain / subspecies:**

Specify:

---

**Search scope:**

Search the whole database    Search among HOST organisms  
 Search in the result of the previous query (logical AND)    Use NCBI taxID  
 Combine with the result of the previous query (logical OR)    Include subtaxons  
 Negate search (find results NOT matching current query)

Previous results: 6 structures: [<ID list>](#)

& display  records per page.

[List of organisms](#)   [Home](#)   [Help](#)

---

**Process taxonomy in NCBI Taxonomy DB** (fields are editable):

Genus:  Species:

# Поиск по библиографии

Found 3 publications. Displayed publications from 1 to 3

[Expand all publications](#) [Show all as text \(SweetDB notation\)](#)

1. (Article ID: 1525)

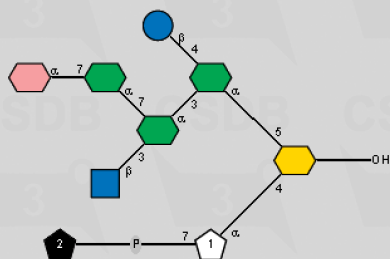
Knirel YA, Lindner B, Vinogradov EV, Shaikhutdinova RZ, Senchenkova SN, Kocha  
**Cold temperature-induced modifications to the composition and structure of Yersinia pestis lipopolysaccharide**  
*Carbohydrate Research* **340(9)** (2005) 1625-1630

Following a report of variations in the lipopolysaccharide (LPS) structure of *Y. pestis* at 10 degrees C and flea (25 degrees C) temperatures, a number of changes to the LPS of the bacterium was observed when the bacterium was cultivated at a temperature of winter-hibernating rodents (6 degrees C). The LPS of the latter differs in: (i) replacement of terminal galactose with terminal mannose; (ii) phosphorylation of terminal oct-2-ulosonic acid with phosphoethanolamine; (iii) the absence of glycine; lipid A differs in the lack of any 4-amino-4-deoxyoctanoic acid side chain (di)oxylation of a fatty acid(s). The data obtained suggest that cold temperature-induced modifications to the mechanism of control of the synthesis of *Y. pestis* LPS.

*Lipopolysaccharide, structure, core, modification, agent, composition, Yersinia pestis, Plague*

The publication contains the following compound(s):

• Compound ID: 4209

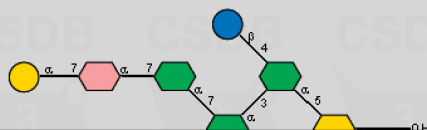


1 = a-Kop  
 2 = EtN

[Show legend](#)  
[Show as text](#)

*Yersinia pestis* KM218  
[CSDB ID 10076](#) (all data & tools)

• Compound ID: 4210



## Search for bibliography

**Authors:** "Knirel YA" OR Toukach [Author index](#) start with: tou

[Help on author/keyword query syntax](#) ä ö ü á é í ó ç š

**Title:** pestis OR plague\*  search also in abstract

(content of title) [Help on title/abstract query syntax](#)

**Keywords:** structure? OR composition?  search also in title

(content of keyword section) [Help on author/keyword query syntax](#)

**Journal:** Carbohydrate Letters  
 Carbohydrate Polymers  
**Carbohydrate Research**  
 Cell  
 Cell Chemical Biology  
 Cell and Tissue Research

**Year:** 1983  
 1984  
**1985**  
 1986  
 1987  
 1988  
 1989

**Vol:** \*  
**Page:** \*

### Search scope:

- Search the whole database
- Search in the result of the previous query (logical AND)
- Combine with the result of the previous query (logical OR)
- Negate search (find results NOT matching current query)
- Publications with structure elucidation only
- Restrict taxonomical domain: All domains

Previous results: 3 publications: 2953,201,1525

[Go!](#) & display 30 records per page.

[PubMed XML](#) [Home](#) [Help](#)

### Author index:

[Toubetto K](#) [Toussaint A](#)  
[Toukach FV](#)

The listed author names start with 'Tou'.  
 Click an author name to copy it to the author field in the caller form.

[Close this window](#)



# Поиск конформаций

## Search for disaccharide conformation maps

Use the following criteria alone or in any combination to search for conformation maps.

Conformation ID:

Model:        
*(only those components are listed for which conformation maps are stored)*

or type dimeric fragment in CSDB encoding

Force field:

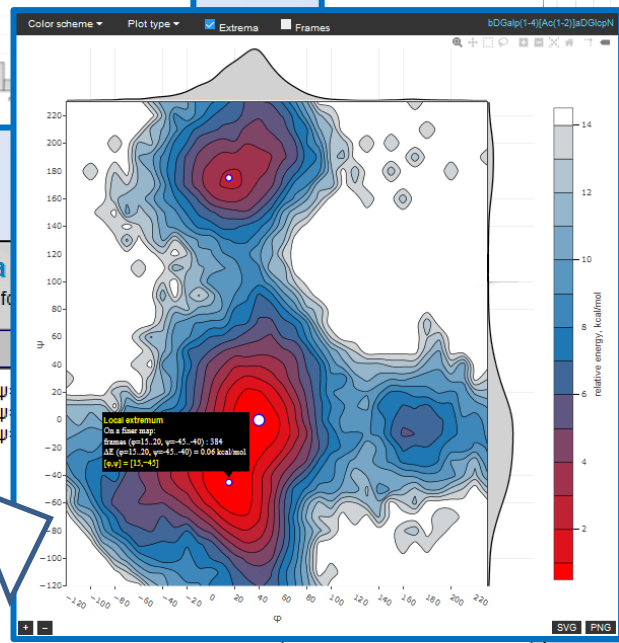
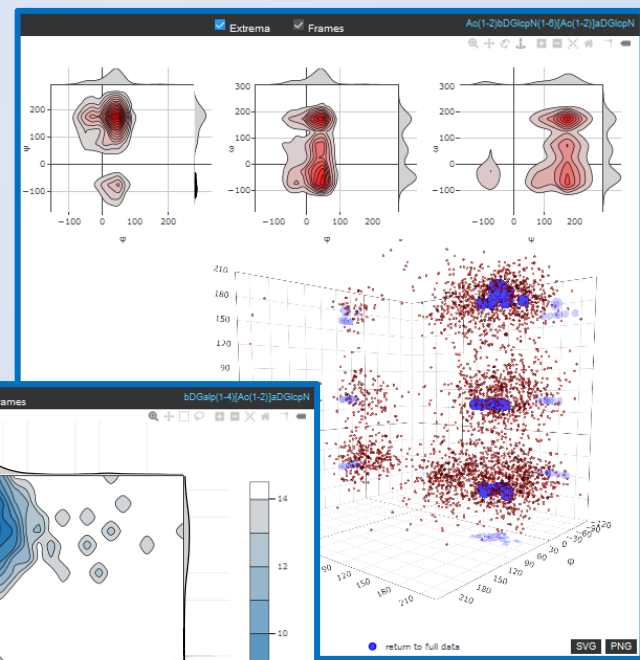
Temperature:

Solvent model:

[Home](#) [Help](#)

**CSDB conformation data**  
12 conformation maps have been found

Model structure	Conformation map	Energy	Force field
		$\varphi = -30, \psi = -10$ $\varphi = -35, \psi = -10$ $\varphi = -25, \psi = -10$	
		$\varphi = 25, \psi = 165, \omega = -75 \Delta E = 0.00 \text{ Kcal/mol}$ $\varphi = 45, \psi = 170, \omega = -60 \Delta E = 0.00 \text{ Kcal/mol}$ $\varphi = 45, \psi = 195, \omega = -75 \Delta E = 0.38 \text{ Kcal/mol}$ $\varphi = 50, \psi = 185, \omega = -60 \Delta E = 0.72 \text{ Kcal/mol}$ $\varphi = 25, \psi = 165, \omega = 180 \Delta E = 0.72 \text{ Kcal/mol}$ $\varphi = 30, \psi = 155, \omega = -60 \Delta E = 0.72 \text{ Kcal/mol}$ $\varphi = 35, \psi = 180, \omega = 180 \Delta E = 0.85 \text{ Kcal/mol}$ $\varphi = 45, \psi = 215, \omega = -60 \Delta E = 0.85 \text{ Kcal/mol}$ $\varphi = 40, \psi = 170, \omega = 165 \Delta E = 0.99 \text{ Kcal/mol}$ $\varphi = 20, \psi = 185, \omega = 165 \Delta E = 0.99 \text{ Kcal/mol}$ $\varphi = 55, \psi = 195, \omega = 165 \Delta E = 0.99 \text{ Kcal/mol}$ $\varphi = 55, \psi = 155, \omega = -60 \Delta E = 1.13 \text{ Kcal/mol}$ $\varphi = 45, \psi = 185, \omega = 165 \Delta E = 1.13 \text{ Kcal/mol}$ $\varphi = 30, \omega = 145, \omega = -60 \Delta E = 1.13 \text{ Kcal/mol}$	Force field: MM3-1996 Solvent model: None MD temperature: 1000 MD duration: 30 ns Frames: 30K MD summary file: <a href="#">download</a>



3D

2D

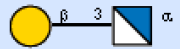
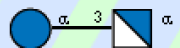
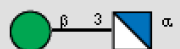
map live view

map live view

# Поиск гликозилтрансфераз

### CSDB glycosyltransferase search

42 glycosyltransferase activities have been identified in the CSDB database.  
Please note that GTR database covers only two species: *Escherichia coli* and *Streptococcus pneumoniae*.

Enzyme	Gene	Activity
Name: WbbD UniProt ID: <a href="#">Q03084*</a>	?	Synthesized dimer: bDGalp(1-3)aDGlcPn  Donor (ID 19342): <a href="#">DGalp(1-P-P-5)nucU</a> Acceptor (ID 19715): <a href="#">Ph(1-11)[Ac(1-2)aDGlcPn(1-P-1)]Subst // Subst = undecan-1,11-diol</a> Status: evidence <i>in vitro</i> <span>?</span> Confirmation methods: <i>in vitro</i> (crude extract) ID: 2053
Name: WbbG UniProt ID: <a href="#">Q0H8C8*</a>		Synthesized dimer: aDGlcP(1-3)aDGlcPn  Status: indirect evidence <i>in vivo</i> <span>?</span> Confirmation methods: mutation (knockout) Notes: Repeating unit of the O148 antigen. ID: 2151
Name: WbaD UniProt ID: <a href="#">Q1L815*</a>	Name: wbaD GenBank ID: <a href="#">7156002*</a>	Synthesized dimer: bDManp(1-3)aDGlcPn  Donor (ID 19855): <a href="#">DManp(1-P-P-5)nucG</a>


### CSDB glycosyltransferase search



Use the following conditions alone or in any combination to search for glycosyltransferases. Any field may be left blank for no restrictions.

**GT names and IDs:** Type enzyme name, e.g. "Orf10". Wildcards (\* and ?) are supported.  
 Enzyme name:

**Organism:** Select species  Type strain/serogroup

**Molecule role:** Filter by target structure

**Synthesized bond:** Type dimeric fragment in CSDB encoding or use tools  
 [Use Wizard](#) 

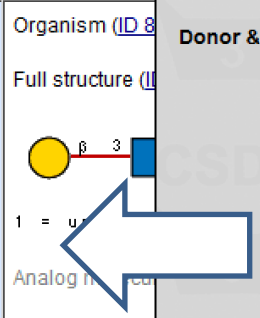
**Donor & acceptor:** Type donor CSDB encoding or use tools  
 [Use Wizard](#)   
Type acceptor CSDB encoding or use tools  
 [Use Wizard](#) 

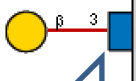
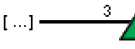
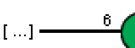
Treat donor/acceptor as fragments

**Confirmation status:** Filter results to those

[Search!](#)

[Home](#)   [Help](#)   [HELP !!!](#) ?



 Full structure (ID 8)	Organism (ID 8)
 Full structure (ID 8)	Organism (ID 8)
CSDB ID(s): <a href="#">11572</a> , <a href="#">21578</a> , <a href="#">23062</a> , <a href="#">26257</a> , <a href="#">27289</a>	Molecule role: O-antigen
Organism (ID 1863): <i>Escherichia coli</i> O77	Zhou et al. 2016 DOI: <a href="#">10.1016/j.carres.2016.02.007</a>
Full structure (ID 4600): 	Wang et al. 2007 DOI: <a href="#">10.1099/mic.0.2007</a>

# Кластеризация таксонов

**Scope settings**

Limit taxonomical scope to: **phylum**

Display groups:  bacteria  archaea  protista  algae  fungi  plants  animals

**Phylum:** (select multiple with CTRL key)

- (unspecified bacteria)
- (unspecified protista)
- Actinobacteria
- Bacteroidetes
- Chlamydiae
- Chloroflexi**
- Crenarchaeota
- Cyanobacteria

**General settings**

species Rank of taxons to compare (should be lower than selected scope). [Specify exact species \(all\)](#)

50 **Taxon population threshold.** Minimal number of structures\* assigned to a taxon or its subtaxons, to include this taxon in calculation (affects selection of taxons). Check to use this filter.

15 % **Normalized taxon population threshold.** Minimal part of structures\* assigned to a taxon or its subtaxons, to include this taxon in calculation (affects selection of taxons). Normalized by the total number of structures\* in the database. Check to use this filter.

50 **Structure abundance threshold.** Minimal number of structures\* in which a fragment should be contained to be qualified as 'present in biota' (affects selection of fragments)

60 **Fragment abundance threshold.** Minimal number of instances\* in which a fragment should be present to be qualified as 'present in biota' (affects selection of fragments)

2 **Fragment presence threshold.** Minimal number of instances\* in which a fragment should be present in organisms of a taxon to be qualified as present in this taxon (affects occurrence codes and thus, taxon dissimilarity)

**two residues** Type of fragments to analyze (dimeric or monomeric)

**only polymers** Type of structures to analyze. Only structures of this type are considered in fragment analysis and where marked (\*). 'Optimized' = only polymers from bacteria, archaea and fungi, and only mono/oligomers from plants.

**R-project** Format of the dissimilarity matrix

**Fragment pool generation settings**

- Combine anomeric forms.** All sugar residues will be treated as 'any anomer'
- Exclude underdetermined residues.** Residues with unknown anomeric, absolute or ringsize configuration will be omitted from analysis.
- Exclude monovalent residues.** Residues like Me, Ac, etc. will be omitted from analysis. Please note, that Ac in N-acetylated amnosugars is a separate residue.
- Exclude superclasses.** Fragments with residues represented by aliases and superclasses will be omitted from analysis.
- Differentiate aliases.** Residue aliases (used for atypical residues) will be differentiated by actual residue names, otherwise they are combined under an alias name.
- Sugars only.** Fragments with non-sugar residues (including monovalent residues, like N-acetyls) will be omitted from analysis.
- Exclude aglycons.** Fragments with atypical residues at non-reducing ends will be omitted from analysis.
- Differentiate location.** The same fragments at different locations (inline, terminal, reducing) will be treated as different.
- Strict comparizon** of fragments. Unknown configurations and ringsizes are always unequal to those known (otherwise a fuzzy comparizon is performed).

## Distance matrix based on fragment presence

The analysis was performed over all cellular organisms

Prepared 20 monomers  
Prepared 32 genera  
Generated occurrence bit-codes. [Show](#)  
Generated dissimilarity matrix. [Show](#)

**Calculation parameters:**

Hamming mode: YES  
Fragment size: monomer  
Fragment abundance filter: instance threshold: 550  
Fragment abundance filter: structures threshold: 500  
Fragment presence threshold: 2  
Differentiate structures of this type: any  
Filter: differentiate monomer positions (inline/terminal/reducing) in structures: NO  
Matrix data format: R

**Coverage data on used taxons:**

(taxons, number of organisms in a taxon, number of structures assigned to these organisms)

Acinetobacter (BA)	68	140
Aeromonas (BA)	64	122
Bacillus (BA)	95	234
Burkholderia (BA)	36	219
Solanum (PF)	46	127

**Matrix-based dendrogram:**

Your job name is `dsmatrix_2014Nov09_21-34-09`

Use these persistent links to download [all job data](#) or [the distance matrix alone in R format](#)

Build a new **unrooted tree** and colorize **12** cluster(s): **Rebuild dendrogram** and **Export Newick tree**



# Предсказание структуры

### Structure generation constraints:

The structure contains 6 residue(s): [Add residue](#)

$\alpha/\beta$	D/L	Residue	Ring form
1. ?	D	galact-2N-uronic acid	pyranose
2.		acetic acid	
3.	D	show all residues	
4.		phosphoric acid	
5. $\alpha$	?	any octose	pyranose
6.	L	alanine	

Allowed linkages:

	C1	C2	C3	C4	C5	C6	C7+
D-GalpNA	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	None
Ac	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
D-Rib-ol	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	None
P	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
$\alpha$ -Octp	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	None
L-Ala	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	None

Advanced options: [Hide](#)

Min in	Max in	Location	Ac at N	Acceptors	Remove
1	2	any	demanded	any	<input checked="" type="checkbox"/>
?	?	any		any	<input checked="" type="checkbox"/>
		terminal		3,5	<input checked="" type="checkbox"/>
?	?	reducing			<input checked="" type="checkbox"/>
?	?	any	forbidden	1	<input checked="" type="checkbox"/>

Search depth: Widespread structures only

Scope:  oligomers  polymers

Advanced scope:  $\beta$ -anomers: = 1 CH<sub>2</sub> carbons: ? no furanoses

### Top 15 matches:

#Rank	Structure	Experimental spectrum	Simulated spectrum	Comments
#1. $\Delta \sim 0.94$ ppm Corr = 1.000 RMS dev = 1.46 ppm Trust = 46%				1 = b-D-GalpNA 2 = L-Ala
#2. $\Delta \sim 0.95$ ppm Corr = 1.000 RMS dev = 1.46 ppm Trust = 46%				
#15. $\Delta \sim 1.42$ ppm Corr = 0.999 RMS dev = 1.99 ppm Trust = 49%				1 = b-D-GalpNA 2 = L-Ala

### Find best matching structures:

Experimental <sup>13</sup>C NMR spectrum in water (24 signals of 24 expected):

17.4 22.9 34.7 50.5 52.4 63.9 64.9 66.2 68.3 70.6 72.3 72.4 72.7 73.3 73.6 76.5  
78.6 78.8 99.2 102.6 171.2 175.2 176.0 176.5

$\pm$  2 signals

Find 15 best-fitting structures

Save generated structures

[Go!](#)

E-mail for results: [why?](#)  
user@gmail.com

