Институт органической химии им. Н.Д. Зелинского РАН, Москва

90 мин

Информатика в науке об углеводах

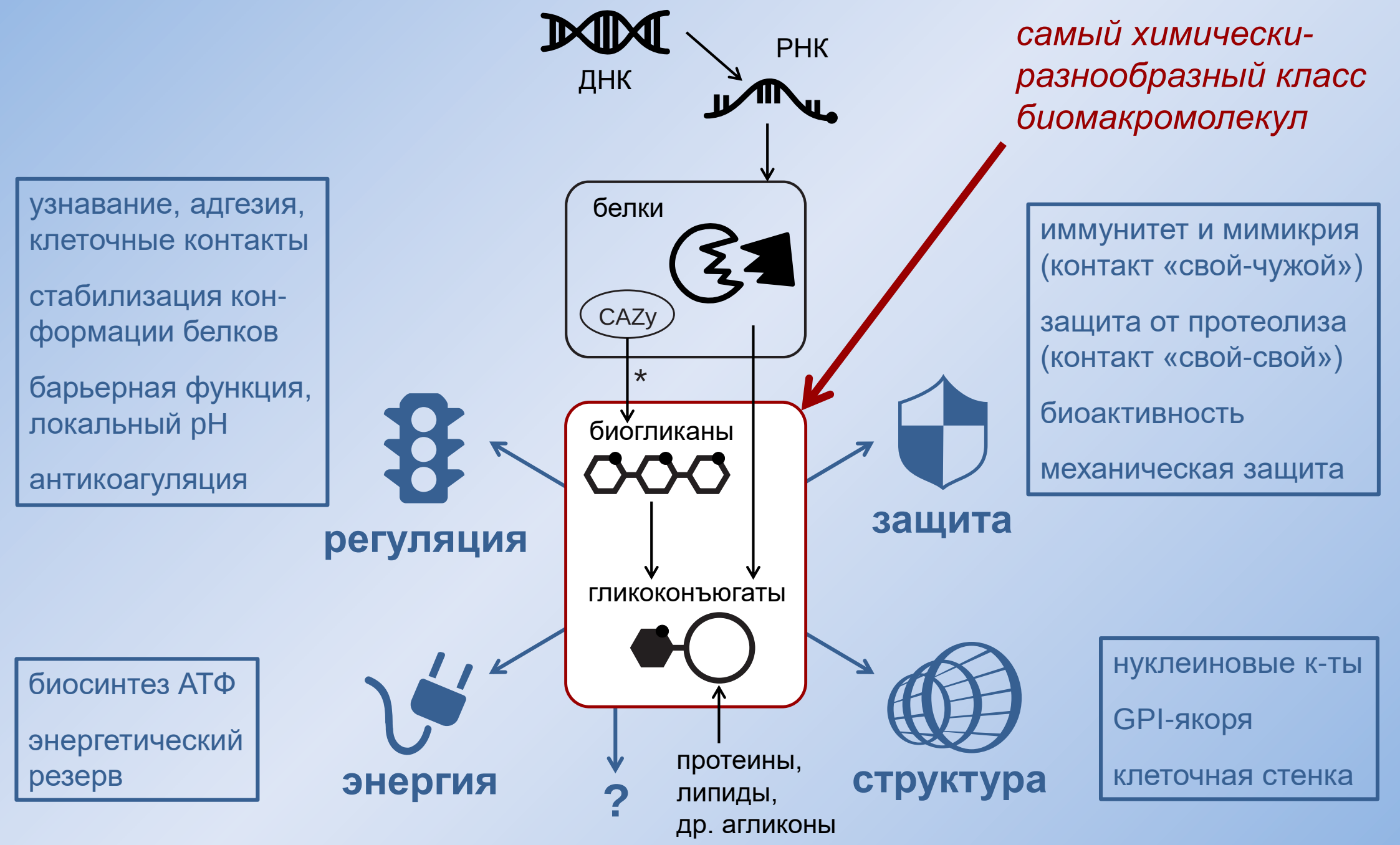
ver. 2024

Филипп Тоукач

вед.н.с. ИОХ РАН, профессор НИУ ВШЭ,
руководитель группы гликоинформатики лаб. 21 ИОХ

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

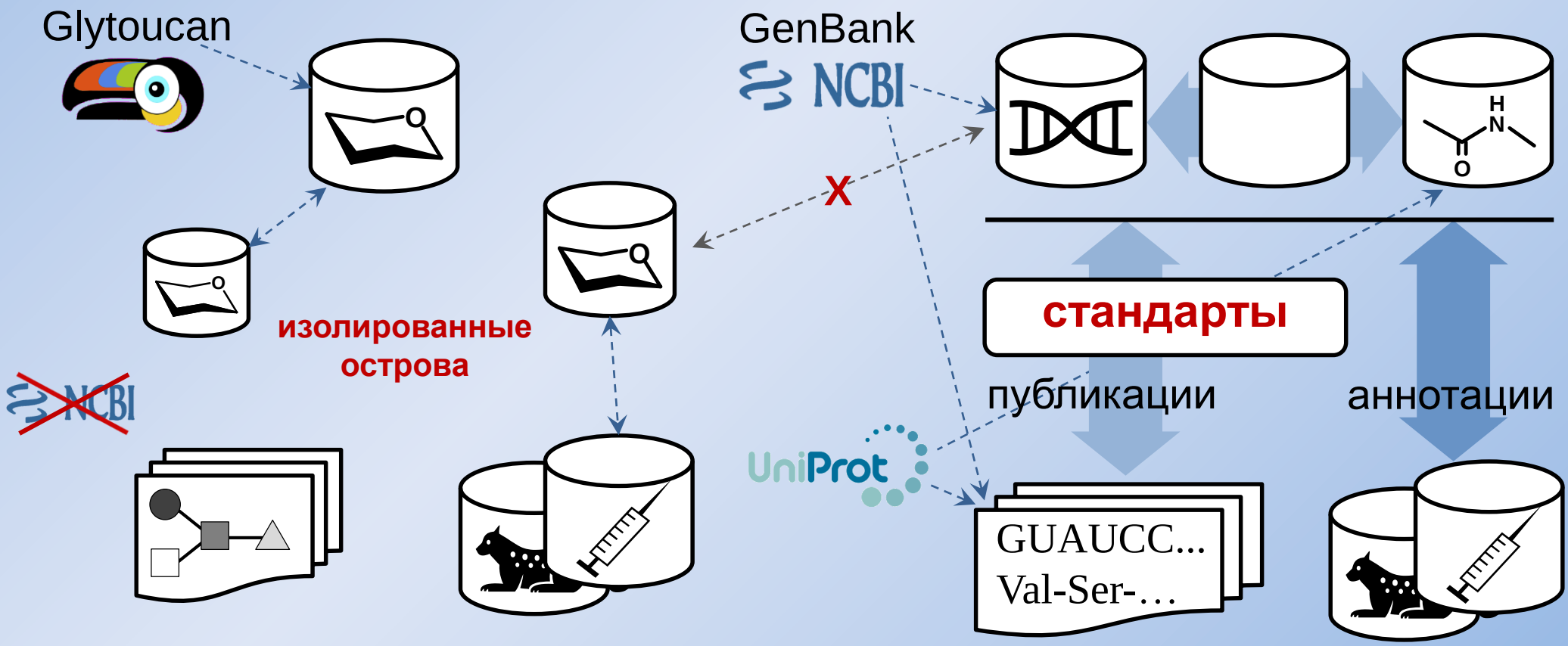
Углеводы в клетках



Гликомика vs. геномика, протеомика

по сравнению с другими –omics:

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



Гликоинформатика

цель - обеспечить исследователей углеводов всей мощью информационных технологий (от концепции до веб-сервиса).

- **Легкий доступ к знаниям и автоматизация исследований**
  Какие природные структуры похожи на заданные? Какие их фрагменты специфичны для заданных биологических видов? Где они опубликованы, в привязке к каким таксонам, болезням, и т.д.? Какие ферменты их синтезируют и с какой достоверностью это показано? На какие гликоэпитопы реагируют антитела?
- **Моделирование свойств молекул**
  Молекулярная геометрия, спектры, биоактивность, …
- **Предсказание структуры по наблюдаемым свойствам**
- **Предсказание свойств таксонов**
  Кластеризация на основании гликомов, поиск схожести и различий таксонов, хемотаксономическая классификация
- **Идентификация, кодирование и визуализация молекул**
- **Методология обработки накопленных знаний**
  Аннотирование публикаций, интеграция проектов, стандартизация знаний

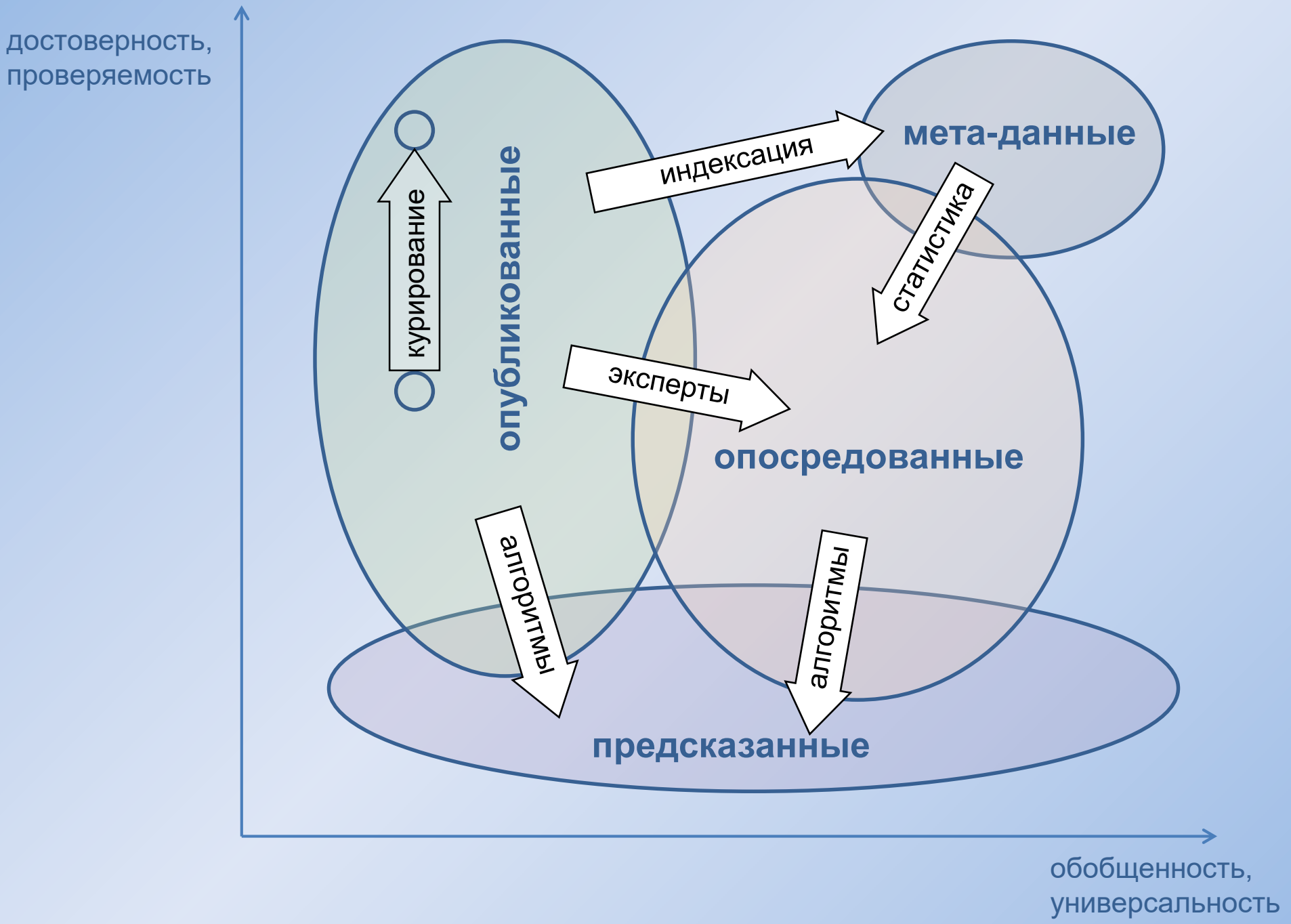
Проблемы в гликоинформатике

- 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

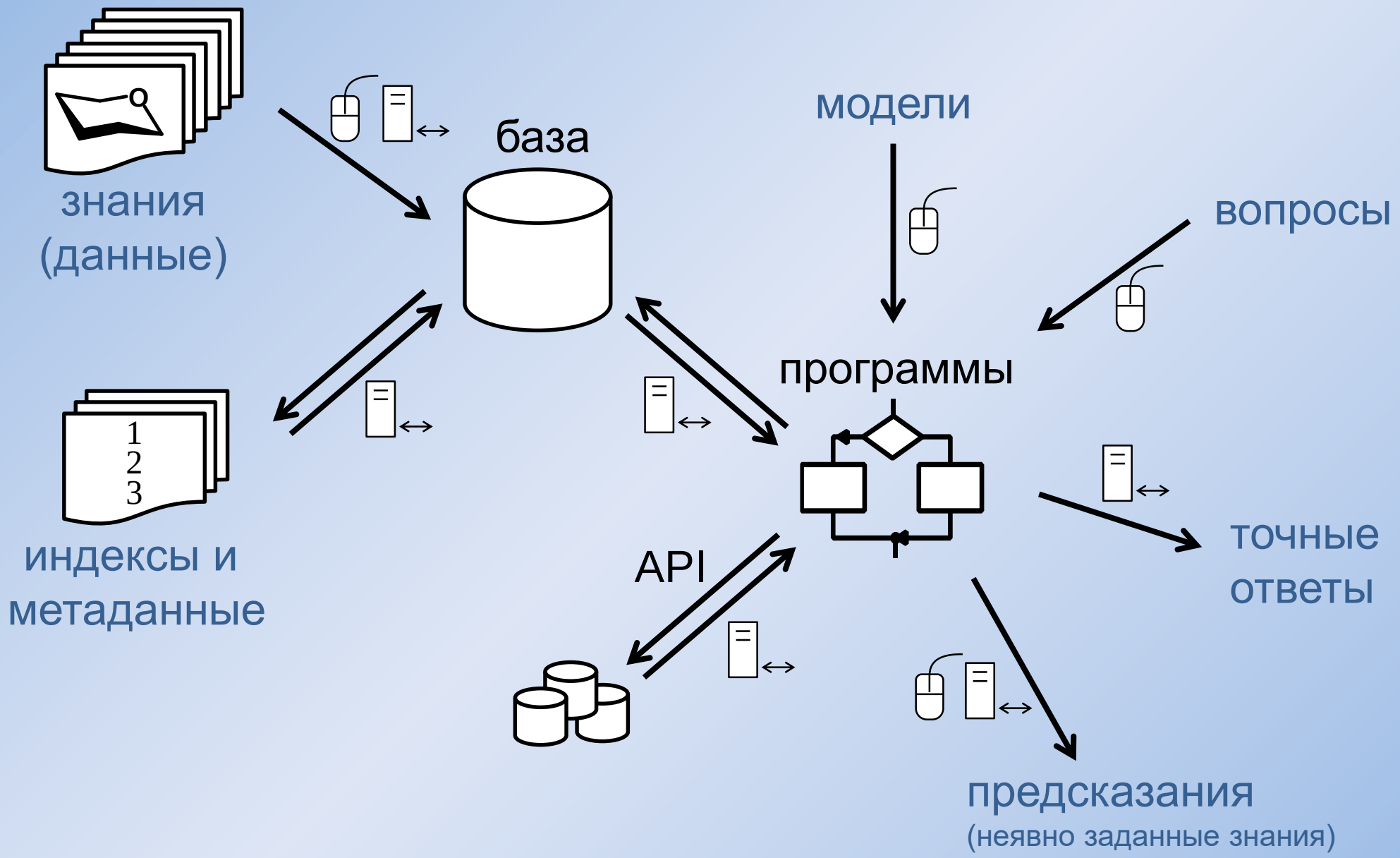
  (нет общепризнанных сервисов, инициативы несовместимы друг с другом)

Aoki-Kinoshita K, *et al.* **Introducing glycomics data into Semantic Web** *J Biomed Semantics 2013, 4: id 39*
Egorova K, Toukach Ph **Bridging isolated islands in the sea of data** *Angewandte Chemie Intl Ed 2018, 57:14986-14990*

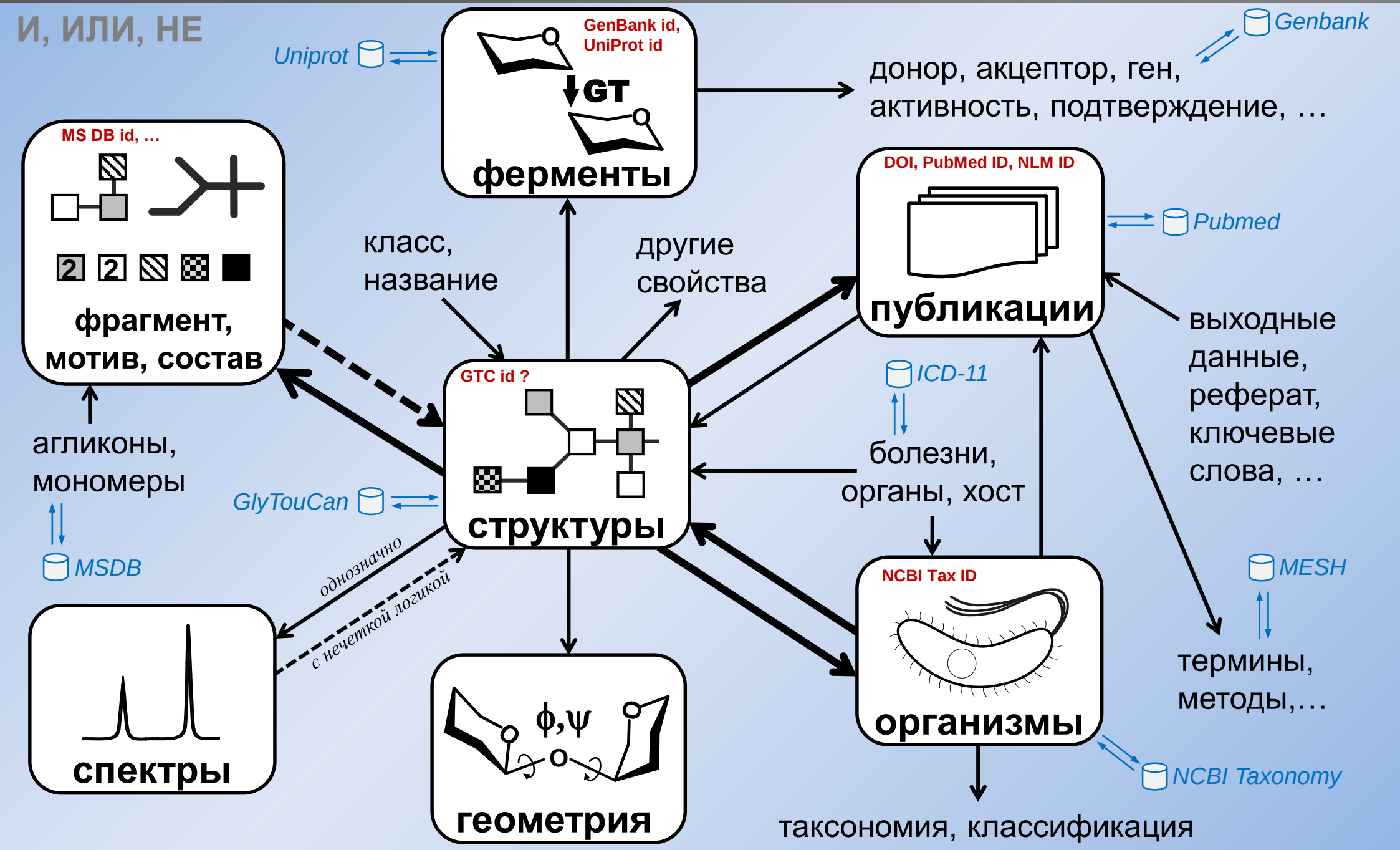
Уровни данных



Базы данных


знания
(данные)
база
модели
вопросы
программы
1
2
3
индексы и
метаданные
API
точные
ответы
предсказания
(неявно заданные знания)
с участием человека
автоматически

Типичные запросы

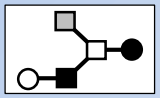


Carbohydrate Structure Database

CSDB

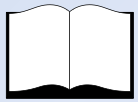
База данных природных углеводов

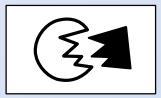
Платформа для сервисов гликоинформатики

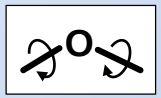

33K
первичные структуры

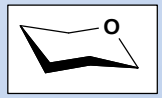

16K
таксономия


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

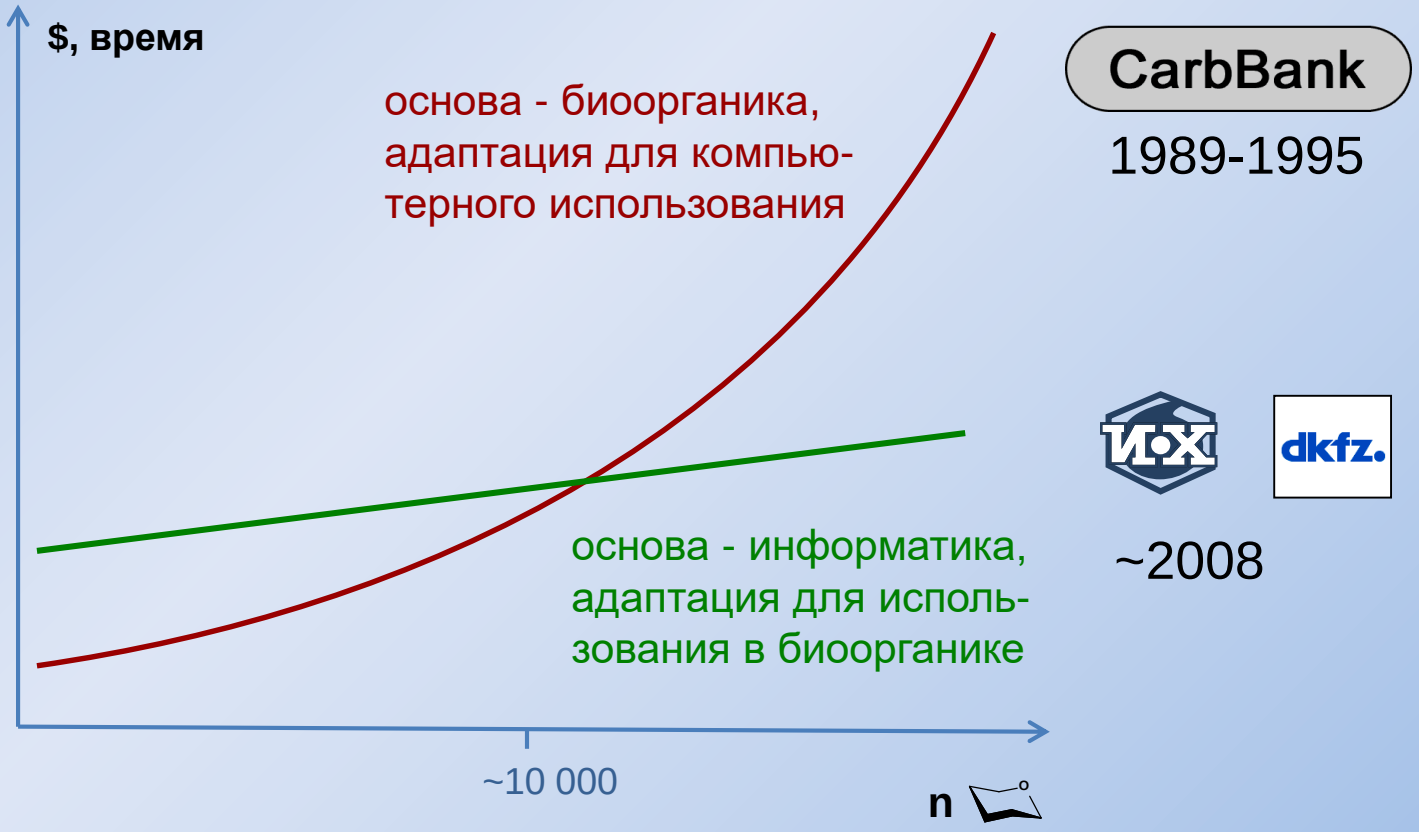
- Конформационные расчеты сахаридов
- Углеводные деревья жизни
- Углеводная нотация CSDB Linear
- Symbol Nomenclature for Glycans
- Онтология GlycoRDF и другие стандарты
- Редакторы, конвертеры, стат. анализаторы, …

http://csdb.glycoscience.ru

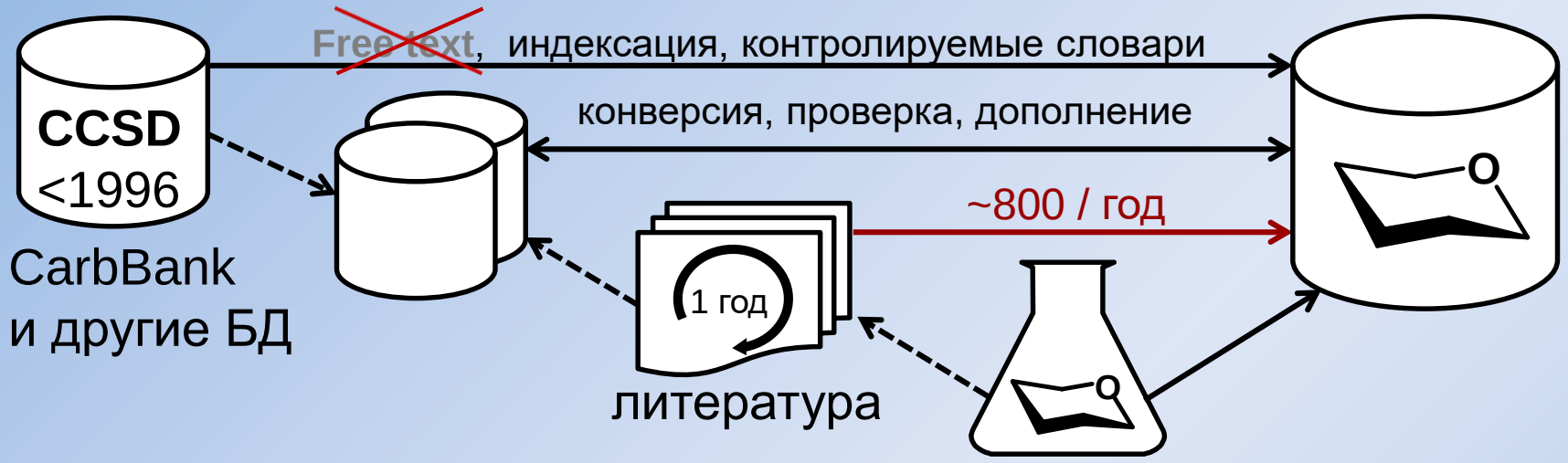
Egorova K, Toukach Ph **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017, 27:285-290*
Toukach Ph, Egorova K **Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts** *Nucl Acid Res 2016, 44:D1229-1236*
Egorova K, Toukach Ph **Expansion of coverage of Carbohydrate Structure Database (CSDB)** *Carbohydr Res 2014, 389:112-114*

Организация разработки

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



Источники данных



CSDB 2023:
32 400 *структур*,
12 600 *публикаций*,
16 200 *организмов*,
19 400 *спектров*,
2 900 *конф. карт*,
2 500 *активностей ГТ*
+взаимосвязи

- выявление публикаций
- ретроспективный анализ
- аннотирование и проверка

полное покрытие по микроорганизмам и грибам:
отрицательный результат поиска = значимая научная информация

первичная структура	род, вид, штамм
тривиальные названия, класс	болезни, органы,
спектры ЯМР, условия съемки	хост, стадия
конформация	библиография
гликопротеины,	*ключевые слова*
биоактивность	*рефераты*
гены, ферменты биосинтеза	*аффилиации*
ссылки на другие БД	методы

10 — реляционная БД, таблица связности для структур, стандартные индексы, человекочитаемый дамп, …

Качество данных

операторские — >2500 в базах — найдено & исправлено — ~900 в статьях — A → B в программах

ошибки, противоречия

исправляемые

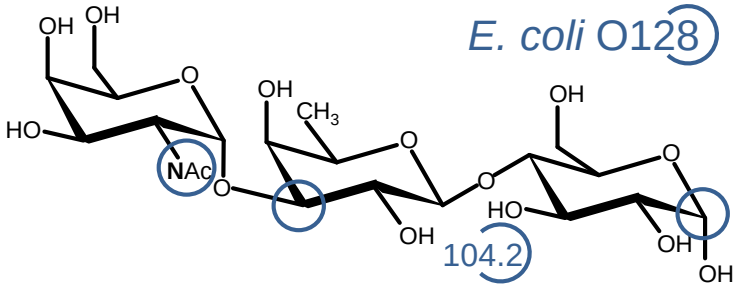
2dGlc → araHex,
α-Rib-ol → Rib-ol,
D-Kdo → Kdo,
1-methyl → 1-Me,
n.m.r. → NMR,
taxid 583 → Proteus,
…

выявляемые

Glc(1-2)GlcN,
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-Gcl, …

невыявляемые

E. coli O127:
aDGal*p*N(1-4)bDFuc*p*(1-4)**b**DGlc*p*
Glc C1 103.2 ppm

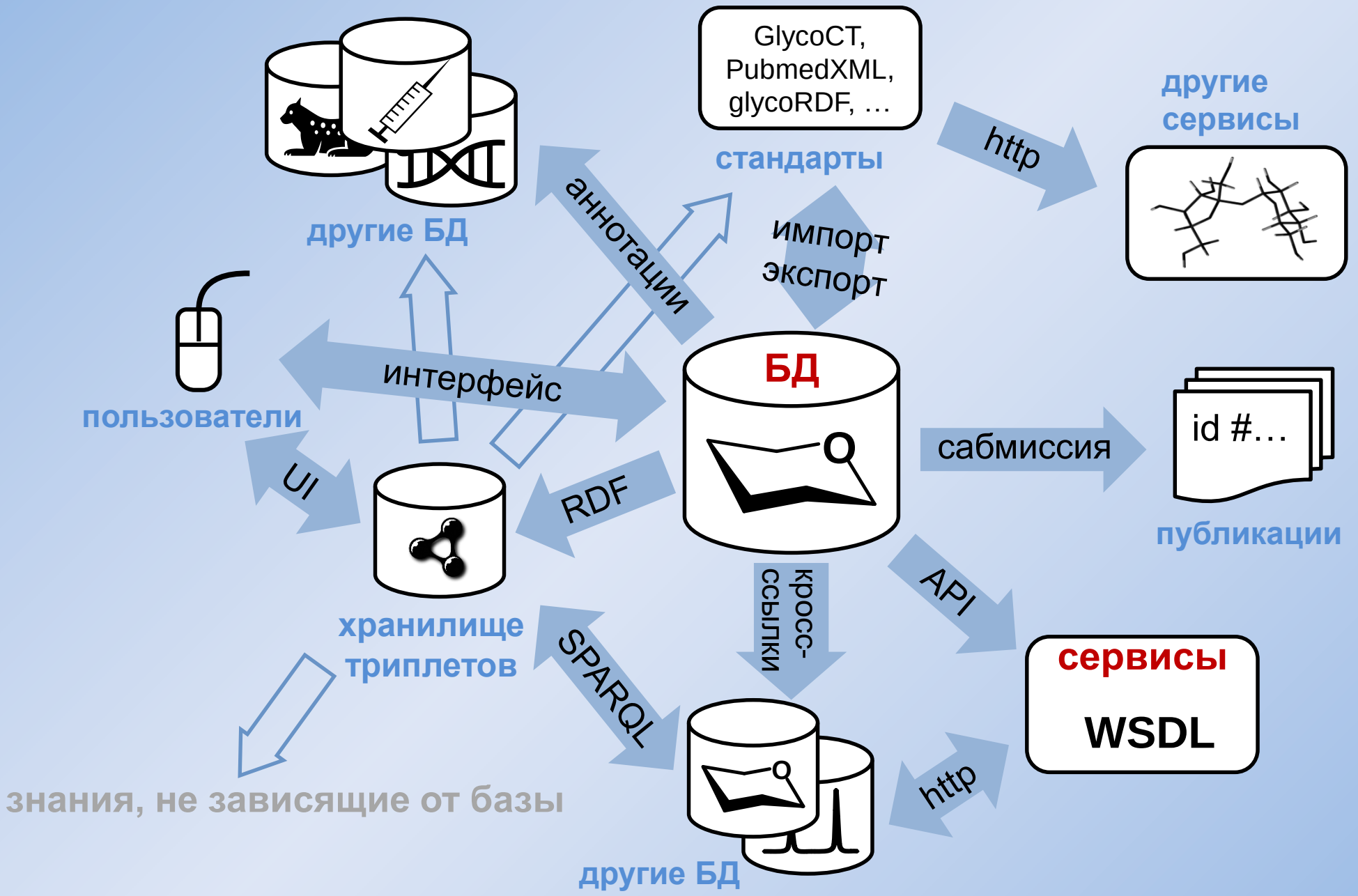


CarbBank ☹ **>50%** (неправильные, отсутствующие, ложно присутствующие структуры, штаммы, аннотации)

→ другие БД

☺ **<10%**

Идеальная интеграция



Aoki-Kinoshita K, *et al.* **Towards an international glycan structure repository** *Glycobiology 2013, 23:1422-1424*
Toukach Ph, Joshi H, Ranzinger R, Knirel Y, von der Lieth C-W **Sharing of worldwide distributed carbohydrate resources** *Nucl Acid Res 2007, 35:D280*

Resource Description Framework

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

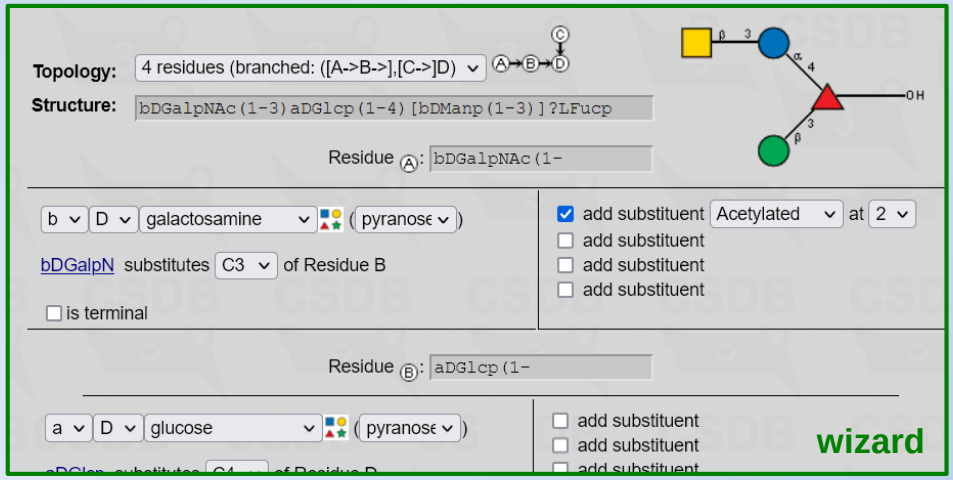
GlycoRDF – первая формальная углеводная онтология (OWL)

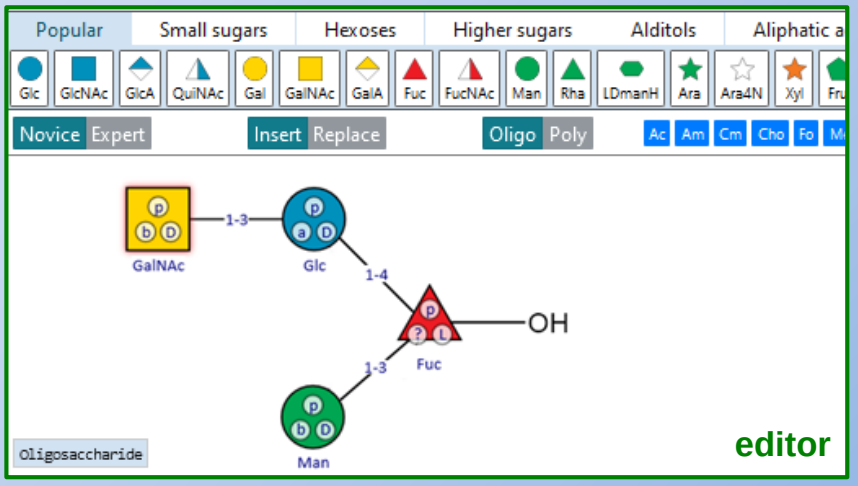
Ranzinger R, *et al.* **GlycoRDF: an ontology to standardize glycomics data in RDF** *Bioinformatics 2015, 31:919-925*
Aoki-Kinoshita K, *et al.* **Introducing glycomics data into the Semantic Web** *J Biomed Semantics 2013, 4:id39*

Интерфейс

- Конверсия данных ↔ другие форматы
  - SNFG, WURCS, GlycoCT, SMILES, MOL, PDB, Glyde II, LinUCS, Sweet-DB GLYCAM, GlycoRDF, DCI XML, …
- Автоматические web-сервисы (WSDL)
- Импорт, экспорт
- Документация, HELP
- Ссылки на записи в других проектах
  - DOI, NCBI Pubmed, NCBI Taxonomy, Uniprot, NCBI Genbank, PubChem, ImmuneEpitopeDB, Glytoucan, ICD-11
- Ввод и вывод структур

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

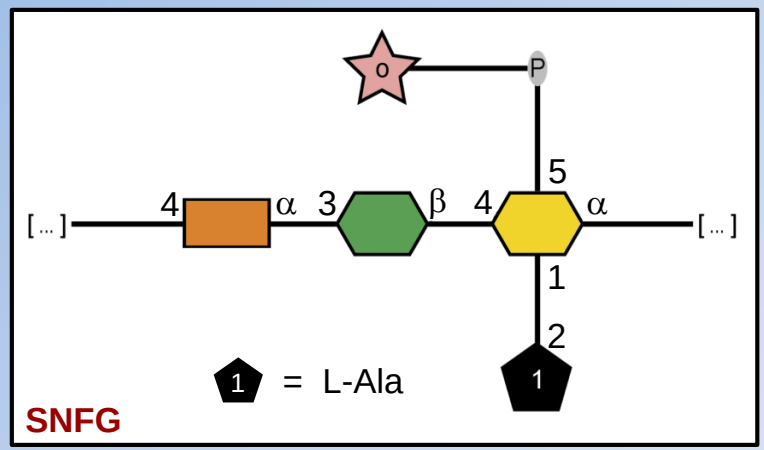




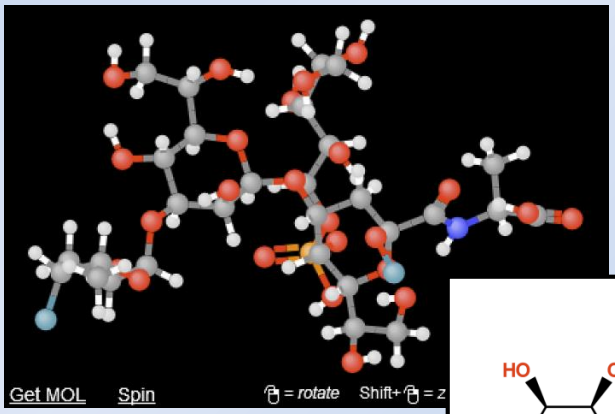
`bDManp(1-3)[Ac(1-2)bDGalpN(1-?)aDGlcp(1-4)]?LFucp` нотация CSDB Linear

Вывод структур

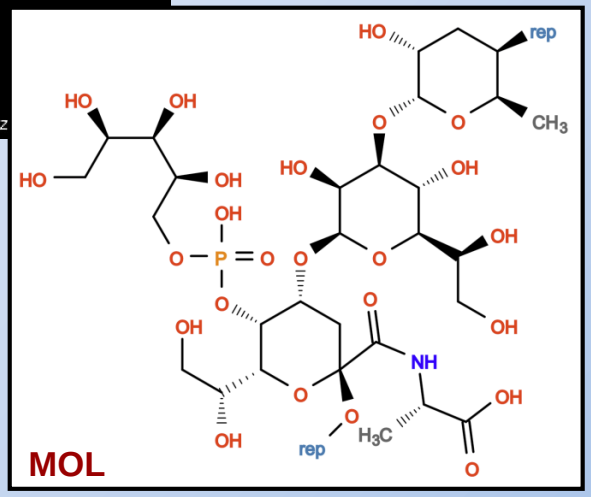
Визуализация в человекочитаемых форматах:



SNFG



MOL (3D)



MOL

```
                                D-Rib-ol-(1-P-5)┐
                                                │
-4)-α-Abep-(1-3)-β-L-gro-D-manHepp-(1-4)-α-Kdop-(2-
                                                │
                                L-Ala-(2-1)┘
```

Sweet-DB

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

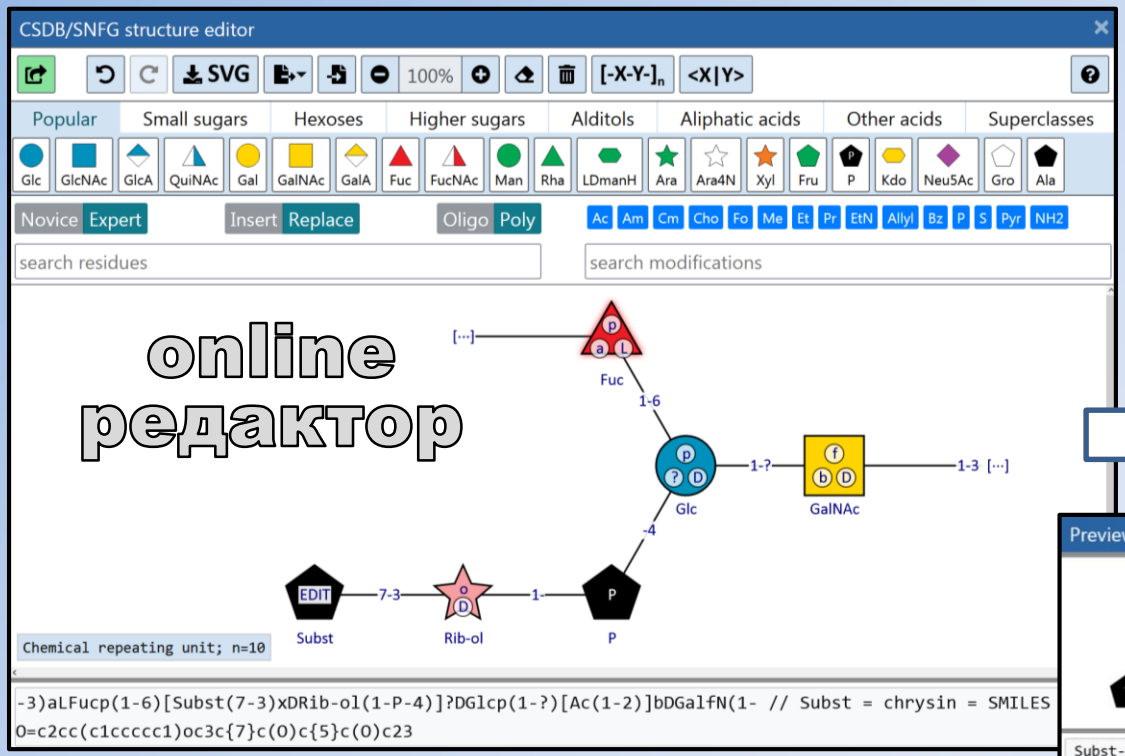
```
-4)aXAbep(1-3)bXLDmanHepp(1-4)[xDRib-ol(1-P-5),xLAla?(2-1)]aXKdop(2-
```

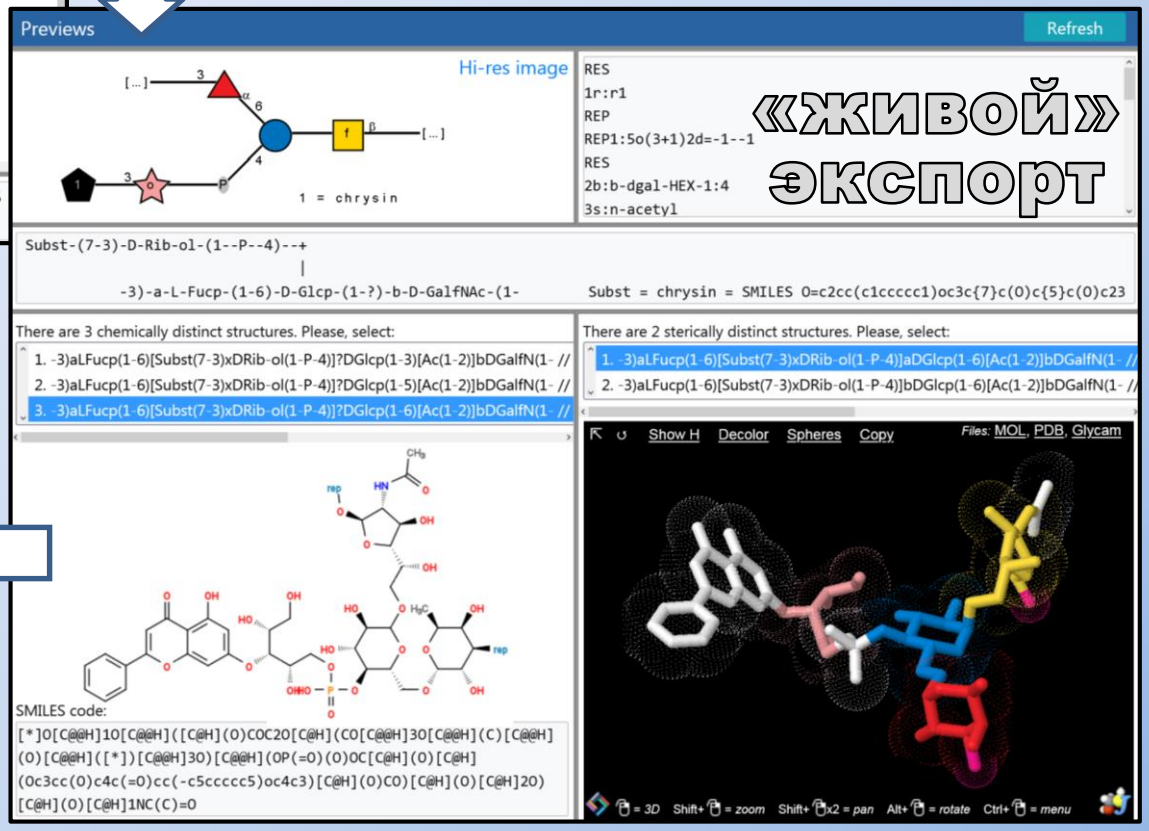
CSDB Linear

Chernyshov I, Toukach Ph **Automated translation of glycans from residue-based notation to SMILES and coordinates** *Bioinformatics 2017, 34:2679*
Varki A, *et al.* **Symbol nomenclature for graphical representations of glycans** *Glycobiology 2015, 25:1323-1324*

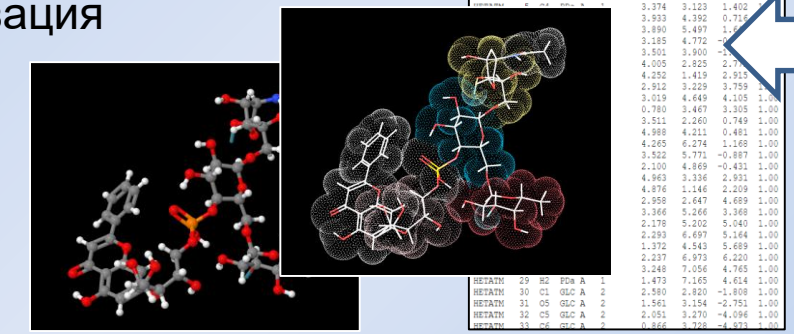
Редактор структур

- 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

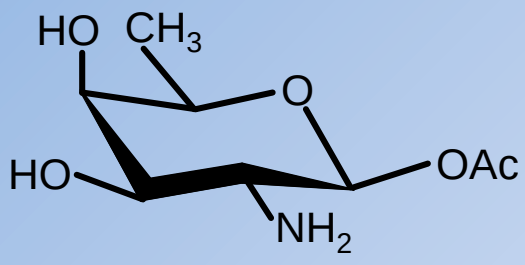




- работа с геометрией в браузере
- экспорт на углеводных и химических языках
- экспорт атомных координат
- визуализация
- цвета SNFG



Трудности перевода



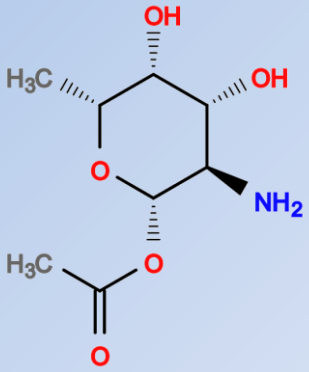
bDFucpN(1-1)Ac (CSDB)

однозначно соотнесено со структурой, но при этом понятно людям

D-Fuc*p*N-β1OAc

beta-fucosamine acetate

1-acetoxy-beta-D-fucopyranosamine



2-deoxy-2-amino-β-D-fucopyranosyl acetate (IUPAC)

β-D-fucosamine acetic ester

β-6-deoxy-D-galactosamine acetate

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

1Ac (CFG)

β-D-фукозамин-1-O-ацетат (на других 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)

WURCS=2.0/1,1,1/[a2112m-1b_1-5_2*N]/1/a1*OCC/3=O (WURCS)

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 α-D-Gal*f*-(1-3)-β-D-Glc*p)*
- Сложный перевод семантического описания в поатомное
  => трудоемкость аннотирования статей
- Нельзя описать структуры с неопределенностями
- Данные визуально не сопоставлены со знаниями

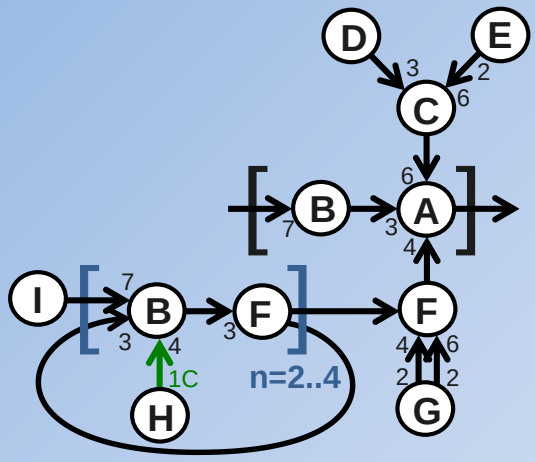
⇩

- 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)

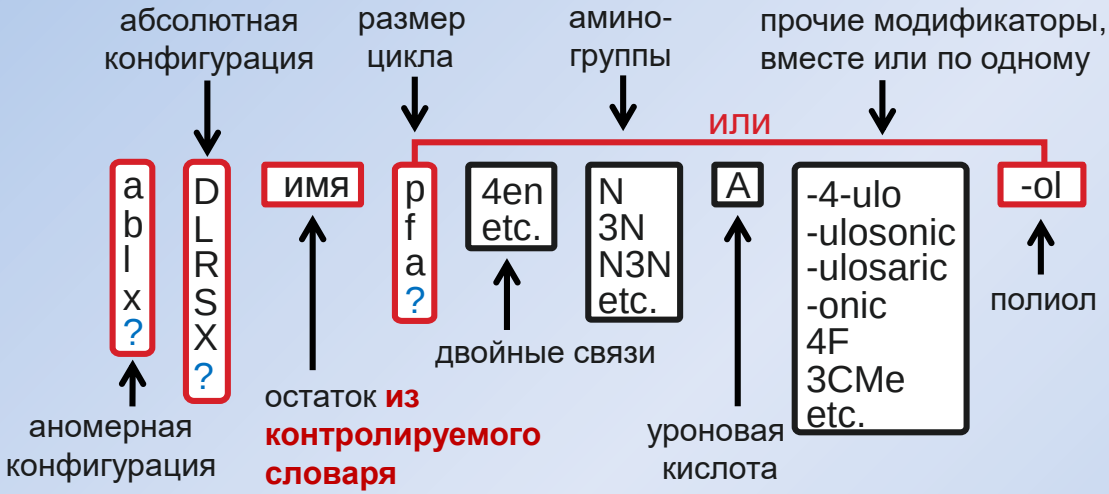
```
52.0606    6.3910   -0.1606 C   0  0  0
51.8591    8.6986   -0.1875 N   0  0  0
52.9844    9.0384    0.7259 C   0  0  1
53.8550    8.9929    0.0662 H   0  0  0
52.9684   10.5530    1.3121 C   0  0  2
52.2705   11.0903    0.6993 H   0  0  0
1117    1  0  0  0  0
1118    1  0  0  0  0
1119    1  0  0  0  0
```

атомы, координаты, связность

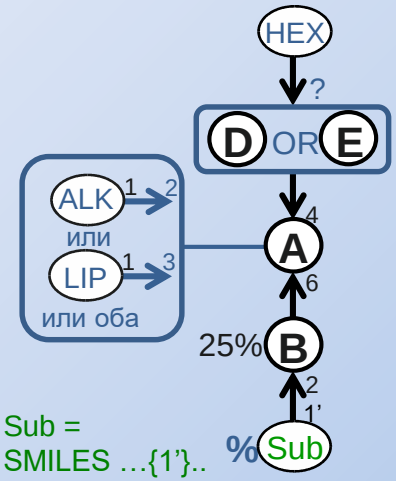
Нотация CSDB Linear



топология

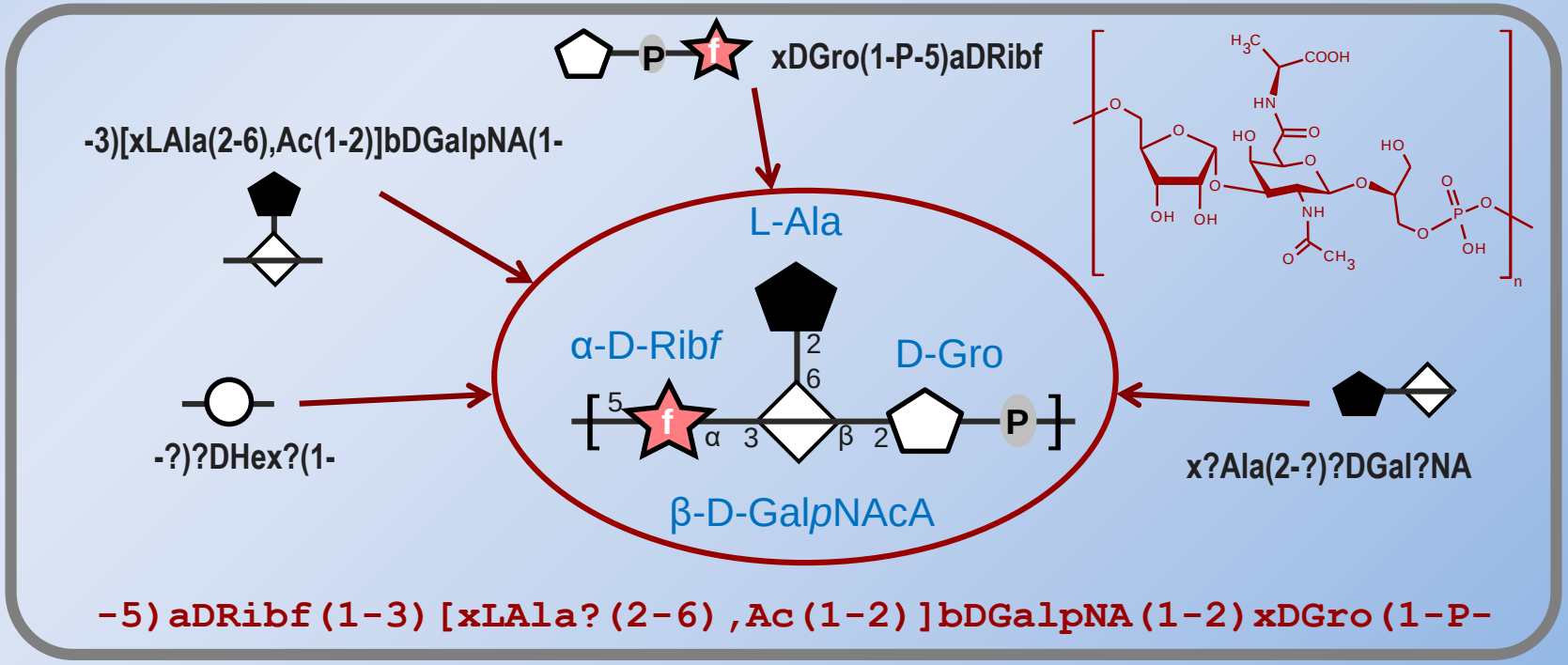


мономеры

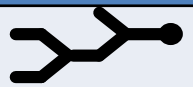
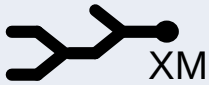
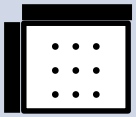
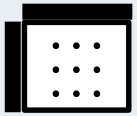
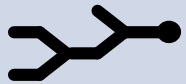
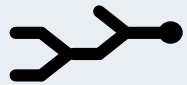
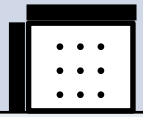
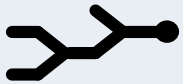


неточности

- полнота
- однозначность
- человекочитаемость
- машиночитаемость
- расширяемость
- атомарный слой
- неопределенности
- конверсия в/из других нотаций



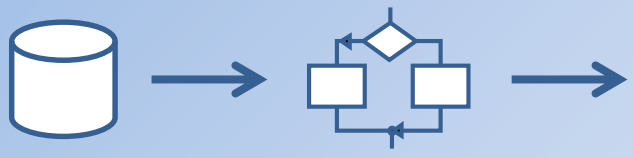
Сравнение языков

	подход	полнота	однознач-ность	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



выводы на основании обработки ВСЕХ данных, в том числе скрининг и статистика

- Анализ путей биосинтеза (база гликозилтрансфераз)
- Конформационные карты олигосахаридов
- Предсказание и отнесение спектров ЯМР ^{13}C, ^{1}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 и агликонов

Гликозилтрансферазы

фермент

WdaF
CAZy GT99
HQ388393
Genbank #

ген, кластер, семейство

WdaF
Uniprot
E7DWQ4
CSDB #2918

идентификаторы и названия

полная по эксп. подтвержденным активностям *A. thaliana, E. coli, S. cerevisiae, A. baumannii*

DOI, PMID

публикации

E. coli O11

вид, штамм, орган

ID 11762

полная структура, роль в клетке

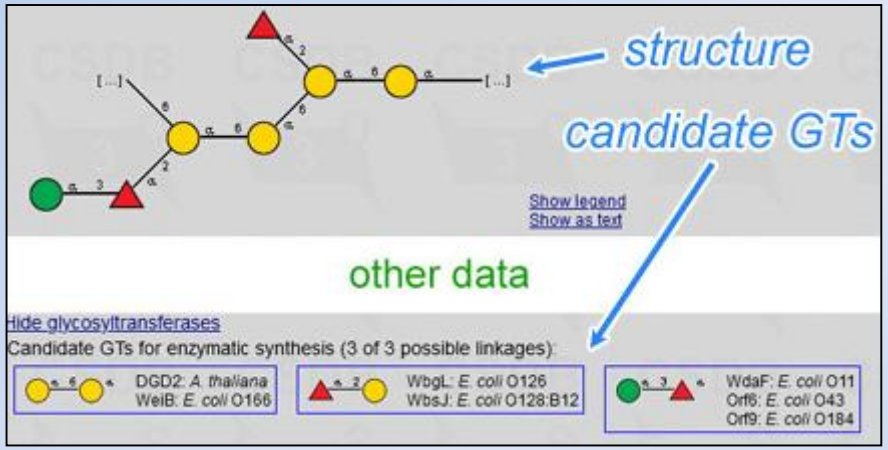
структура или подструктура

Man—UDP

донор и акцептор

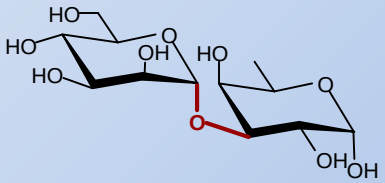
CSDB GT

Glycosyltransferase GT Database



1α 3

синтезируемый фрагмент



- *in vivo*
- *semidirect*
- *indirect*
- *in vitro*
- *suggested*
- *in silico*

уровень достоверности

активность — объект

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*

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

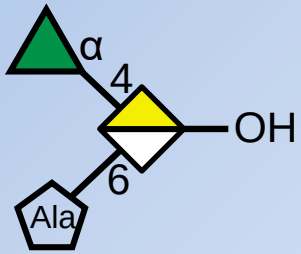
- Нам нужен инструмент «из коробки», для не-информатиков

- → стандартизованные модели структур в растворе
- → автоматическая генерация, обширные пред-расчеты
- → экспорт в атомарные форматы

Анализ конформаций

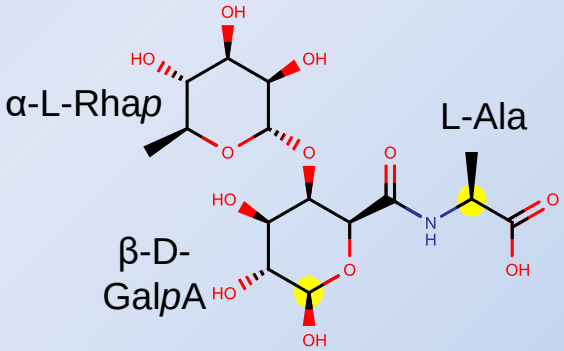
aLRhap(1-4)[x?Ala?(2-6)]?DGalpA

структуры,
в т.ч. неполные

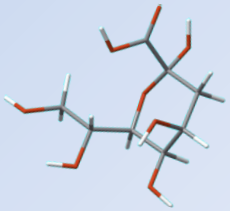


другие варианты
(α-GalA, D-Ala, и т. д.)

SMILES

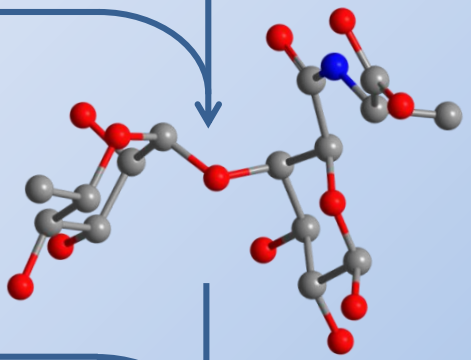


выгодные
конформеры
~1000 остатков

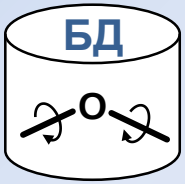


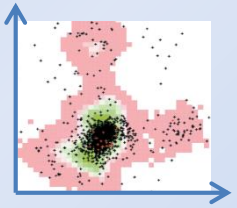
$^4C_1$,$^1C_4$, …

«креслификация»



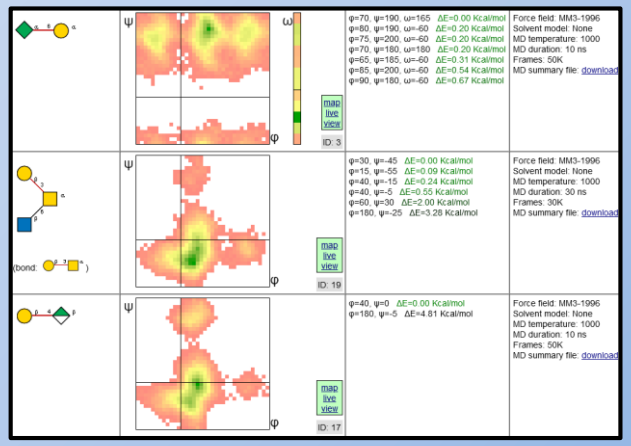
заселенные
состояния
мостиков





φ,ψ,ω,θ

выбор минимумов
ММ-релаксация



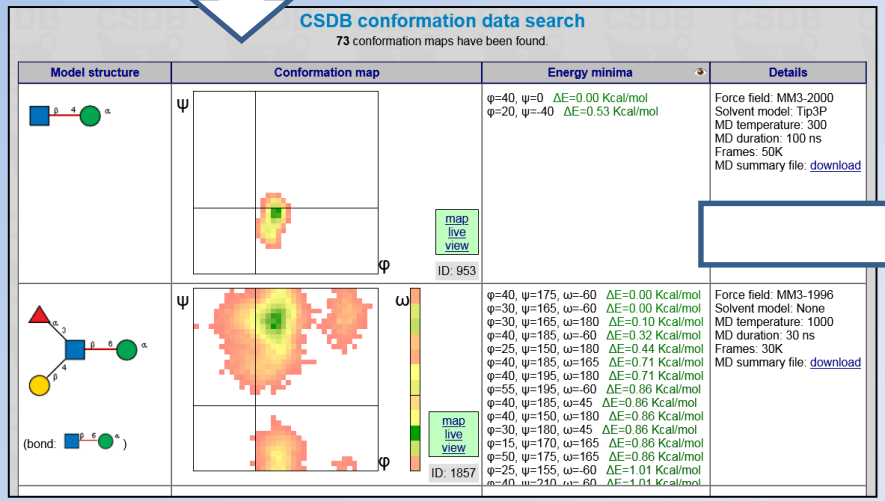
конформеры
+ их энергии

мол. динамика
300K, 100нс, H_2O

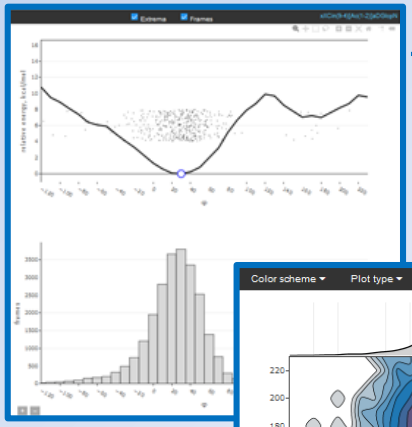
Конформационный модуль CSDB

~2700 ди- и трисахаридов:
100-нс MD в явной H_2O, 300K

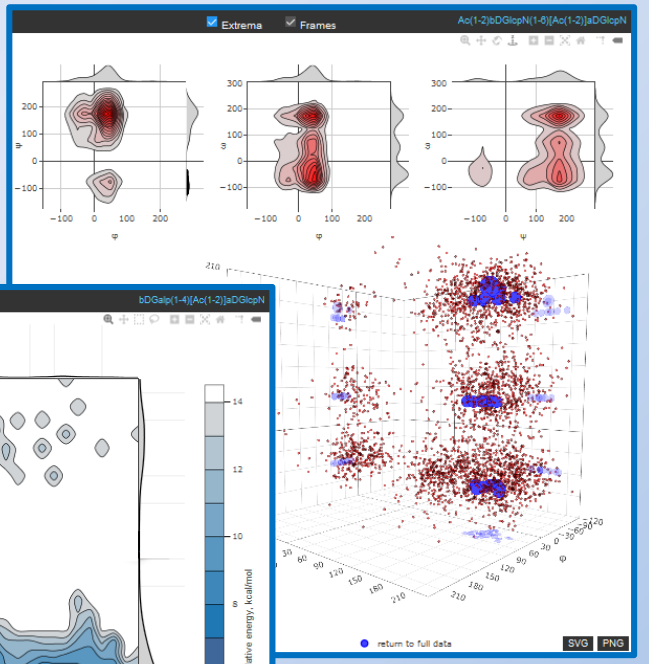
- поиск ID, структур, параметров расчета



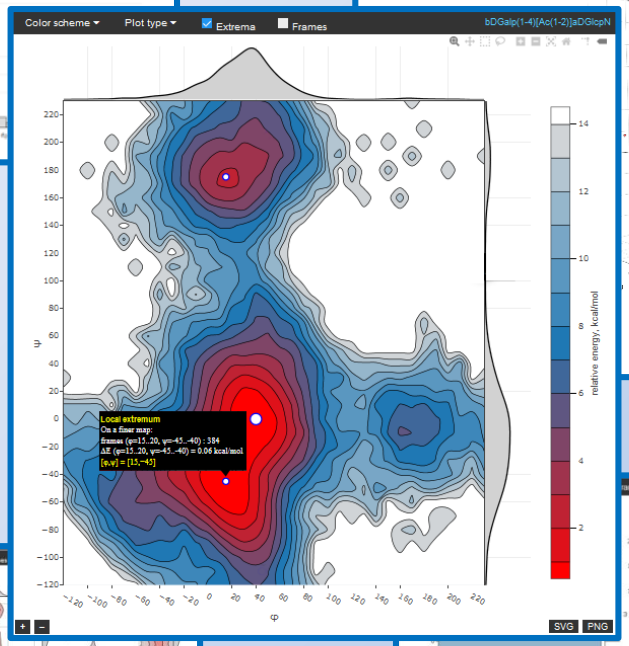
- работа с картами энергий и заселенности
- проекции, экстремумы, экспорт
- до 4 углов / связь
- интерактивность (прокрутка, масштаб, вращение, контроль плотности, цвет, слои)



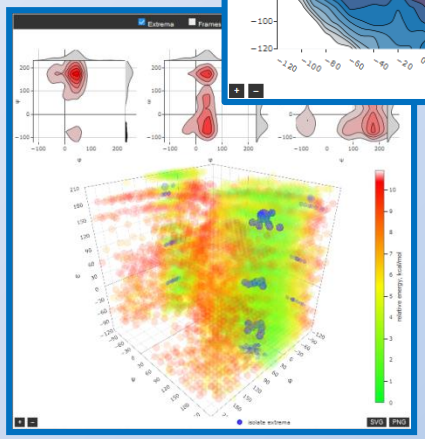
1D



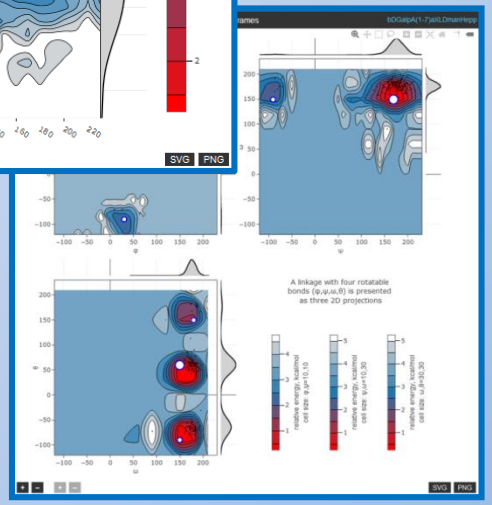
2D



3D



3D, ΔE



4D

Scherbinina S, Frank M, Toukach Ph **Carbohydrate Structure Database oligosaccharide conformation tool** *Glycobiology 2022, 32:460-468*
Scherbinina S, Toukach Ph **Three-dimensional structures of carbohydrates and where to find them** *Int J Mol Sci 2020, 21: id 7702*

Заполнение базы

анализ встречаемости фрагментов

автоматический конвейер (pipeline) для МД-расчетов

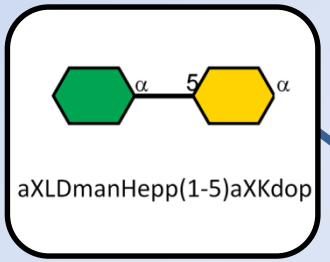
~300 / год

связи с 1..4 углами вращения в ди-, три- и тетрамерных фрагментах *(заполнение базы)*

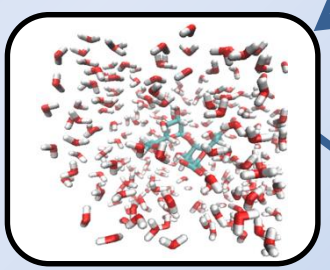
импорт GlycoMapsDB

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

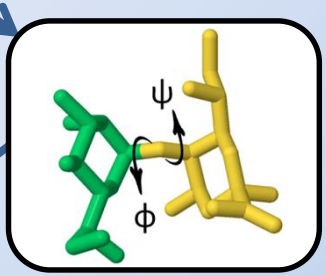
первичная структура *CSDB Linear*



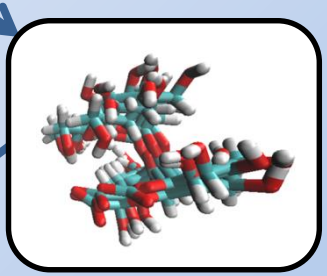
100-нс мол. динамика в явной H_2O *Packmol/Tinker*



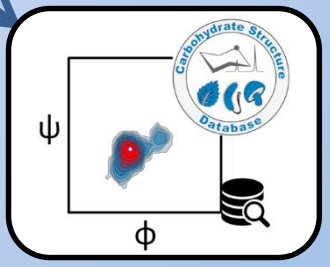
торсионные углы, экстремумы *own scripts*



начальная геометрия *REStLESS*



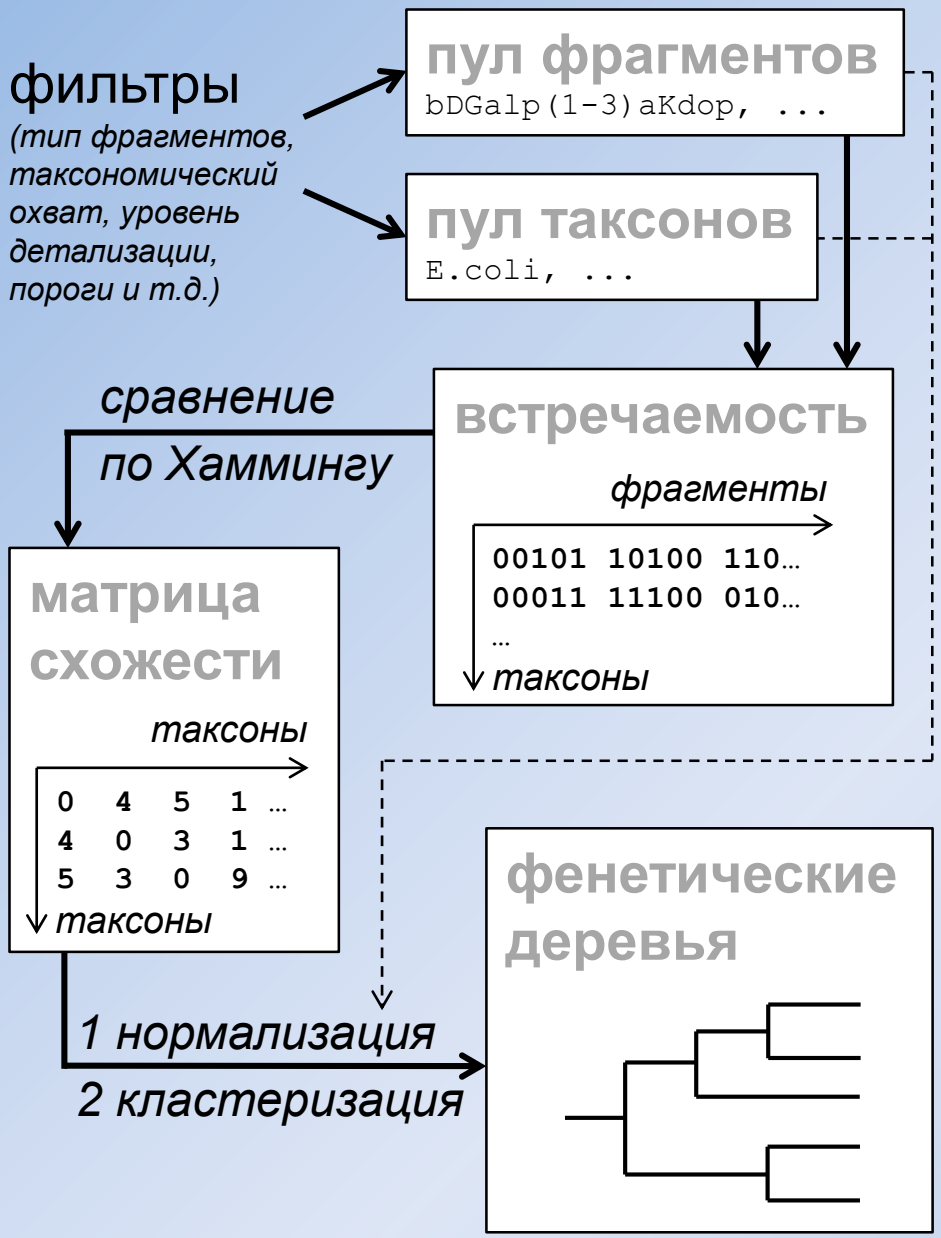
анализ траекторий *CAT*

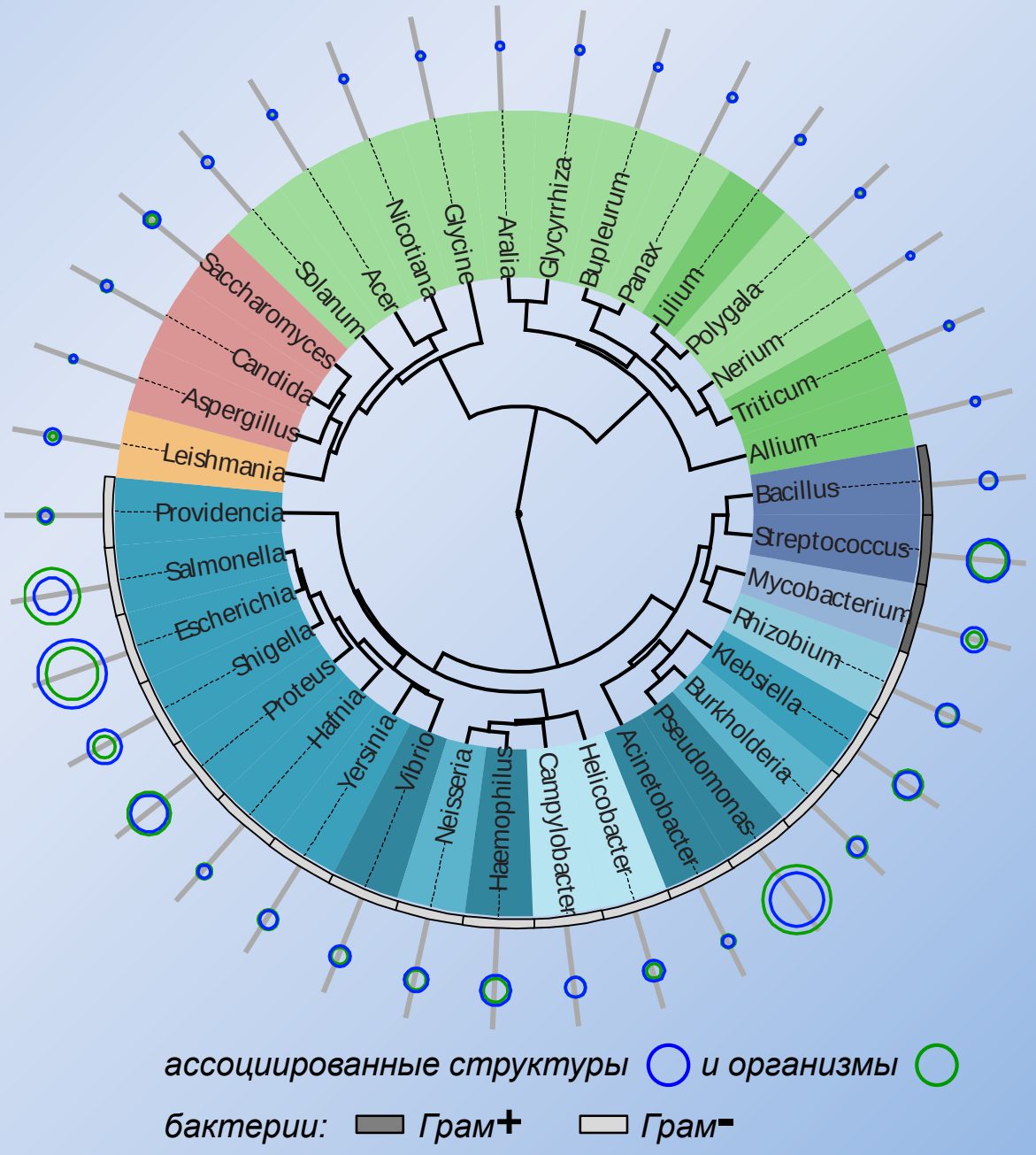


импорт карт в базу *CSDB importer*

Scherbinina S, Frank M, Toukach Ph **Carbohydrate Structure Database oligosaccharide conformation tool** *Glycobiology 2022, 32:460-468*
Stroylov V, Panova M, Toukach Ph **Comparison of methods for bulk simulation of glycosidic bond conformations** *Int J Mol Science 2020, 21:id7626*

Кластеризация таксонов





Структура → спектр ЯМР

Статистические

химические сдвиги ^{13}C и ^{1}H
HOSE на уровне атомов или остатков

☺ Опираются на существующие базы
☺ Прослеживается до публикаций
☹ Медленно (~ 1 мин)

Эмпирические

химические сдвиги, форма сигналов

☺ Очень быстро (~ 0.01 сек)
☹ Требуется модель
(для разных классов – разная)
☹ Требуются специальные БД

Квантовомеханические

геометрия + GIAO (все параметры ЯМР)

☺ Не зависят от баз данных
☹ Низкая точность (>3 м.д. в ^{13}C)
☹ Очень медленно (~ 1 неделя)
☹ Качество зависит от конформационной лабильности

Нейросетевые

химические сдвиги

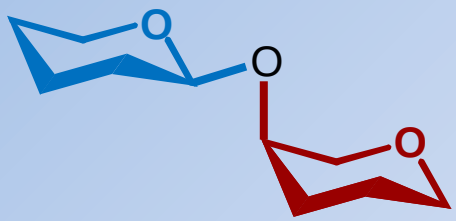
☺ Универсально
☺ Модно, можно получить грант
☹ Требуется обучение и референсная база
☹ Недоказательно

Малоизученные

(регрессия, ММ-QM)

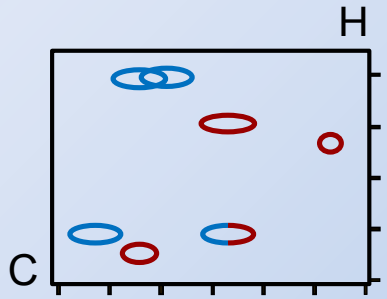
ЯМР-моделирование

ЯМР – основной метод анализа первичной структуры в гликобиологии

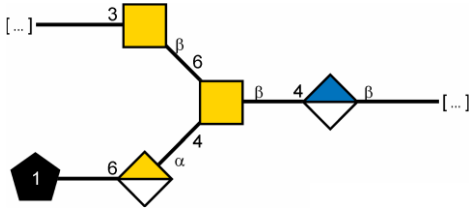


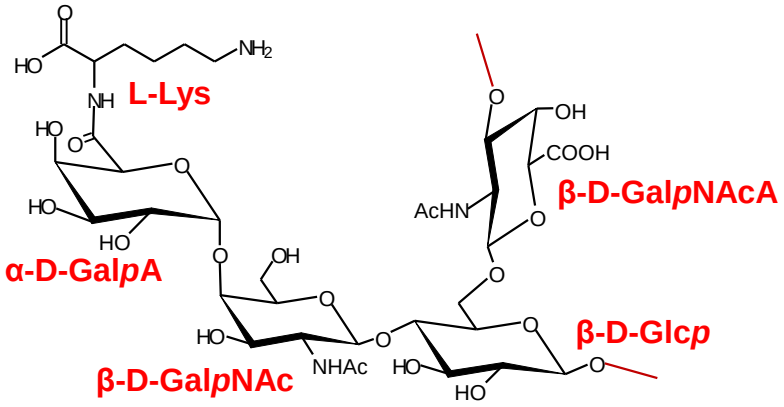
алгоритм **GODDESS**

Glycan-Optimized Database-Driven
Empirical Spectrum Simulation

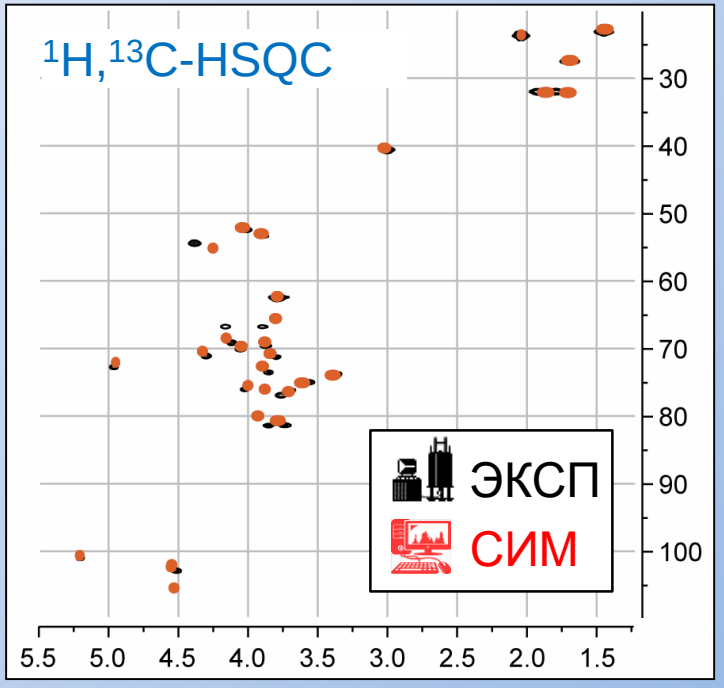


Proteus mirabilis G1



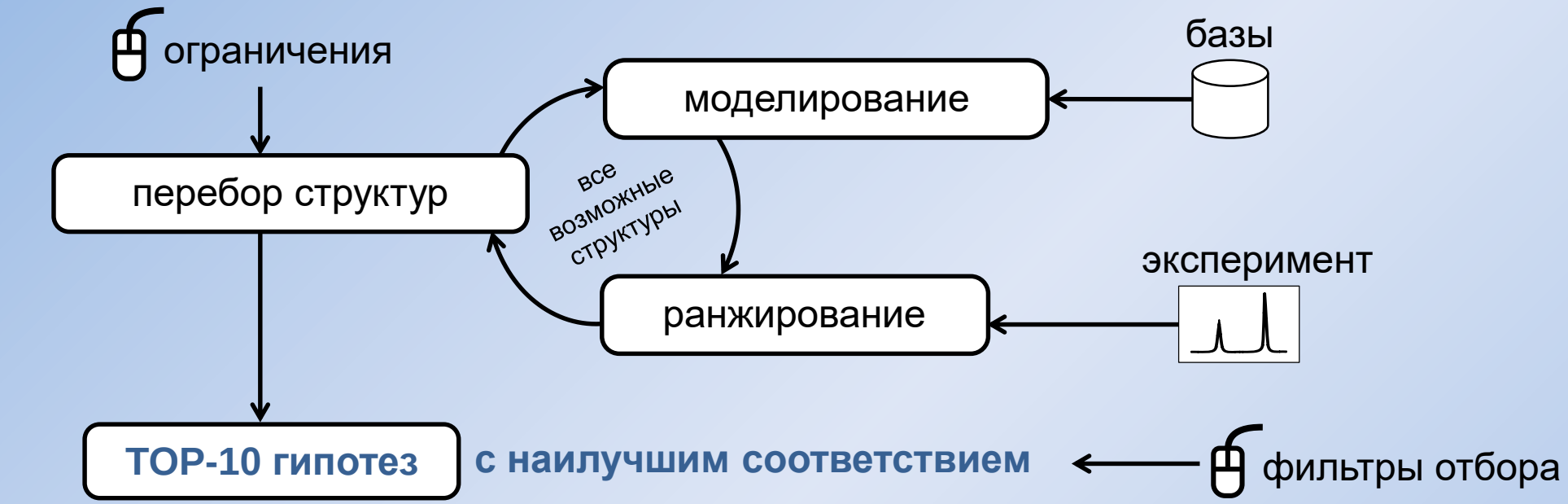


- ЯМР-симуляция (^{13}C ~ 0.7 ppm, ^{1}H ~ 0.06 ppm, COSY, TOCSY, HSQC, HMBC)
- Помощь эксперту
- Отнесение сигналов и проверка гипотез
- Верификация расчетной геометрии

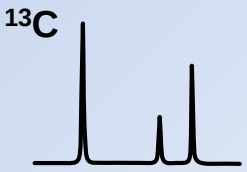


Toukach F, Ananikov V **Computational predictions of NMR parameters for structure elucidation of carbohydrates** *Chem Soc Rev 2013, 42:8376-8415*
Kapaev R, Egorova K, Toukach Ph **Carbohydrate structure generalization for simulation of experimental observables** *J Chem Inf Model 2014, 54:2594*
Kapaev R, Toukach Ph **Improved carbohydrate structure generalization scheme for ^{1}H and ^{13}C NMR simulations** *Analyt Chem 2015, 87:7006-7010*
Kapaev R, Toukach Ph **Simulation of 2D NMR spectra of carbohydrates using GODESS software** *J Chem Inf Model 2016, 56:1100-1104*

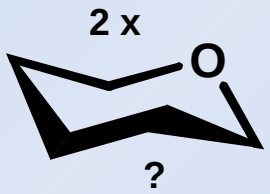
Спектр → структура



неотнесенный спектр ЯМР

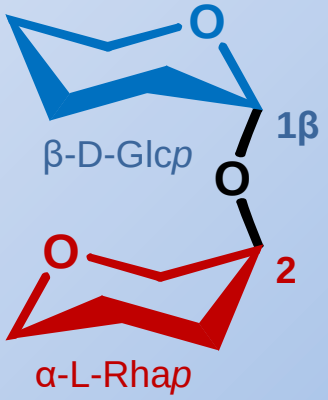


известные данные о структуре



алгоритм **GRASS**

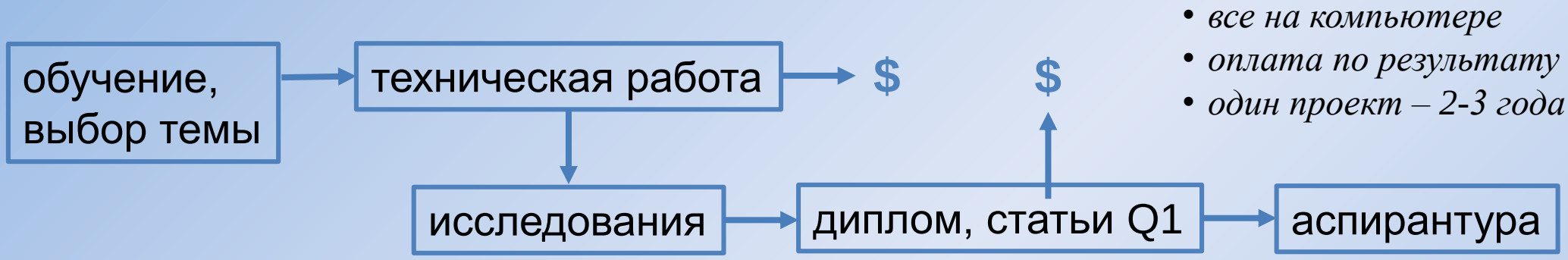
Generation, Ranking and Assignment of Saccharide Structures



полная структура (оставшиеся неизвестные)

мономеры, конфигурации, модификации, позиции замещения, последовательность

Чем можно заняться?



- *все на компьютере*
- *оплата по результату*
- *один проект – 2-3 года*

- молекулярные расчеты (автоматизация), генерирование конформационных карт
- нечеткое сравнение структур* и поиск структур с похожими мотивами
- симуляция масс-спектров, интеграция с программами их анализа
- предсказание стерических свойств и NOE (на основании конформаций фрагментов)
- автоматический сбор неструктурированных данных из других баз и Интернет (data-mining, аннотирование и верификация), алгоритмы выявления ошибок
- статистические исследования структур, выявление корреляций с таксономией, кластеризация организмов
- выявление углеводных компонентов в молекулярных графах* (напр., PDB)
- распознавание образов (иллюстраций в статьях), семантический анализ SNFG* и WURCS* сравнение и предсказание спектров ЯМР (оптимизация)
- другое (Ваши идеи?)

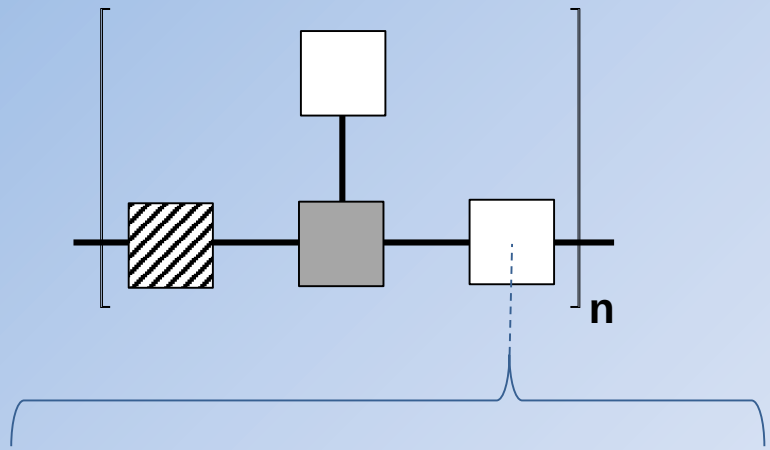
** программирование, предсказуемо, относительно быстро*

презентация группы гликоинформатики: http://toukach.ru/files/invi_2023.pdf

Дополнения

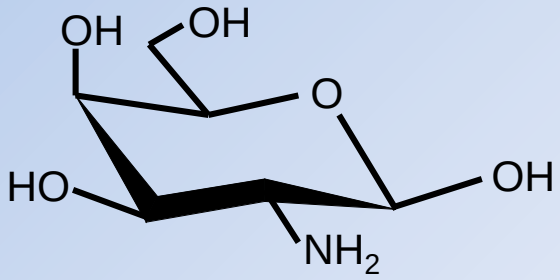
(для ответов на вопросы)

Структура углеводов

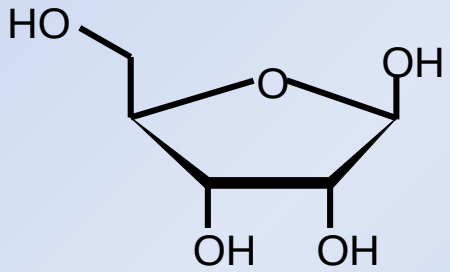


Полная структура

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



пример *альдо*-пиранозы (β-D-Gal*p*N)

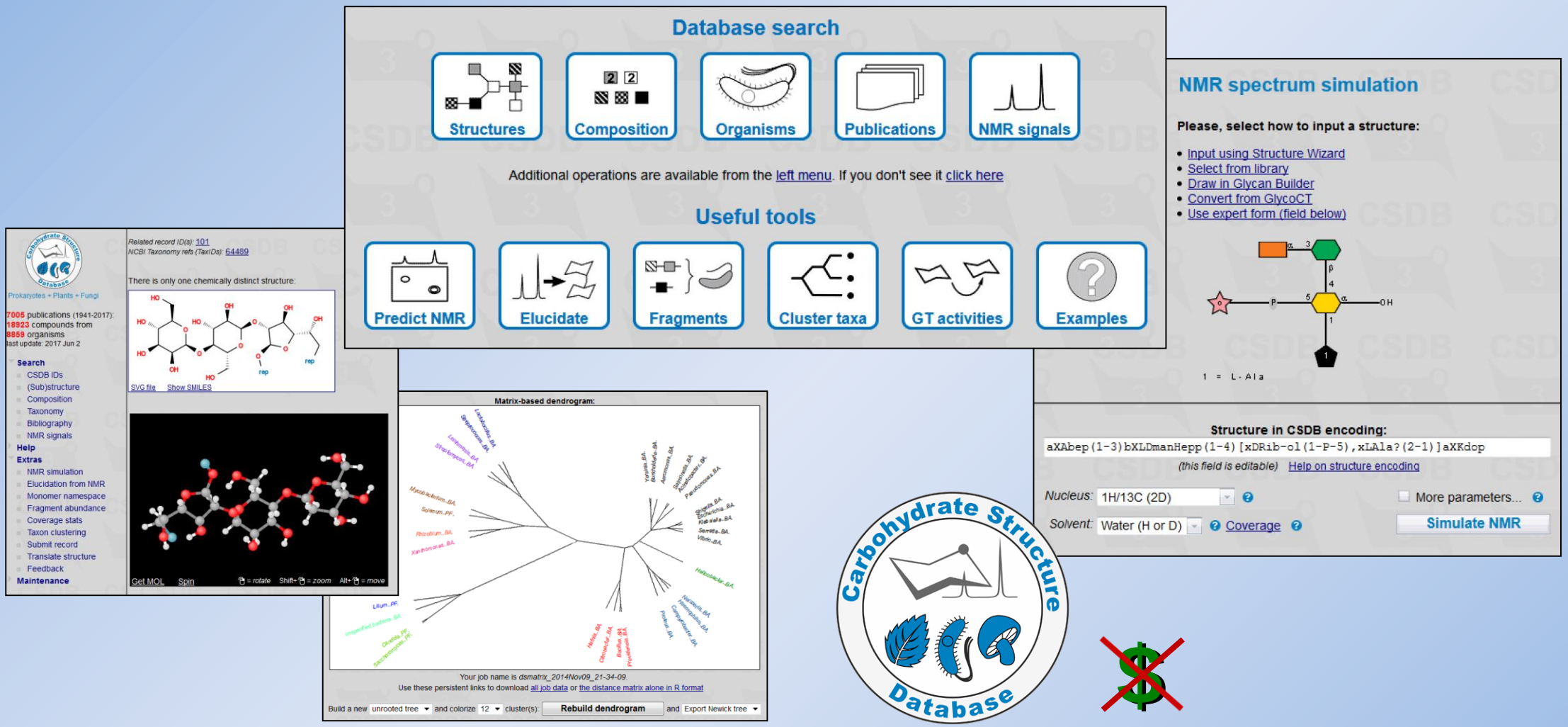


пример *альдо*-фуранозы (β-D-Rib*f*)

Структура остатка

- размер скелета (4-10)
- стереохимия всех центров (=*мономер*)
- размер цикла (*p/f/a, aldo/keto*)
- аномерная форма (α/β)
- абсолютная конфигурация (D/L)
- модификации (-NH_2,-COOH, deoxy)

Сайт в Интернете



http://csdb.glycoscience.ru

- свободный доступ
- подробная документация
- примеры решения задач

Примеры вопросов к CSDB

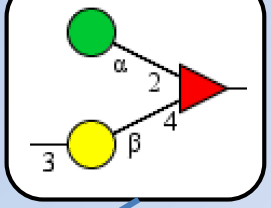
- Как введение аминогруппы влияет на химические сдвиги ЯМР в лактозном фрагменте?
- Какие структуры О-антигенов, содержащие галактуроновую кислоту и еще как минимум одну гексозу, были опубликованы после 2005-го года?
- Какие гликозиды, выделенные из растений рода паслёна, содержат агликон соланидин?
- Какие углеводы, кроме октозосодержащих, имеют сигнал ЯМР ^{13}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}C 3-О-α-абеквозил-6-деокси-β-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}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?

CSDB: составной запрос

фрагмент

Lys / Ala
↓ 2
↓ 6
SugA

фрагмент



состав

2 x KDO
1 x HexN
+ …

ИЛИ

И

И

класс / название

И

IDs

таксономия

Proteus O1

носитель:
человек

И

И

И

НЕ

ЯМР фильтр

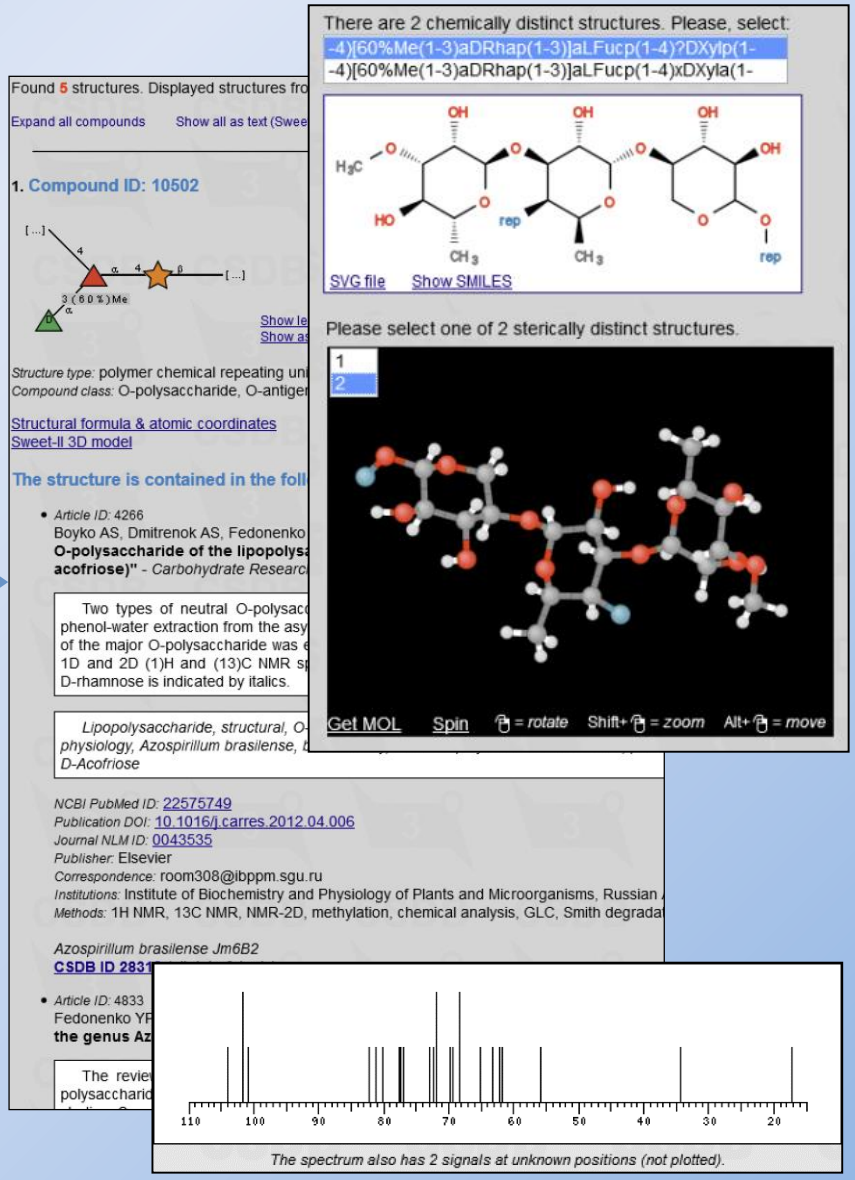
$\delta^{13}C$:
18.3 ± 0.8

библиография

≥ 2005

ключевое
слово: LPS

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



Структурные базы

Все БД

CarbBank — 23
☺ полная до 1996
☹ архитектура, % ошибок
ORIG

GlycomeDB → **GLYTOUCAN** — 120
☺ мета-репозиторий
☹ неполные аннотации
☹ нет агликонов

CFG glycan
млекопитающие, **> 6**

SweetDB, SugaBase → **glycoSCIENCES.DE** — **27 / 21**
PDB

$ **KEGG GLYCAN** — **11**

Eurocarb DB
☺ архитектура
☹ только модель

BCSDB — PFCSDB → Carbohydrate Structure Database, GlycosylTrans Database
15 / 6 (бактерии, археи)
17 / 6 (грибы, растения)
☺ полная
по прокариотам и грибам
☺ курируемая
ORIG
GTR

$ **GlycoSuite** By Proteome Systems
ORIG
млекопитающие+…
10 / 1
☺ полная до 2005

> 70
jCGGDB AIST
☺ коллекция баз
☹ слабо аннотированы

$ **UniCarbKB** — **4 / 1**
LC
☺ курируемая

0.7 O- & N- **nibrt** GLYCOBASE
HPLC
ORIG

44
Glycoconjugate Data Bank — Organized by Hokkaido University
PDB

0.2 *E. coli* **EcoDAB** — Escherichia coli O-Antigen Database
ORIG
GTR

34 (человек, мышь, крыса, вирусы)
GlyGen

0.3 животные
$ **GlycoBase**
ORIG

Специальные базы

JCGGDB AIST

- углеводные гены человека — GGdb — ~0.2
- $MS^{2,3,4}$ N- и O-гликанов — GMDB — ~0.2
- химические реакции — GlycoNAVI — ~3.0 (4.4 str)
- конъюгаты & агликоны — 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 — ~120

адгезия к патогенам — SugarBindDB (Pathogen Sugar-Binding Database) — ~0.9

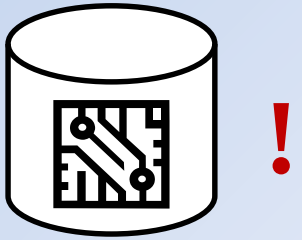
O-glycBase, O- и C-гликопротеины — CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS — ~0.2

GlycoMaps, расчетные конформационные карты — ~2.6

MSDB моносахариды и номенклатура — ~0.8

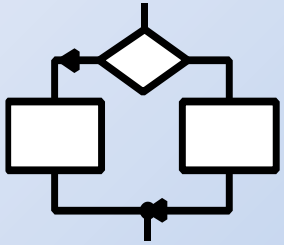
Критерии оценки

- **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

исправляемость

Редакторы структур

Углеводные, SNFG

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	++	++	+

Химические

программа	плагин	мономеры	результат	оптимизация
Hyperchem	Sugar Builder	30+26+user	3D, можно задать φ,ψ,ω	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*

Bochkov A, Toukach Ph **An online glycan builder with 2D and 3D structure visualization** *J Chem Inf Model 2021, in press*
Alocci D, Suchankova P, Costa R, Hory N, Mariethoz J, Toukach Ph, Lisacek F **SugarSketcher: quick online glycan drawing** *Molecules 2018, 23: id 3206*

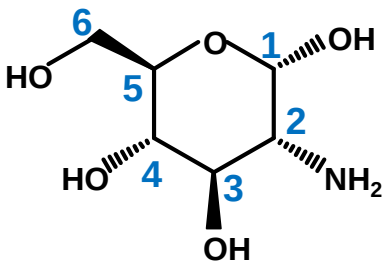
От семантики к атомам

анализ CSDB Linear или другой нотации

xRPyr(2-4:2-6)[xLAla?(1-2)]bDGlcpN
2 1 0

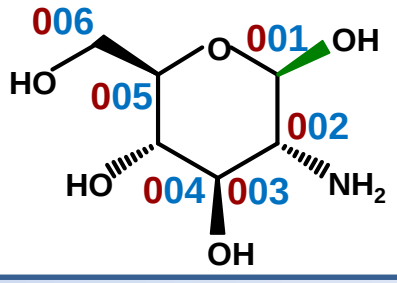
→ SMILES прототипов

0 α-D-Glc*p*N



стереоконфигурации + метки

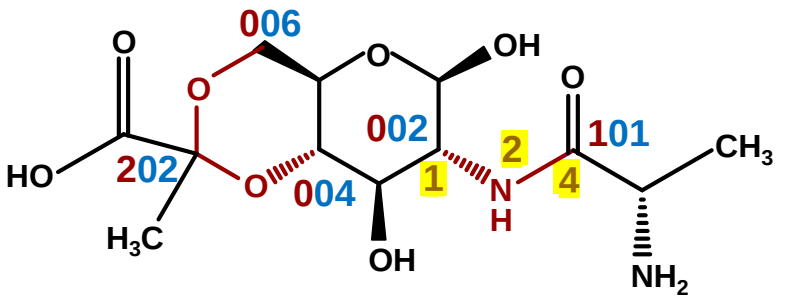
0 β-D-Glc*p*N



[006CH2](O)[005C@@H](O1)
[004C@@H](O)[003C@H](O)
[002C@@H](O)[001C@H](O)1

SMARTS связей между остатками

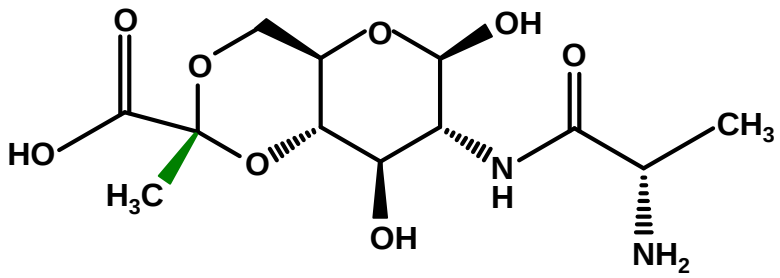
0+1, 1+2 Pyr-(2-4:6)-[L-Ala-(1-2)]-β-D-Glc*p*N



[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]

сборка + новые стереоцентры

R-Pyr-(2-4:6)-[L-Ala-(1-2)]-β-D-Glc*p*N



SMILES: 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

Любые операции на атомарном уровне
+
Внешние химические программы

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, (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

Как воспользоваться?

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

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

http://glytoucan.org

http://jcggdb.jp/index_en.html

http://glycosciences.de

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

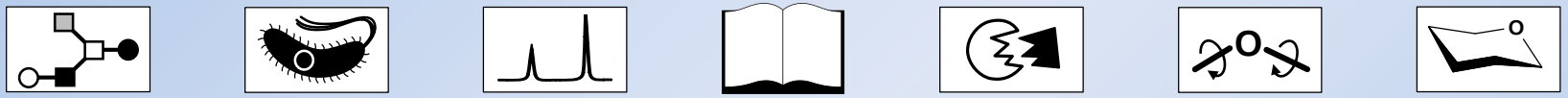
http://www.unicarbkb.org/

http://csdb.glycoscience.ru

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

Участники и спонсоры

База данных природных углеводов
Платформа для сервисов гликоинформатики



программирование	Роман Капаев, Андрей Бочков, Иван Чернышов, …
аннотирование и проверка данных	Ксения Егорова, Надежда Калинчук, Кирилл Казанцев, ...
общая поддержка и сбор данных	Юрий Книрель
интеграция, онтология	Рене Ранцингер, Кийоко Аоки-Киношита, Томас Люттеке, ...
конформационный анализ	Виктор Стройлов, Софья Щербинина, …
идеи, язык, архитектура, интерфейс, программирование, координация	Филипп Тоукач
партнеры	GlycomeDB, JCGGDB, UniCarbKB, GLYCOSCIENCES.DE

Российский Фонд Фундаментальных Исследований	Международный Научно-Технический Центр	Комиссия по грантам при президенте РФ	Немецкий Центр Исследования Рака	Фонд Содействия Отечественной Науке	Российский Научный Фонд
2005-2020 (x4)	2004-2005	2005-2007	2007-2011 (x4)	2008-2009	2018-2022 (x2)

Дальнейшее развитие

● сделано в CSDB ● предстоит сделать ● близко к завершению

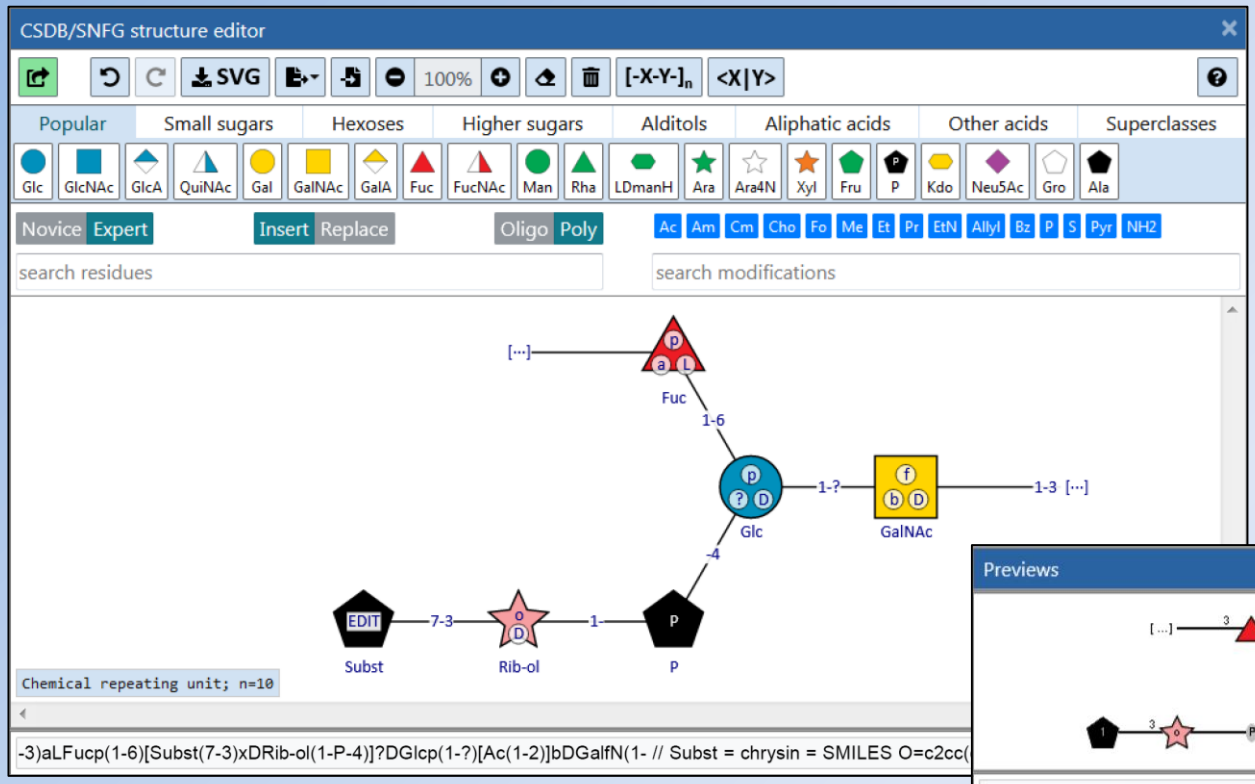
- **Стандартный человекочитаемый язык** (SNFG, CSDB Linear, …)
- **Расширение онтологий** (интеграция через GlycoRDF и GlycoCoO)
- **Кросс-проектные сервисы**

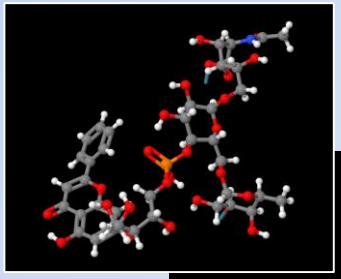
  (ввод-вывод структур, конформационные расчеты, предсказание свойств)
- **Стандартные индексы**

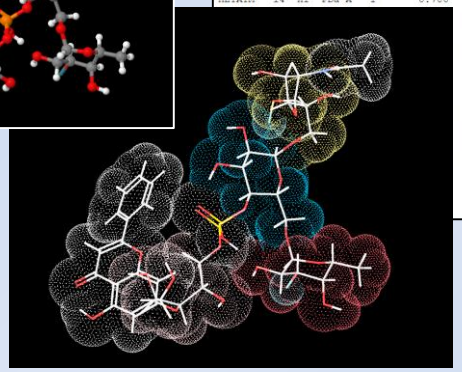
  (Glytoucan ID, MSDB, PMID, DOI, TaxID, ICD-11, PDB id, Genbank, …)
- **Стандартные протоколы** (API, WSDL, SPARQL, REST, …)
- **Идеологическая замена CarbBank** 
- **Требование включать ID в публикации** (Glytoucan ID?)

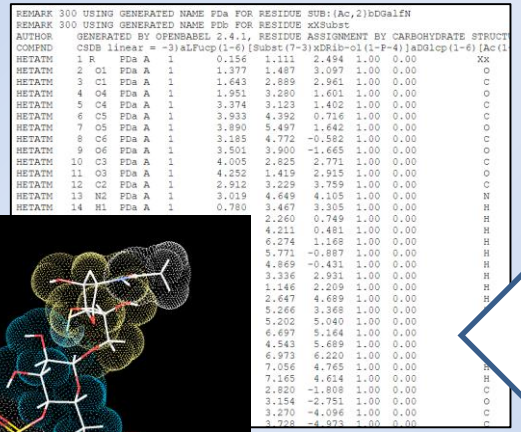
  (кто будет чистить базы от неопубликованных/ошибочных данных?)

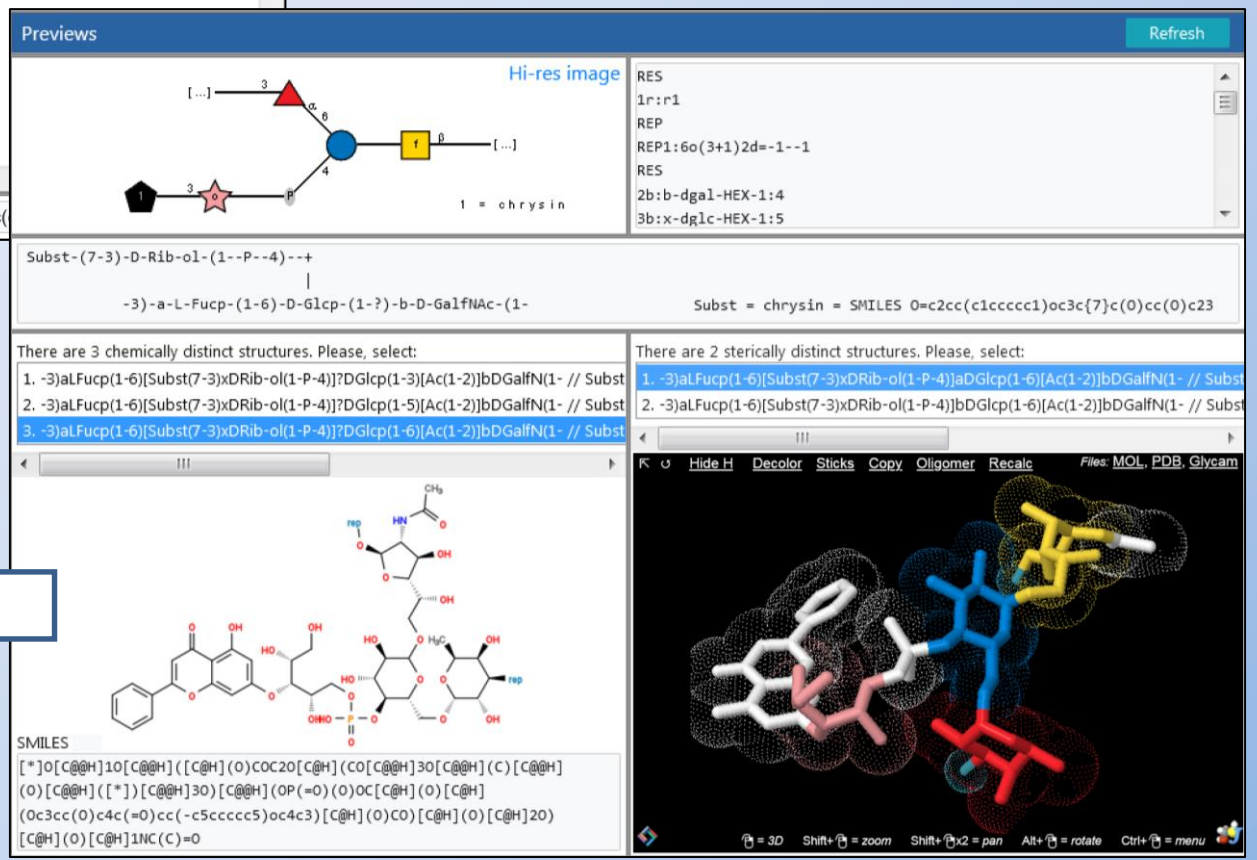
Ввод и вывод структуры









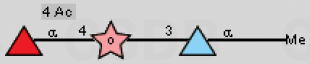


Поиск по фрагменту структуры

Structure wizard

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

Structure: aLFucp4Ac(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me



Residue Ⓐ: aLFucp4Ac(1-

a L fucose (pyranose)
aLFucp
substitutes C4 of Residue B
is terminal

add substitution
add substituent Acetylated at 4
add substituent
add substituent
add substituent

Residue Ⓑ: xDRib-ol(1-

D ribose (alditol)
DRib-ol
substitutes C3 of Residue C

add substituent
add substituent
add substituent
add substituent

Residue Ⓒ: a?6dTal?(1-1)Me

a ? 6-deoxytalose (?)
a?6dTal?
has aglycon : methyl

add substituent
add substituent
add substituent
add substituent

Structure in CSDB encoding:

Ac(1-4)aLFucp(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me

Return the structure to the search page and close this window

Home Help

Glycan Builder

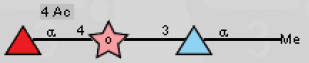
File Edit Structure View Help

Linkage ? 1 -> --- 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 (aDFucp3N)
- Use expert form (field below)



Structural fragment in CSDB encoding:

Ac(1-4)aLFucp(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me

(this field is editable) Help on structure encoding

Only those containing text: factor-2A' 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 fragment (substructure)
Search for molecule types: All molecule types
Search for structures with published NMR data only
Restrict compound class: inner core component
Restrict taxonomical domain: All prokaryotes

Previous results: 122 structures: <ID list>

Go! & display 30 records per page.

Predict NMR Sweet 3D model GLYCAM 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)

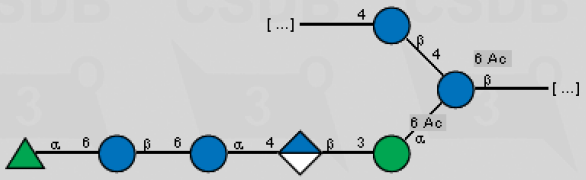
1. (*Organism ID:* 1005)

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:

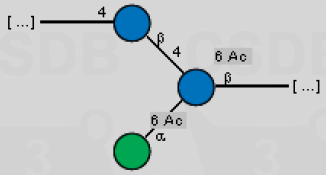
- *Compound ID:* 1717



Show legend
Show as text

Carbohydrate Research 2004, "Synergistic interactions between the genetica
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: Absidia, **Acetobacter**, Acholeplasma, Acidithiobacillus, Acinetobacter, Acremonium, Actinobacillus

Species: *Any*, sp., diazotrophicus, methanolicus, tropicalis, **xylinum**

Specify: *

Strain / subspecies: *Any*, B42, CKE5, CKEP, CR1/4, IFO 13693

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)

☐ Search among HOST organisms
☐ Use NCBI taxID
☑ Include subtaxons

Previous results: 6 structures: <ID list>

Go! & display 30 records per page.

List of organisms Home Help

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

Genus: Acetobacter *Species:* xylinum Process

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 str of Yersinia pestis

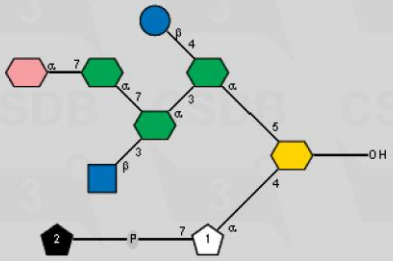
Carbohydrate Research **340(9)** (2005) 1625-1630

Following a report of variations in the lipopolysaccharide (LPS) structure degrees C) and flea (25 degrees C) temperatures, a number of changes to the the bacterium was cultivated at a temperature of winter-hibernating rodents (6 known Y. pestis LPS types, LPS of a new type was isolated from Y. pestis KM21 the latter differs in: (i) replacement of terminal galactose with termin 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-deox (di)oxygenation of a fatty acid(s). The data obtained suggest that cold temp 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

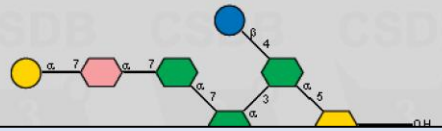


1 = a-Kop
2 = Et N

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 č 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

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)

Previous results: 3 publications: 2953,201,1525

Publications with structure elucidation only

Restrict taxonomical domain: All domains

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: *Type CSDB conformation ID or range, e.g. 1-5,10,12* — ANY — Search!

Model: *Use selectors or wizard* — α | L-Rhap | →4 | β | D-GlcpNAc
(only those components are listed for which conformation maps are stored)

or type dimeric fragment in CSDB encoding
aLRhap(1-4)[Ac(1-2)]bDGlcpN

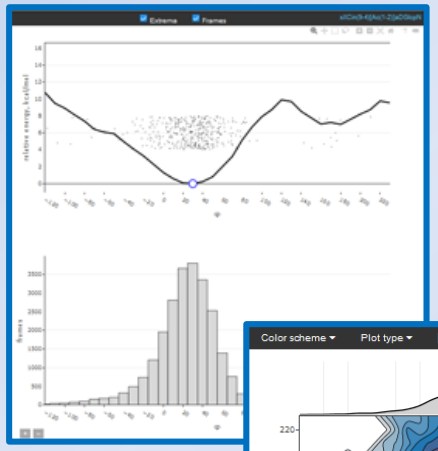
Force field: *Filter by MD method* — MM3-2000

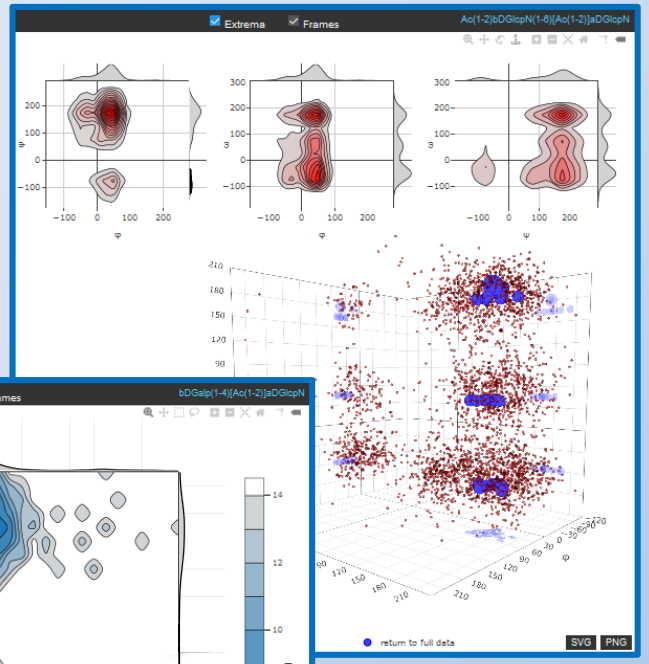
Temperature: *Filter by MD temperature* — 1000

Solvent model: *Filter by solvent model* — any

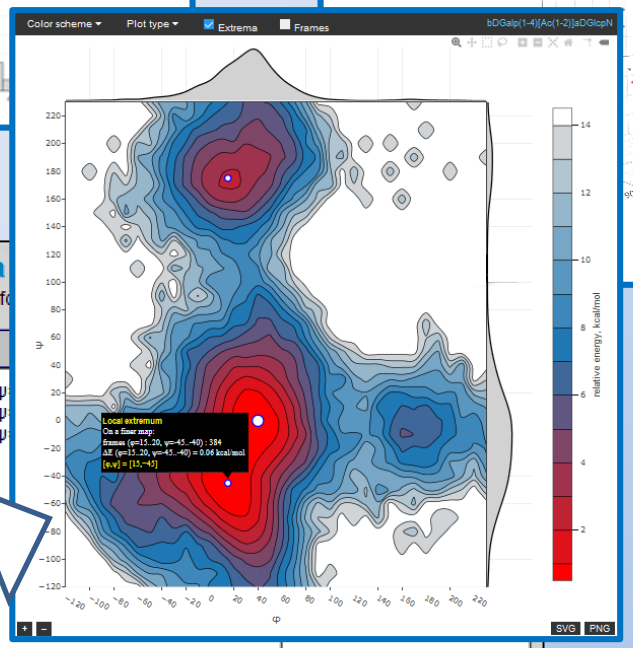
Home Help

1D





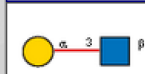
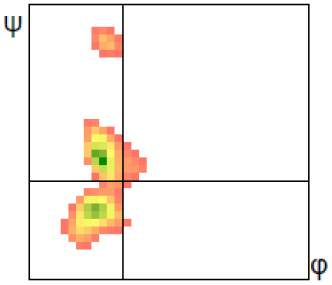
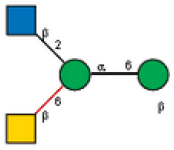
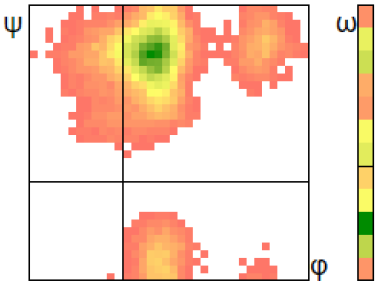
3D



2D

CSDB conformation data

12 conformation maps have been f

Model structure	Conformation map	
	 map live view ID: 1610	φ=-30, ψ= φ=-35, ψ= φ=-25, ψ=
 (bond: )	 map live view ID: 907	φ=25, ψ=165, ω=-75 ΔE=0.00 Kcal/mol φ=45, ψ=170, ω=-60 ΔE=0.00 Kcal/mol φ=45, ψ=195, ω=-75 ΔE=0.38 Kcal/mol φ=50, ψ=185, ω=-60 ΔE=0.72 Kcal/mol φ=25, ψ=165, ω=180 ΔE=0.72 Kcal/mol φ=30, ψ=155, ω=-60 ΔE=0.72 Kcal/mol φ=35, ψ=180, ω=180 ΔE=0.85 Kcal/mol φ=45, ψ=215, ω=-60 ΔE=0.85 Kcal/mol φ=40, ψ=170, ω=165 ΔE=0.99 Kcal/mol φ=20, ψ=185, ω=165 ΔE=0.99 Kcal/mol φ=55, ψ=195, ω=165 ΔE=0.99 Kcal/mol φ=55, ψ=155, ω=-60 ΔE=1.13 Kcal/mol φ=45, ψ=185, ω=165 ΔE=1.13 Kcal/mol φ=30, ψ=145, ω=-60 ΔE=1.13 Kcal/mol	Force field: MM3-1996 Solvent model: None MD temperature: 1000 MD duration: 30 ns Frames: 30K MD summary file: download

Поиск гликозилтрансфераз

CSDB glycosyltransferase search

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

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

Organism: *Select species* Escherichia coli — *Type strain/serogroup*

Molecule role: *Filter by target structure* O-antigen

Synthesized bond: *Type dimeric fragment in CSDB encoding or use tools* ANY(1-?)?DGlcpN Use Wizard

Donor & acceptor: *Type donor CSDB encoding or use tools* ANY Use Wizard

Type acceptor CSDB encoding or use tools ANY Use Wizard

Treat donor/acceptor as fragments

Confirmation status: *Filter results to those* evidenced or predicted

Search!

Home Help HELP !!!

CSDB glycosyltransfer

42 glycosyltransferase activities ha
Please note that GTR database covers only two species: *A*

Enzyme	Gene	Activity		
Name: WbbD UniProt ID: Q03084*	?	Synthesized dimer: bDGalp(1-3)aDGlcpN Donor (ID 19342): DGalp(1-P-P-5)nucU Acceptor (ID 19715): Ph(1-11)[Ac(1-2)aDGlcpN(1-P-1)]Subst // Subst = undecan-1,11-diol Status: evidence *in vitro* Confirmation methods: *in vitro* (crude extract) ID: 2053	Organism (ID 8 Full structure (I Analog i	
Name: WbbG UniProt ID: Q0H8C8*		Synthesized dimer: aDGlcp(1-3)aDGlcpN Status: indirect evidence *in vivo* Confirmation methods: mutation (knockout) Notes: Repeating unit of the O148 antigen. ID: 2151	Organism (ID 8 Full structure (I CSDB ID(s): 11572, 21578, 23062, 26257, 27289 Molecule role: O-antigen	
Name: WbaD UniProt ID: Q1L815*	Name: wbaD GenBank ID: 7156002*	Synthesized dimer: bDManp(1-3)aDGlcpN Donor (ID 19855): DManp(1-P-P-5)nucG	Organism (ID 1863): *Escherichia coli* O77 Full structure (ID 4600):	**Zhou et al. 2016** DOI: 10.1016/j.carres.2016.02.007 Wang et al. 2007 DOI: 10.1099/mic.0.2007

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 occurence 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 by (*). '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 occurence bit-codes. Show
Generated dissimilarity matrix. Show
```

Calculation parameters:

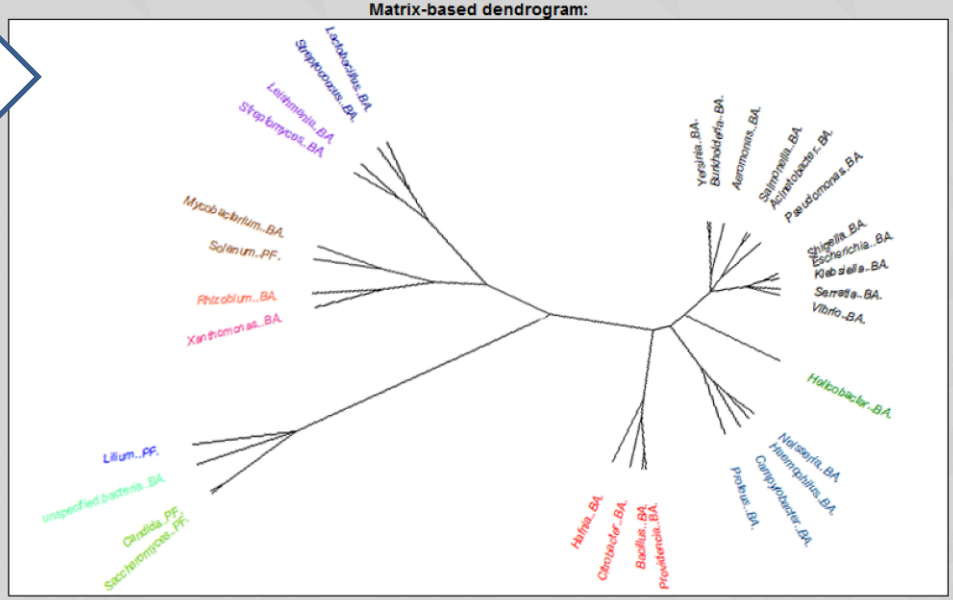
```
Hamming mode: YES
Fragment size: monomer
Fragment abundance filter: instance threshold: 550
Fragment abundance filter: structures threshold: 500
Fragment presence threshold: 2
                                  (atypical residues): NO
                structures of this type: any
Filter: differentiate 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

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

Allowed linkages:

	C1	C2	C3	C4	C5	C6	C7+
D-Gal*p*NA			✓	✓		✓	None
Ac	✓						
D-Rib-ol	✓	✓	✓	✓		None	
P							
α-Oct*p*			✓	✓	✓	✓	None
L-Ala	✓	✓		None			

Advanced options: Hide

	Min in	Max in	Location	Ac at N	Acceptors	Remove
D-Gal*p*NA	1	2	any	demanded	any	☒
Ac					1	
D-Rib-ol	?	?	any		any	☒
P			terminal		3,5	☒
α-Oct*p*	?	?	reducing			☒
L-Ala	?	?	any	forbidden	1	☒

Search depth	Scope	Advanced scope
Widespread structures only	✓ oligomers ☐ polymers⚠	β-anomers: = 1	CH_2 carbons: ?	☐ no furanoses

Find best matching structures:

Experimental ^{13}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
```

± 2 signals

✓ Find 15 best-fitting structures

✓ Save generated structures

Go!

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

Top 15 matches:

#Rank **Mean deviation** Linear correlation RMS deviation Trustworthiness	Structure Experimental spectrum Simulated spectrum Comments
#1. **Δ ~ 0.94 ppm** Corr = 1.000 RMS dev = 1.46 ppm Trust = 46%	Sim assignment Structure as text 1 = b-D-GalpNA 2 = L-Ala Expt: Sim:
#2. **Δ ~ 0.95 ppm**	
#15. **Δ ~ 1.42 ppm** Corr = 0.999 RMS dev = 1.99 ppm Trust = 49%	Sim assignment Structure as text 1 = b-D-GalpNA 2 = L-Ala