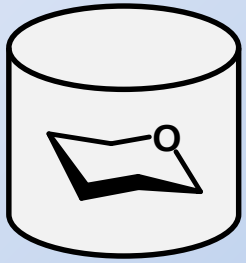
лаб. 30, 21

Филипп Тоукач

Информационные технологии в структурной гликохимии и гликобиологии

доклад для защиты диссертации д.х.н. (2018) с дополнениями (2020)



2005-н.в.

вер. 1.7

В этом докладе:

введение
- Зачем информатизировать гликомику?

глико-информатика
- Углеводные базы данных
- Правила гликоинформатики на примере CSDB

структуры
- Формализация углеводных структур
- Молекулярная геометрия углеводов

надстройки
- Гликотаксономия
- Гликозилтрансферазы

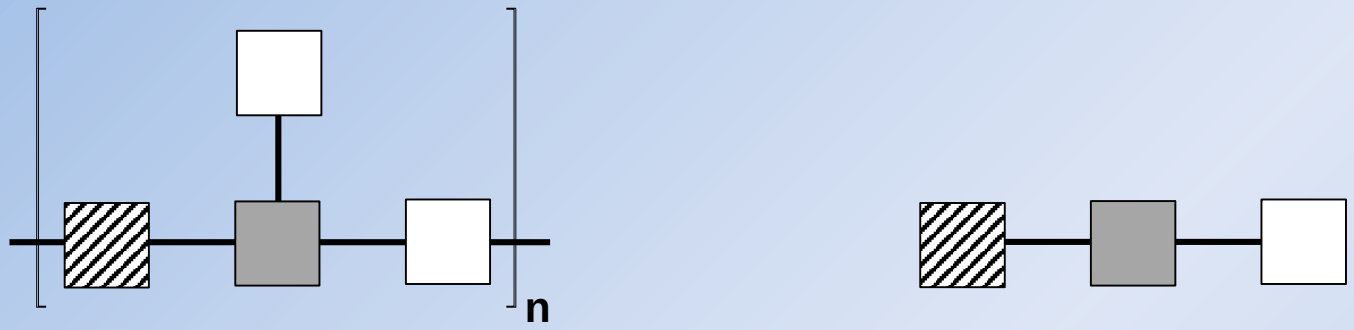
глико-ЯМР
- Моделирование спектров ЯМР
- Предсказание первичной структуры

выводы
- Что сделано, опубликовано, внедрено?

http://toukach.ru/rus/glycoinf.htm

ссылки указаны в случае, если статьи опубликованы в соавторстве с докладчиком

Гликохимия в биологии



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

Гликоинформатика

- **Роли углеводов:** гликозилирование белков, антигены микроорганизмов, межклеточные контакты, клеточная стенка, биоактивные гликозиды.

- Интерес рос, объём информации увеличивался, но **использовать её было проблематично**.

- **Глобальная задача гликоинформатики** - обеспечить исследователей углеводов всей мощью информационных технологий.

>2000 г.

Частные задачи гликоинформатики

- **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 и визуализация молекул** (в т. ч. в публикациях)

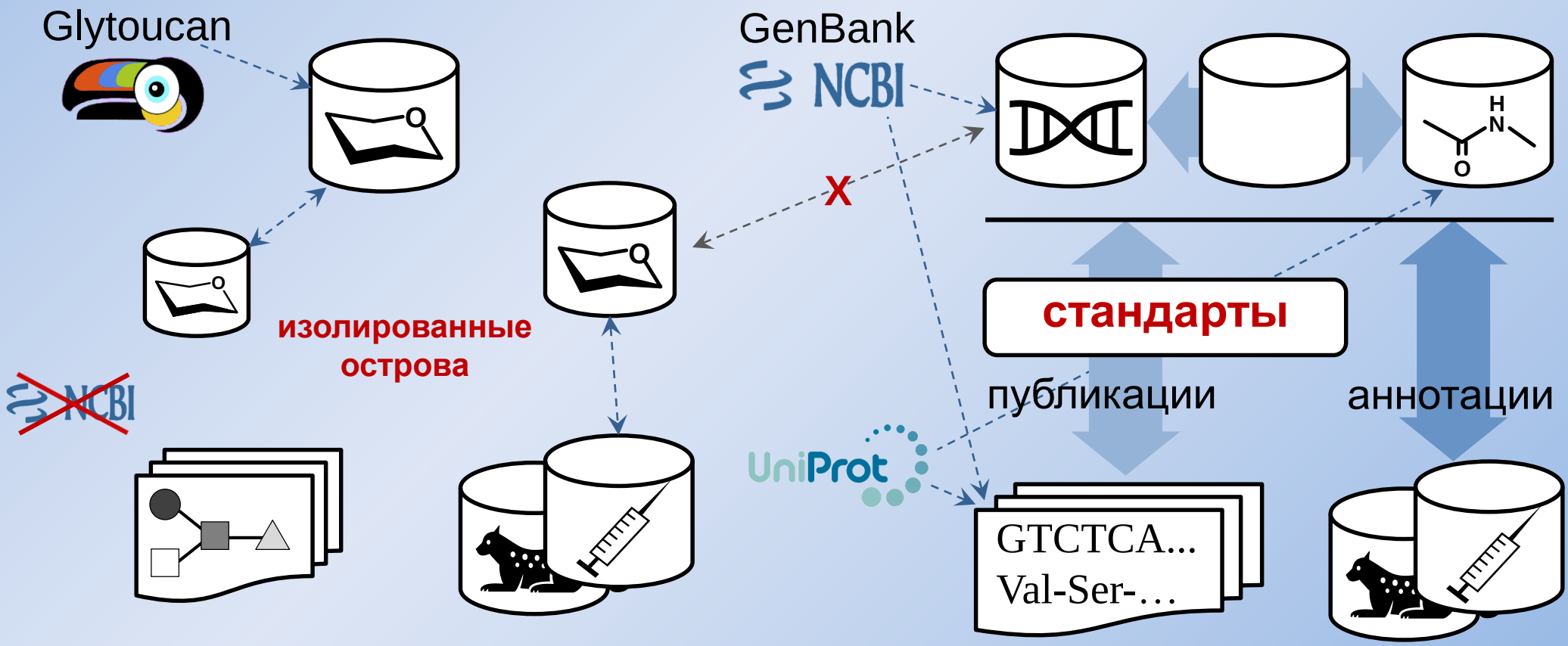
- **Методология обработки накопленных знаний**

  Аннотирование публикаций, интеграция проектов, стандартизация знаний

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



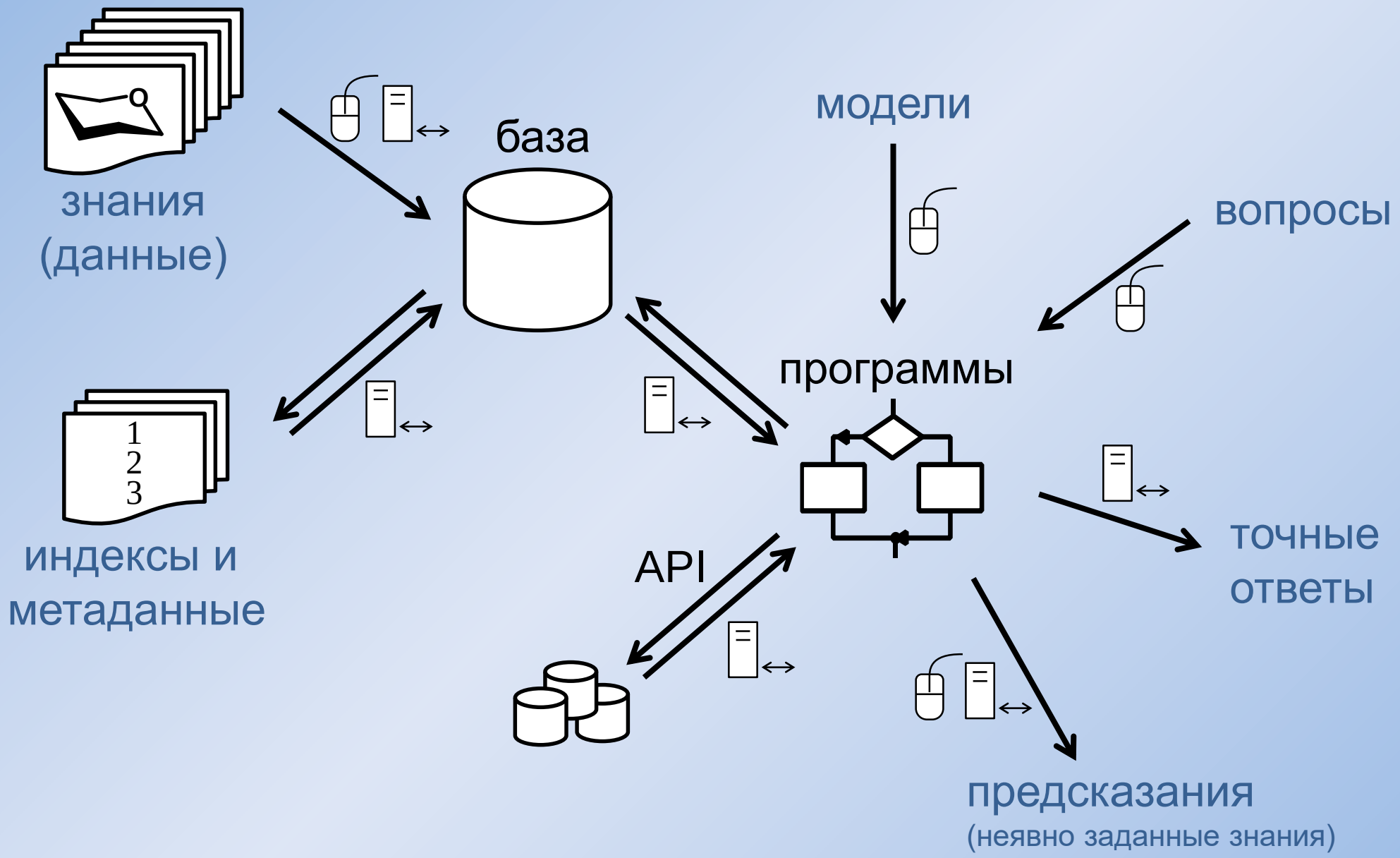
Проблемы в гликоинформатике

- 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*
Egorova K, Toukach Ph **Bridging isolated islands in the sea of data** *Angewandte Chemie Intl Ed 2018*

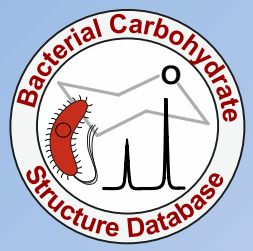
Базы данных



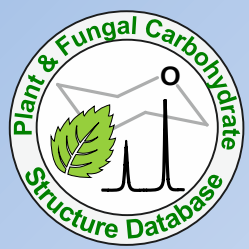
с участием человека

автоматически

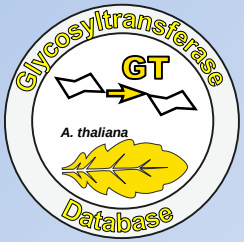
Carbohydrate Structure Database



с 2005

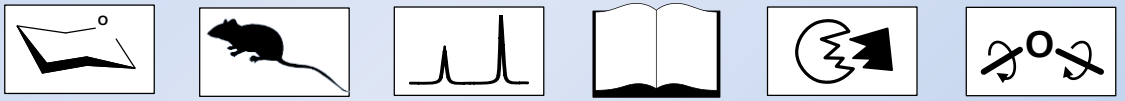


с 2012

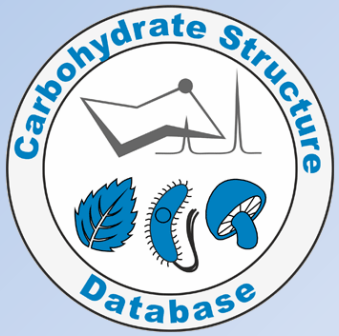


с 2017

Комплекс баз данных + Платформа для сервисов



25K 13K 14K 10K 2K 3K

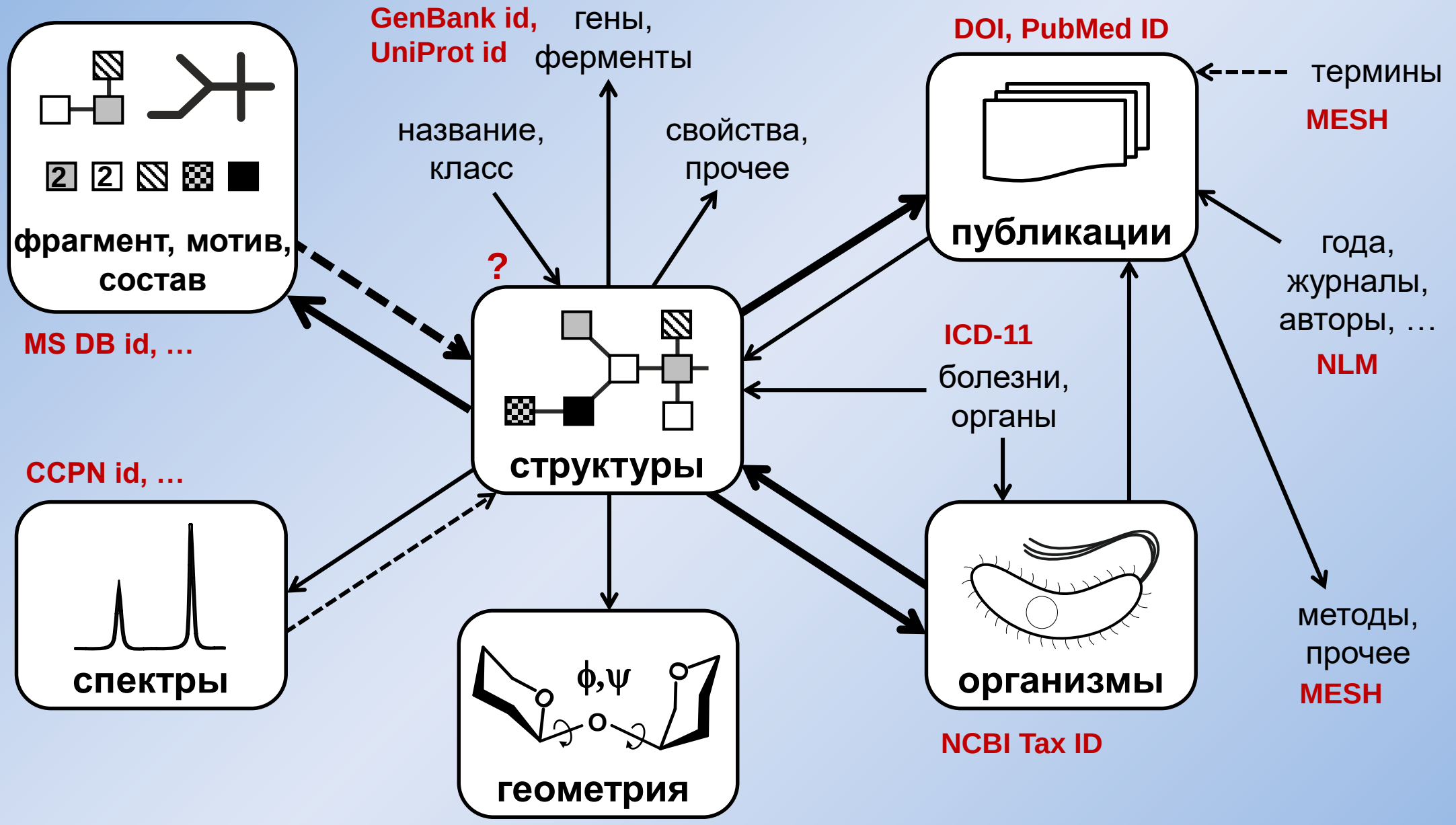


CSDB

- регулярное пополнение
- расширяемая архитектура
- множество инструментов обработки данных
- проверка контента (15% = CarbBank, 85% = литература)
- полное покрытие (бактерии, грибы)
- интеграция с другими базами

Egorova K, Toukach Ph **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017*
Toukach Ph, Egorova K **Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts** *Nucl Acid Res Database Issue 2016*
Egorova K, Toukach Ph **Expansion of coverage of Carbohydrate Structure Database (CSDB)** *Carbohydr Res 2014*
Toukach Ph **Bacterial Carbohydrate Structure Database 3: principles and realization** *J Chem Inf Model 2011*

Типичные запросы



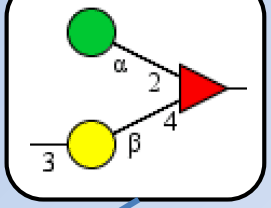
→ однозначные переходы ⇢ нечеткие переходы комбинации: **И, ИЛИ, НЕ**

CSDB: составной запрос

фрагмент

Lys / Ala
↓ 2
↓ 6
SugA

фрагмент



ИЛИ

состав

2 x KDO
1 x HexN
+ …

И

И

класс / название

И

IDs

таксономия

Proteus O1

носитель: человек

И

И

И

ЯМР фильтр

НЕ

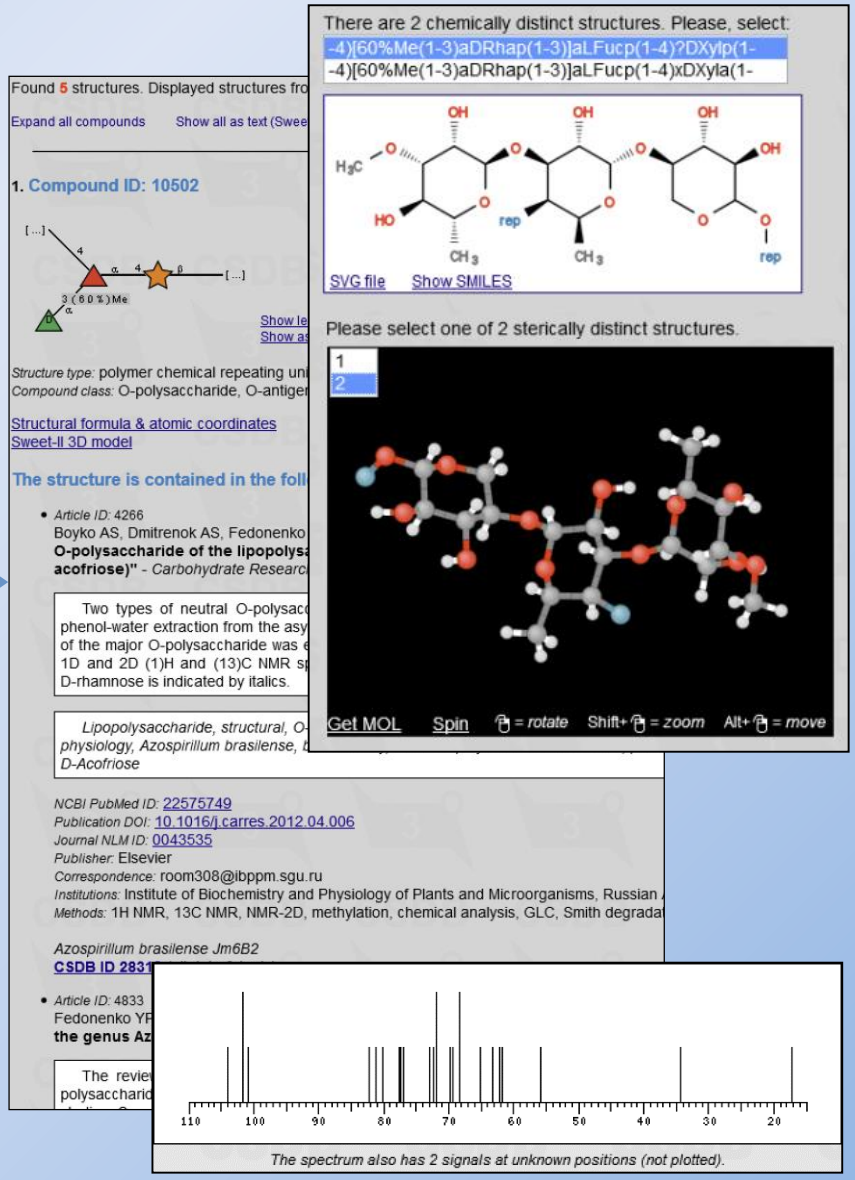
δ^{13}C:
18.3 ± 0.8

библиография

≥ 2005

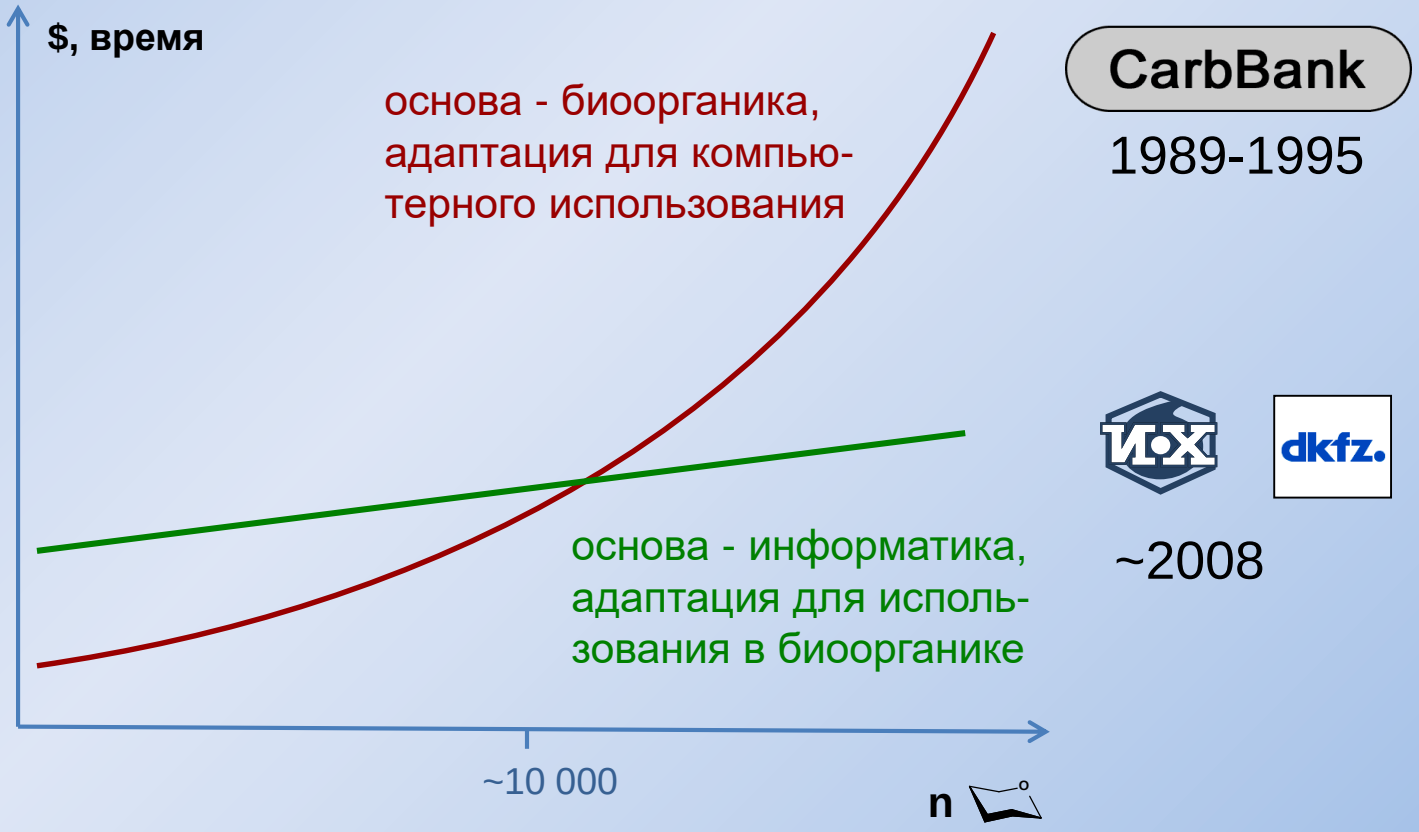
ключевое слово: LPS

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



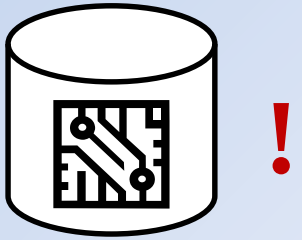
Организация разработки

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



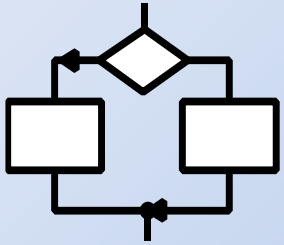
Критерии оценки

- **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

Архитектура

UniCarbKB GlycomeDB Eurocarb DB Carbohydrate Structure Database

- 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

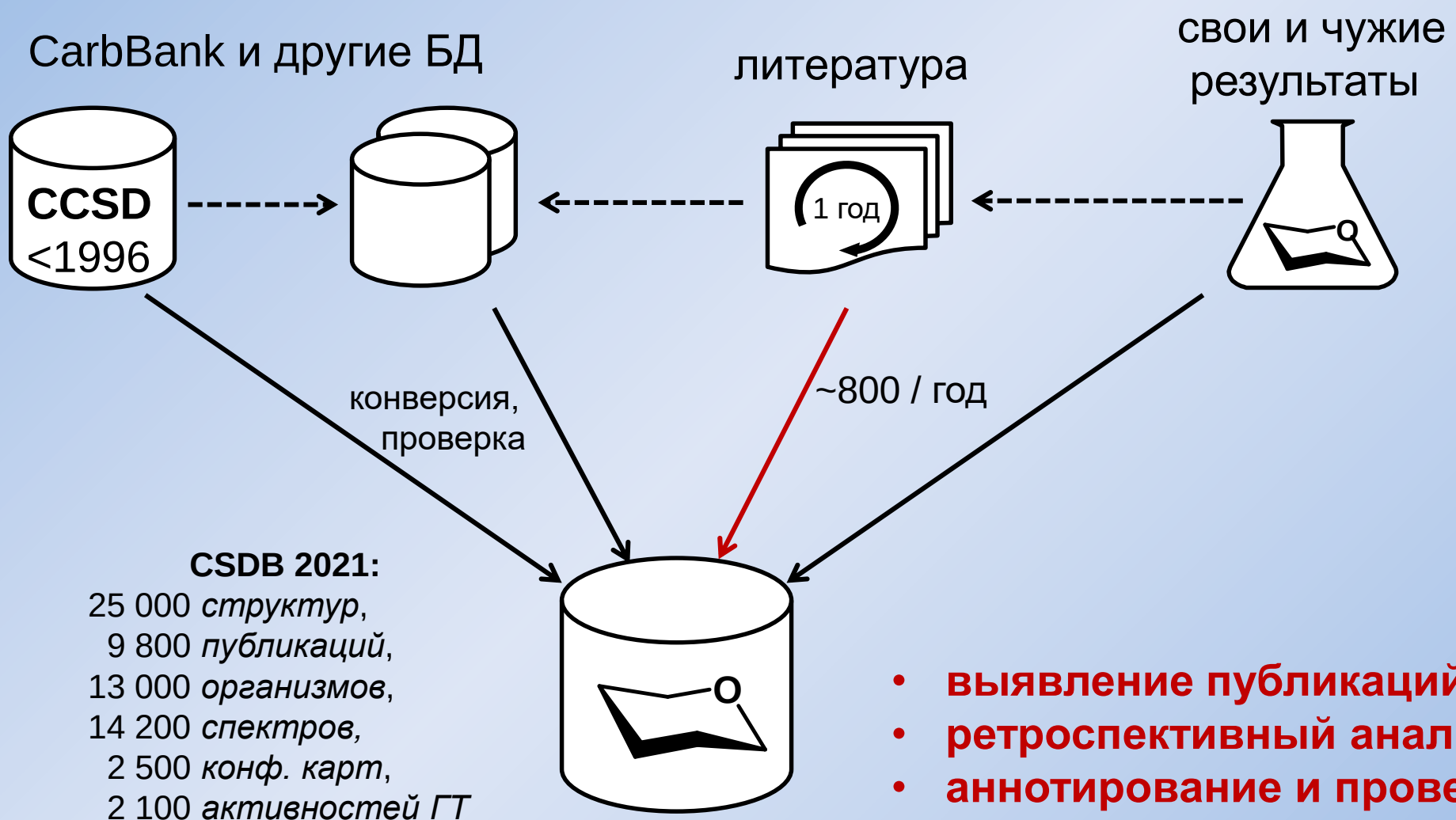
дополнительно

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

болезни
органы, ткани
генотип, стадия

ключевые слова
рефераты
институты
методы

Источники данных



CSDB 2021:

25 000 *структур*,
9 800 *публикаций*,
13 000 *организмов*,
14 200 *спектров,*
2 500 *конф. карт*,
2 100 *активностей ГТ*
+взаимосвязи

- **выявление публикаций**
- **ретроспективный анализ**
- **аннотирование и проверка**

полное покрытие по прокариотам:
отрицательный результат поиска =
значимая научная информация

UniCarbKB
GlycomeDB
Carbohydrate Structure Database

Качество данных

операторские | >2000 в базах | найдено & исправлено | ~350 в статьях | A → B в программах

ошибки, противоречия

исправляемые

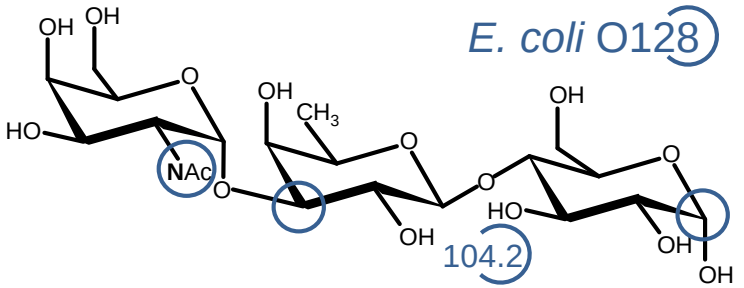
2dGlc → araHex,
α-Rib-ol → Rib-ol,
D-Kdo → Kdo,
1-methyl → 1-Me,
n.m.r. → NMR,
taxid 583 → Proteus,
…

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)bDGlc*p*
Glc C1 103.2 ppm



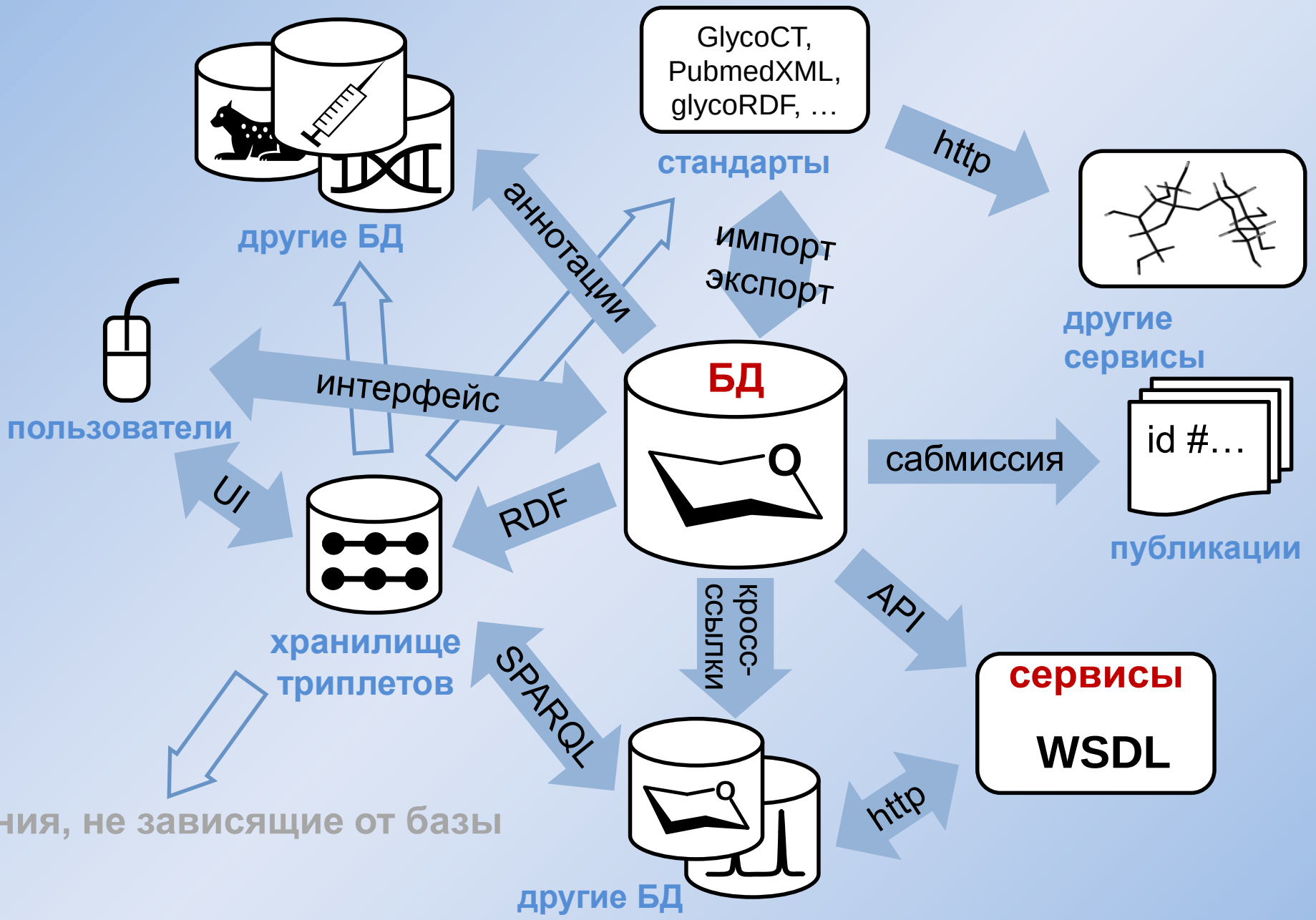
CarbBank ☹ **>50%** (неправильные, отсутствующие, ложно присутствующие структуры, штаммы, аннотации)

→ другие БД

☺ **<10%**

Egorova K, Toukach Ph **Critical analysis of CCSD data quality** *J Chem Inf Model 2012*

Идеальная интеграция



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

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*
Aoki-Kinoshita K, *et al.* **Introducing glycomics data into the Semantic Web** *J Biomed Semantics 2013*

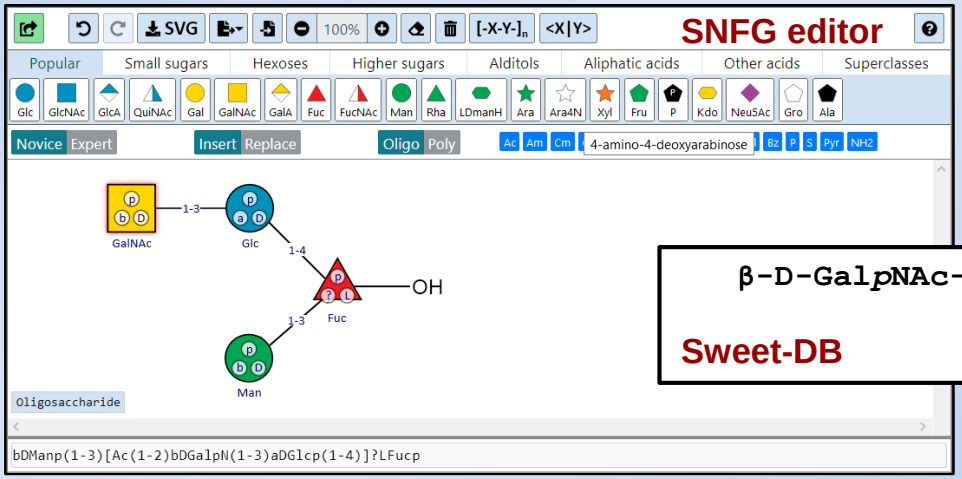
Интерфейс

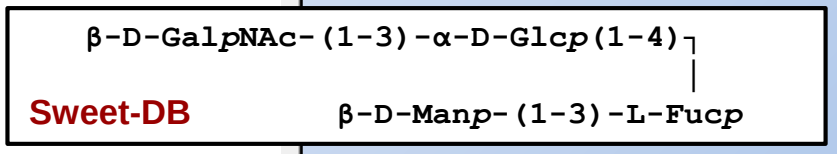
- Конверсия данных ↔ другие форматы
- Автоматические web-сервисы (WSDL)
- Импорт, экспорт
- Документация, HELP
- Дружественность, быстродействие
- Ссылки на записи в других проектах (запросы, индексы, данные)
- Ввод-вывод структур

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

NCBI PubMed, DOI, NCBI Taxonomy, Uniprot, Genbank, Glycosciences.DE, MonosaccharideDB, Glytoucan

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

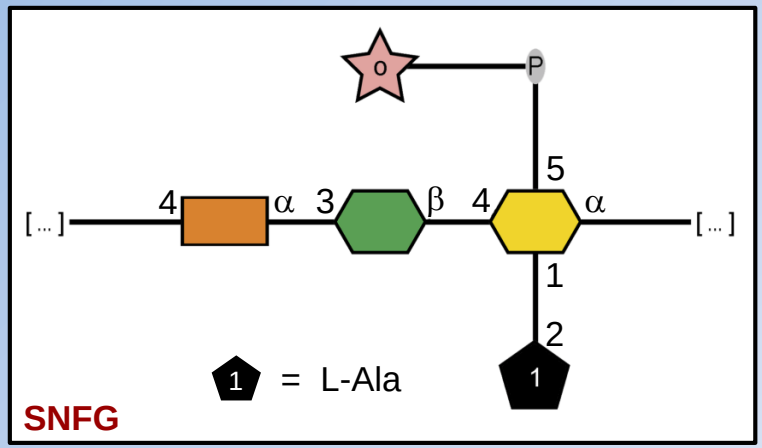




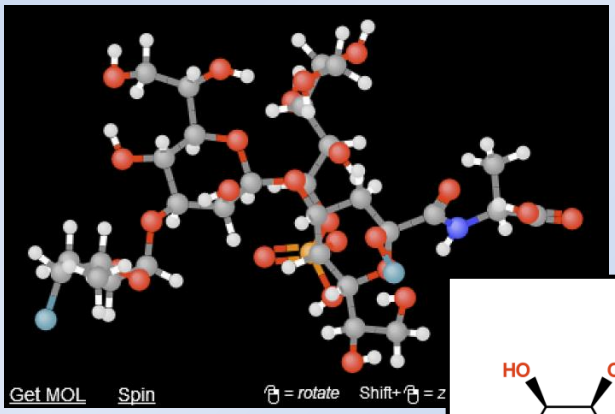
bDManp(1-3)[Ac(1-2)DGalpN(1-?)aDGlcp(1-4)]?LFucp CSDB

Вывод структур

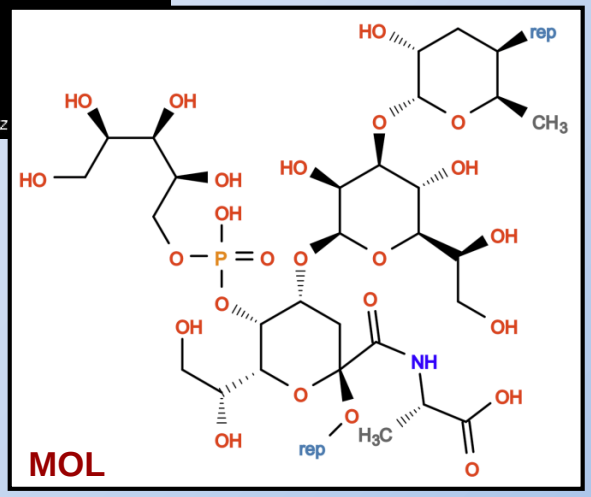
Визуализация в человекочитаемых форматах:



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-
```

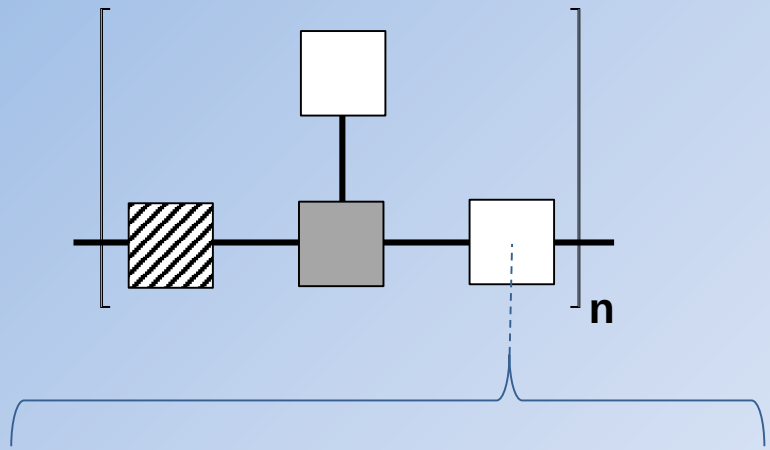
CSDB Linear

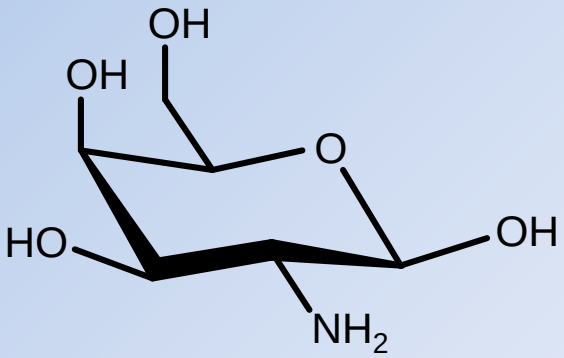
Chernyshov I, Toukach Ph **Automated translation of glycans from residue-based notation to SMILES and atomic coordinates** *Bioinformatics 2017*

Varki A, *et al.* **Symbol nomenclature for graphical representations of glycans** *Glycobiology 2015*

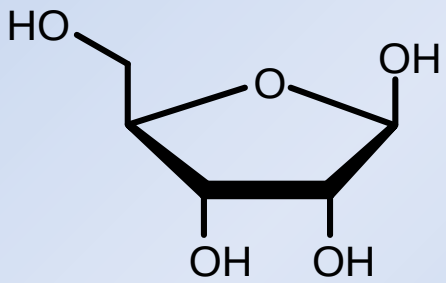
Alocci D, Suchankova P, Costa R, Hory N, Mariethoz J, Toukach Ph, Lisacek F **SugarSketcher: quick intuitive online glycan drawing** *Molecules 2018*

Структура углеводов





пример *альдо*-пиранозы (β-D-Gal*p*N)



пример *альдо*-фуранозы (β-D-Rib*f*)

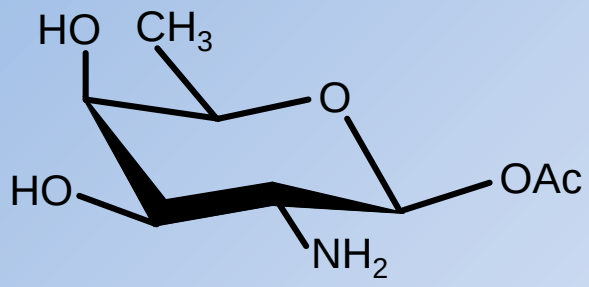
Полная структура

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

Структура остатка

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

Неоднозначность номенклатуры



(CFG)

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)

β-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)

структуры

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

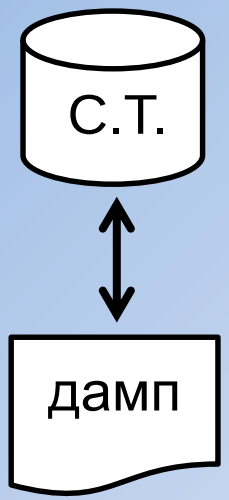
⇩

- 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

Особенности структур и язык

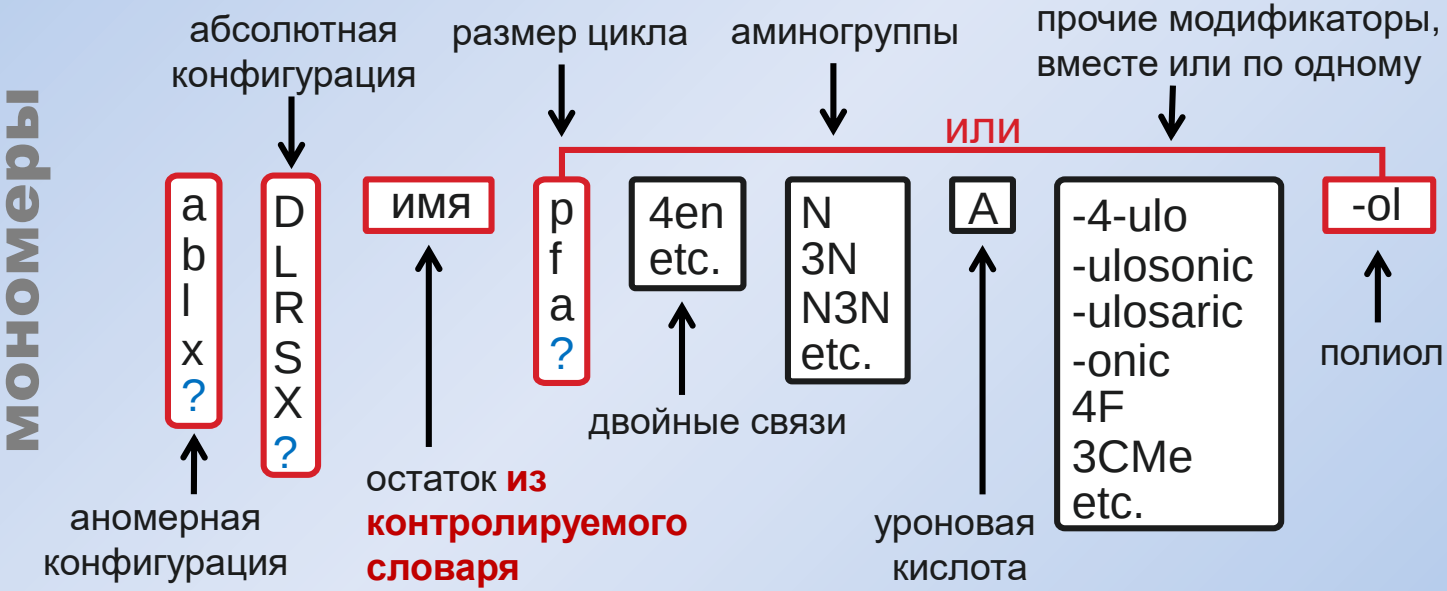


углеводный язык
CSDB Linear

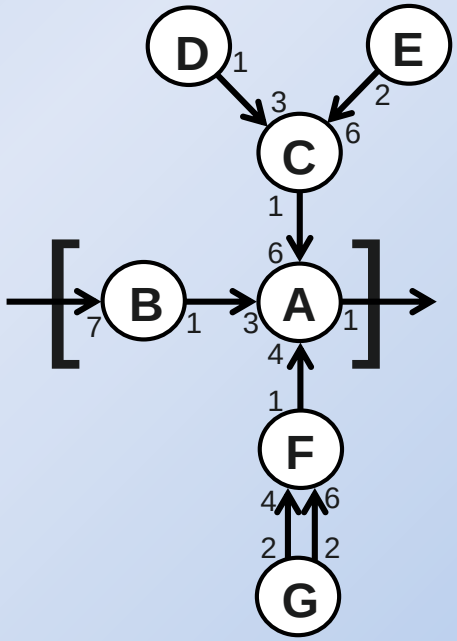
аннотаторы, операторы,
другие языки
(IUPAC, GlycoCT, WURCS, …)

- полнота
- однозначность
- человекочитаемость
- машиночитаемость
- неточные структуры

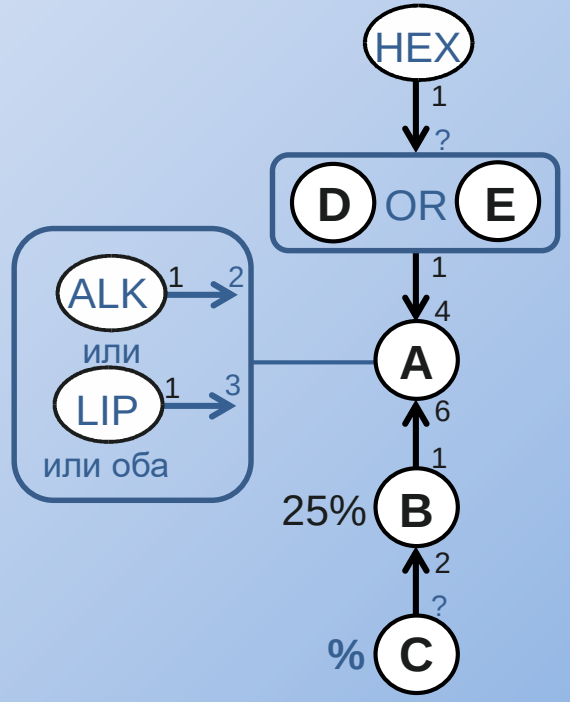
мономеры



топология



неточности



Уровни абстракции

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

точный фрагмент и его связи

xDGro(1-P-5)aDRibf

не указана топология и места присоединения

L-Ala(2-6)+
|
-5)α-D-Ribf(1-3)β-D-GalpNAcA(1-2)D-Gro(1-P-

-?)?Dhex(1-

указаны только класс и 1-связь

приблизительный мотив

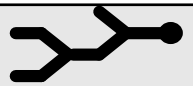
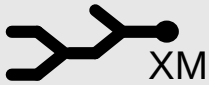
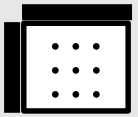
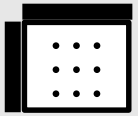
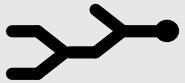
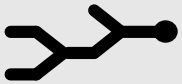
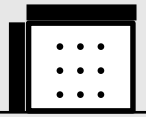
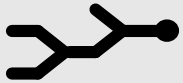
x?Ala(2-?)?DGal?NA

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

HEX, xDRib?, PEP

частичный состав

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

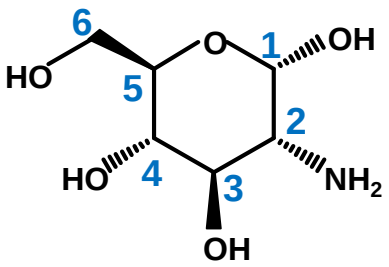
От семантики к атомам

анализ CSDB Linear или другой нотации

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

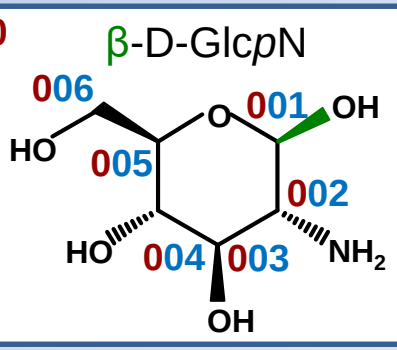
→ SMILES прототипов

0 α-D-Glc*p*N



стереоконфигурации + метки

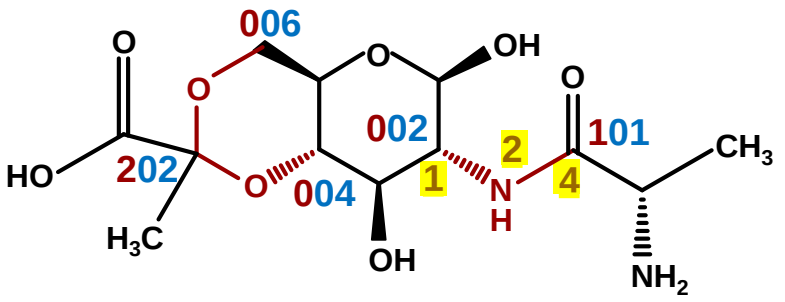
0 β-D-Glc*p*N



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

SMARTS связей между остатками

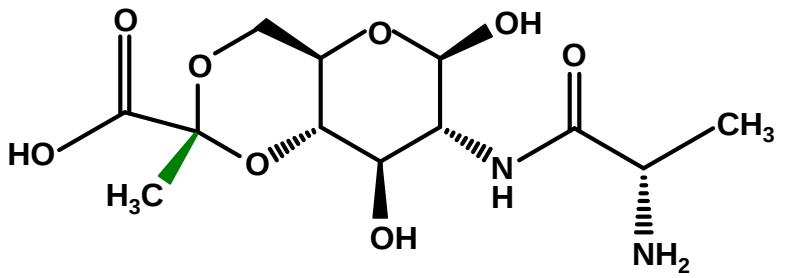
0+1, 1+2 Pyr-(2-4:6)-[L-Ala-(1-2)]-β-D-Glc*p*N



[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

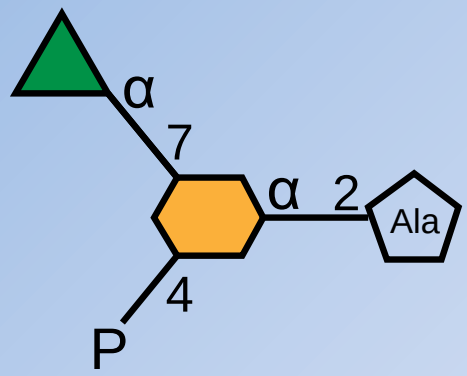
R-Pyr-(2-4:6)-[L-Ala-(1-2)]-β-D-Glc*p*N



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

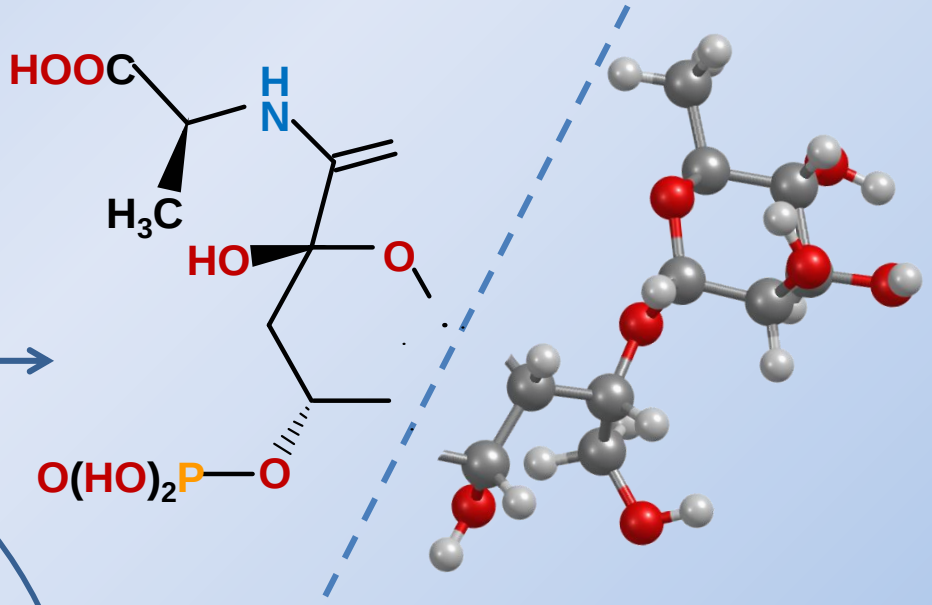
Любые операции на атомарном уровне
+
Внешние химические программы

Молекулярная геометрия

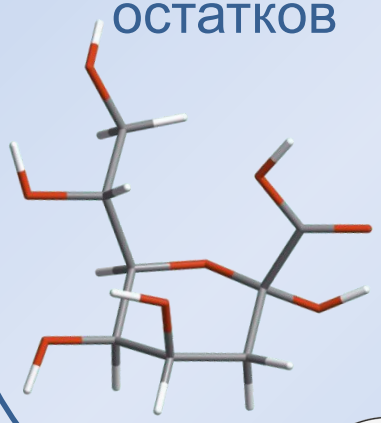


алгоритм
RESтLESS

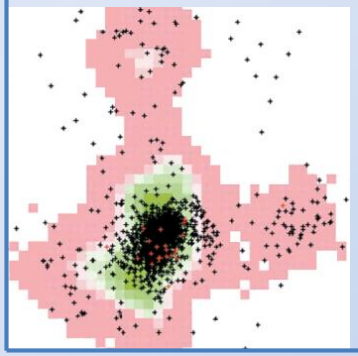
Residues as SMILES,
Linkages as SMARTS



конформеры остатков



φ-ψ карты



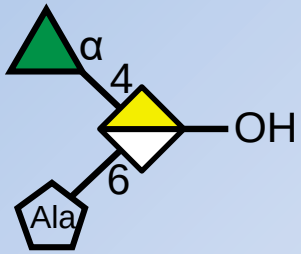
начальные геометрии

исходные данные для докинга, QSAR и др. расчетов структура-активность

Моделирование конформеров

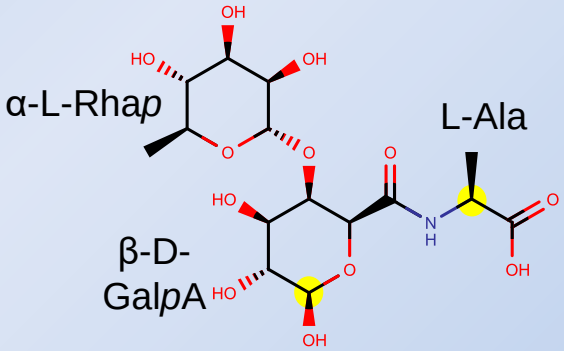
aLRhap(1-4)[x?Ala?(2-6)]?DGalpA

структуры,
в т.ч. неполные

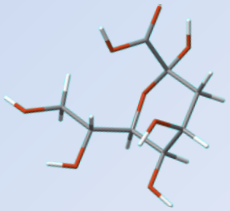


другие варианты
(α-GalA, D-Ala, и т. д.)

SMILES

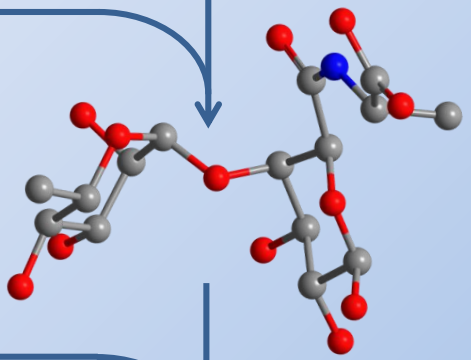


выгодные
конформеры
~1000 остатков

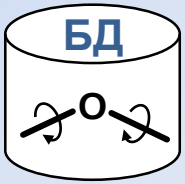


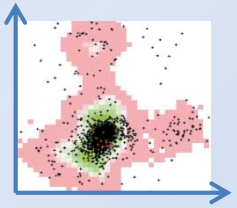
$^4C_1$,$^1C_4$, …

«креслификация»



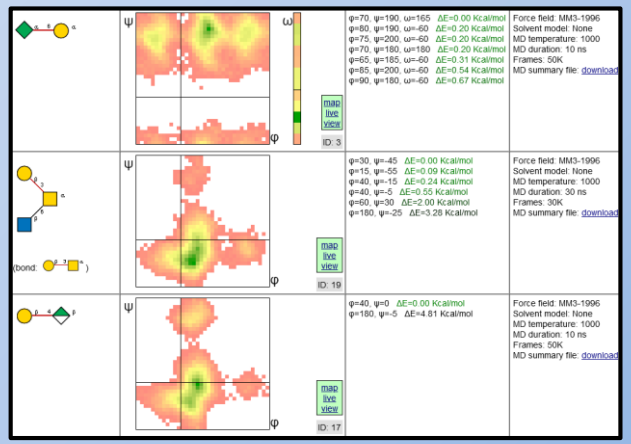
заселенные
состояния
мостиков





φ,ψ,ω,θ

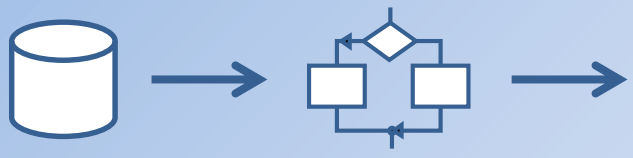
выбор минимумов
ММ-релаксация



мол. динамика
300K, 100нс, H_2O

конформеры
+ их энергии

Надстройки

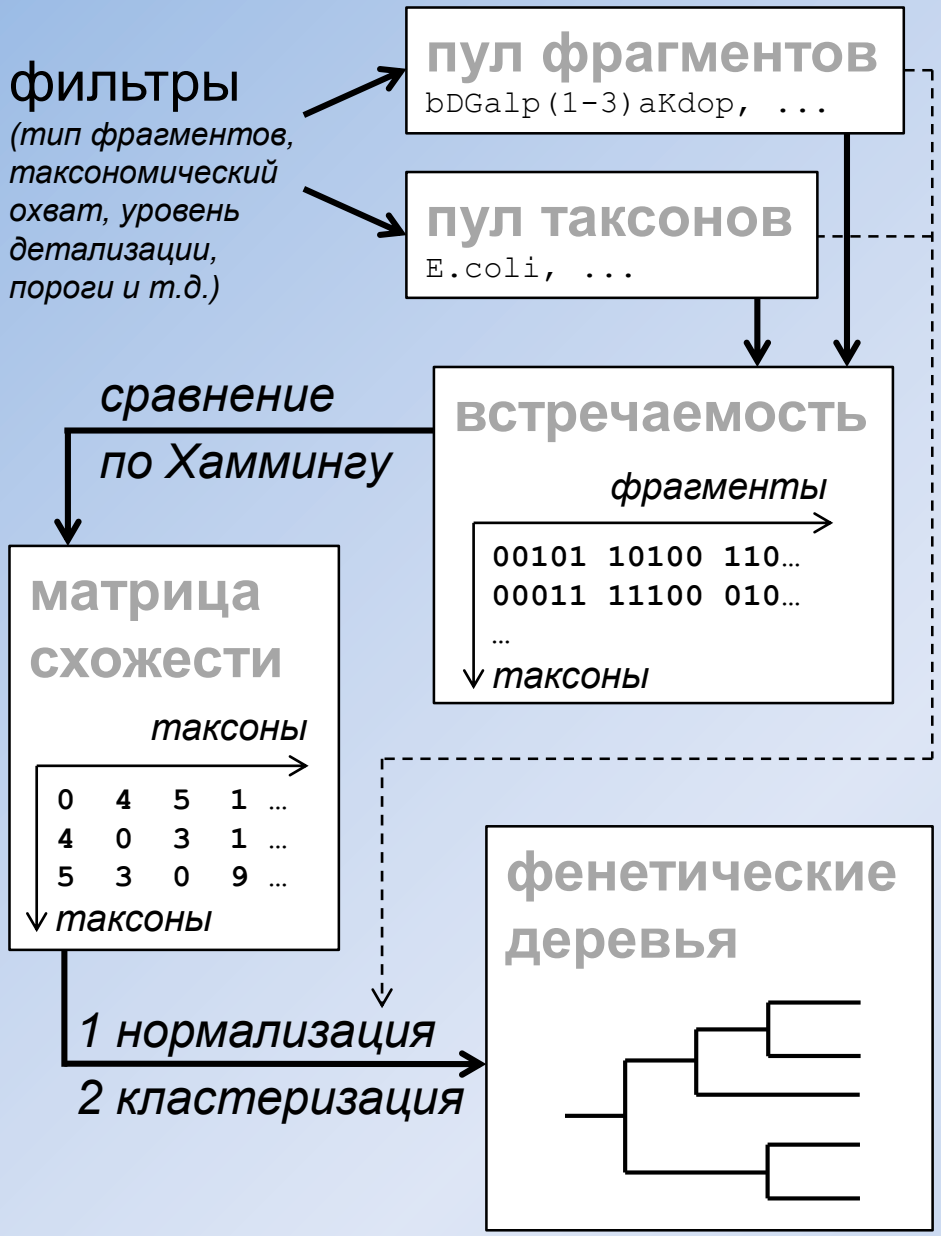


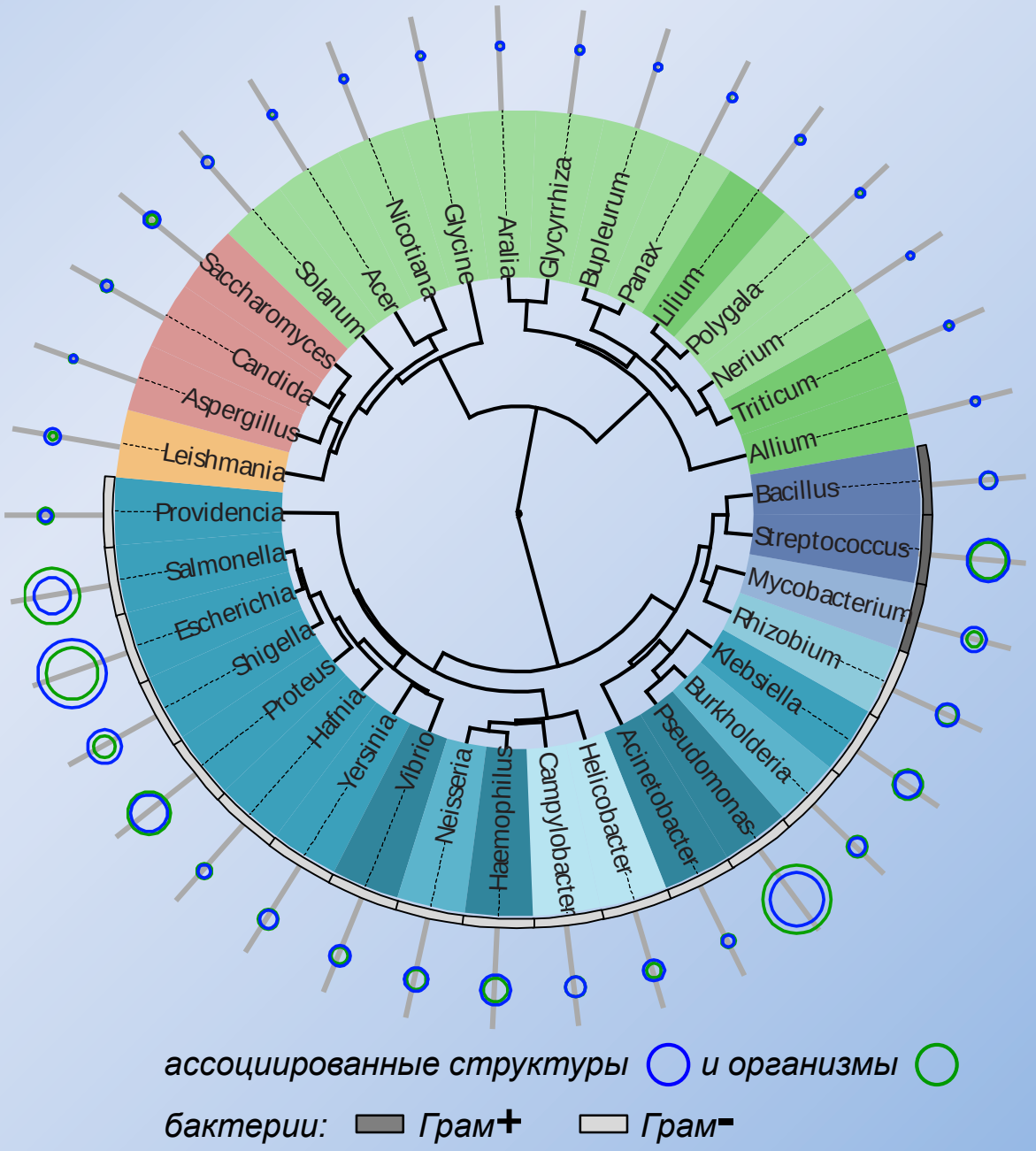
знания надстройки

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

Кластеризация таксонов

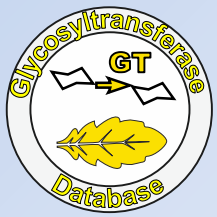




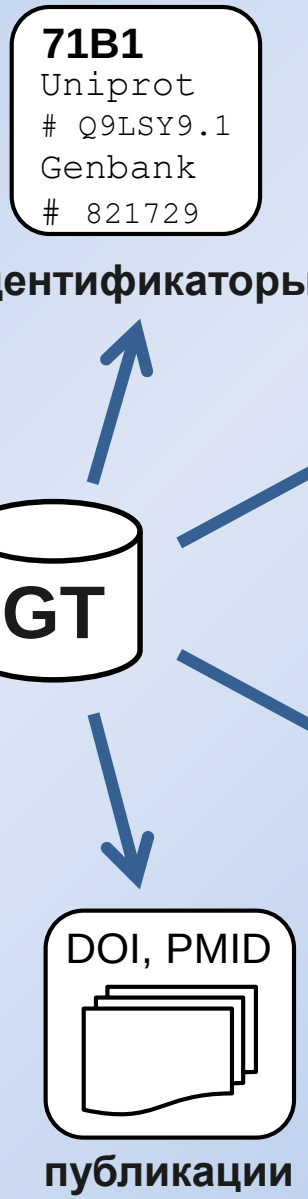
Гликозилтрансферазы

Критерии:
(в любом сочетании)

- идентификаторы в базах
- название фермента / группа
- название гена / кластер
- семейство CAZy
- организм (вид, штамм)
- синтезируемая связь
- донор (или его фрагмент)
- акцептор (или его фрагмент)
- роль объекта в клетке
- уровень достоверности

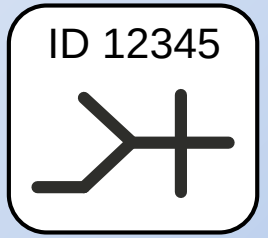


CSDB GT



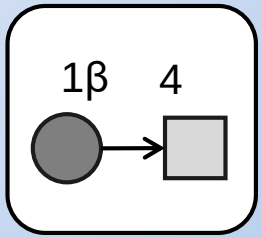
Объект:

организм, орган, ткань

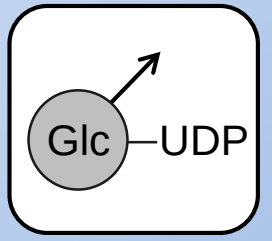


полная структура

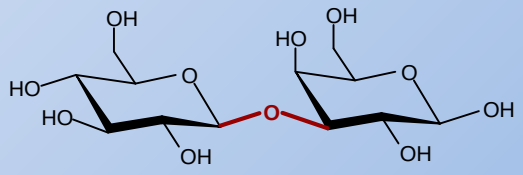
Активность:



синтезируемый фрагмент



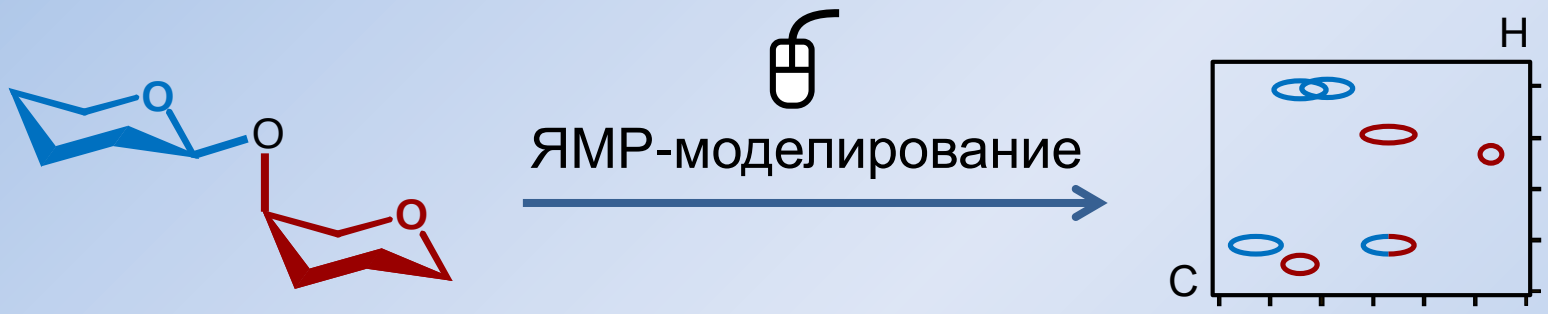
донор и акцептор



Egorova K, Toukach Ph **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017*
Egorova K, Toukach Ph **Expanding CSDB_GT glycosyltransferase database with *Escherichia coli*** *Glycobiology 2019*
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 2020*

ЯМР-моделирование

ЯМР – основной метод анализа первичной структуры в гликобиологии



- Помощь эксперту в установлении структуры
- Отнесение сигналов и проверка гипотез
- Перебор структур и сравнение симуляции с экспериментом
- Верификация расчетной геометрии молекул

алгоритм **GODDESS** Glycan-Optimized Database-Driven Empirical Spectrum Simulation

Как предсказать спектр ЯМР?

Статистически

химические сдвиги ^{13}C и ^{1}H

☺ Существует база (CSDB)
☺ Отслеживается до источников
☹ Медленно (~минуты)

адаптация для углеводов:
GODDESS

Эмпирически

химические сдвиги и КССВ

☺ Очень быстро (~миллисекунды)
☹ Требуется модель
☹ Нужны специальные базы

адаптация для углеводов:
GODDESS, BIOPSEL, CASPER

Квантово-механически

геометрия + все параметры ЯМР

☺ Не зависит от базы
☹ Низкая точность (>3 м.д.)
☹ Очень медленно (~месяцы)

адаптация для углеводов: не нужна

Иначе

(нейронная сеть, регрессия, ММ-QM)

☹ Интегрально малопригодно

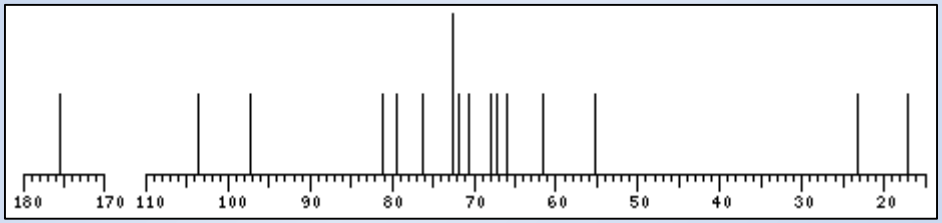
адаптация для углеводов: нет

Статистическое моделирование

- Использует регулярно обновляемую базу CSDB (>9000 спектров)
- Обобщает химическое окружение предсказываемого атома,

  пока не будет найдено достаточно похожих фрагментов в базе
- Работает на уровне групп атомов и специальных дескрипторов
- Оценивает достоверность и позволяет отследить источники
- Смешивает эмпирические и статистические результаты

применимо к предсказанию любых атомарных свойств, представленных в базе



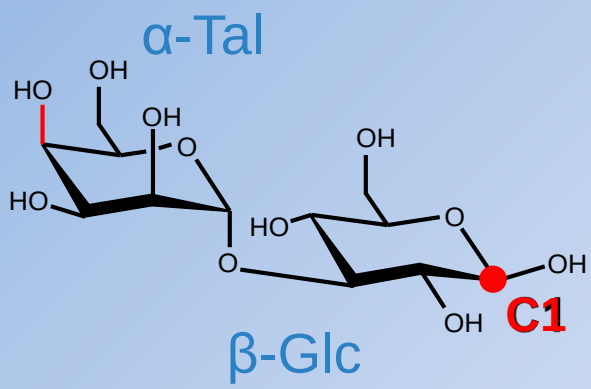
β-D-GlcpNAc-(1-3)-α-D-Fucp-(1-P-3)-D-Gro

^{13}C NMR data: Solvent: Water (H or D) Recalc

Linkage	Residue	C1	C2	C3	C4	C5	C6	Accuracy
0	xDGro	67.3	72.6	66.0				1.17
	trustworthiness->	0	0.58	2.93				
	NMR references->	19	17	19				
3	xXP?	-						
3,0	aDFucp	97.2	72.5	79.5	70.6	68.0	17.2	2.98
	trustworthiness->	2.83	2.62	3.8	1.61	3.39	3.63	
	NMR references->	3	3	3	18	13	4	
3,0,3	bDGlcpN	103.7	55.1	81.2	71.9	76.2	61.5	3
	trustworthiness->	3	3	3	3	3	3	
	NMR references->	1	1	1	1	1	1	
3,0,3,2	Ac	175.4	23.3					3.68
	trustworthiness->	3.65	3.71					
	NMR references->	19	21					

Обобщение структуры

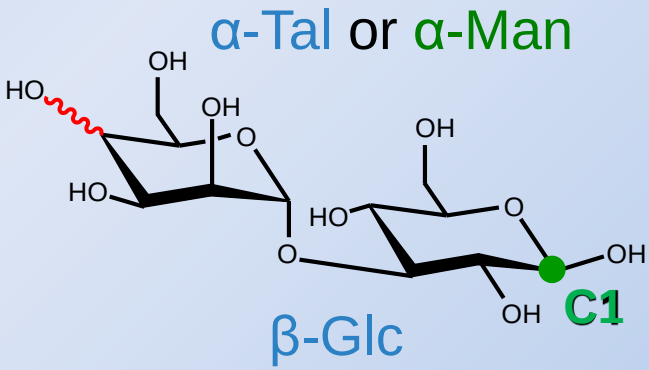
фрагмент структуры



пример обобщения

конфигурация C4 донора

обобщенные фрагменты

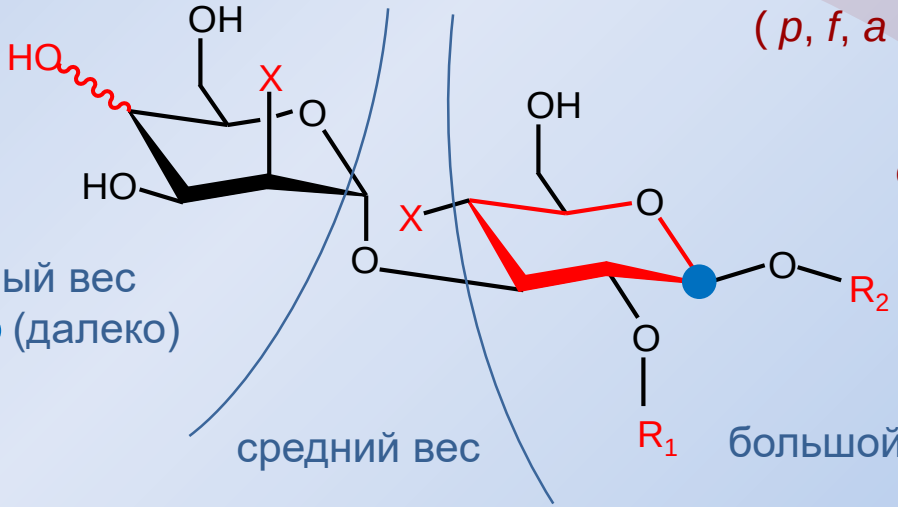


обобщаемые дескрипторы

типы заместителей
(X = OH, NH_2, H, …)

R/S всех центров

циклизация
(*p*, *f*, *a*)



соседние фрагменты (R_n) и их положения

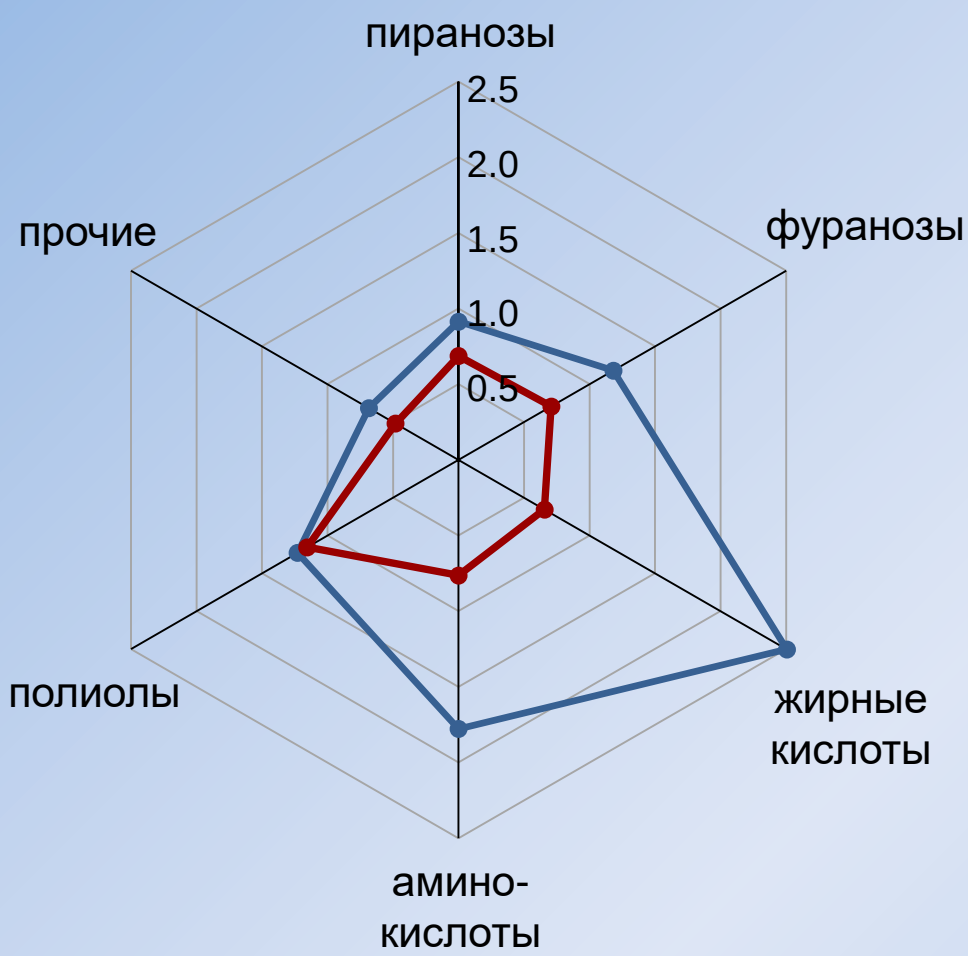
малый вес для ● (далеко)

средний вес

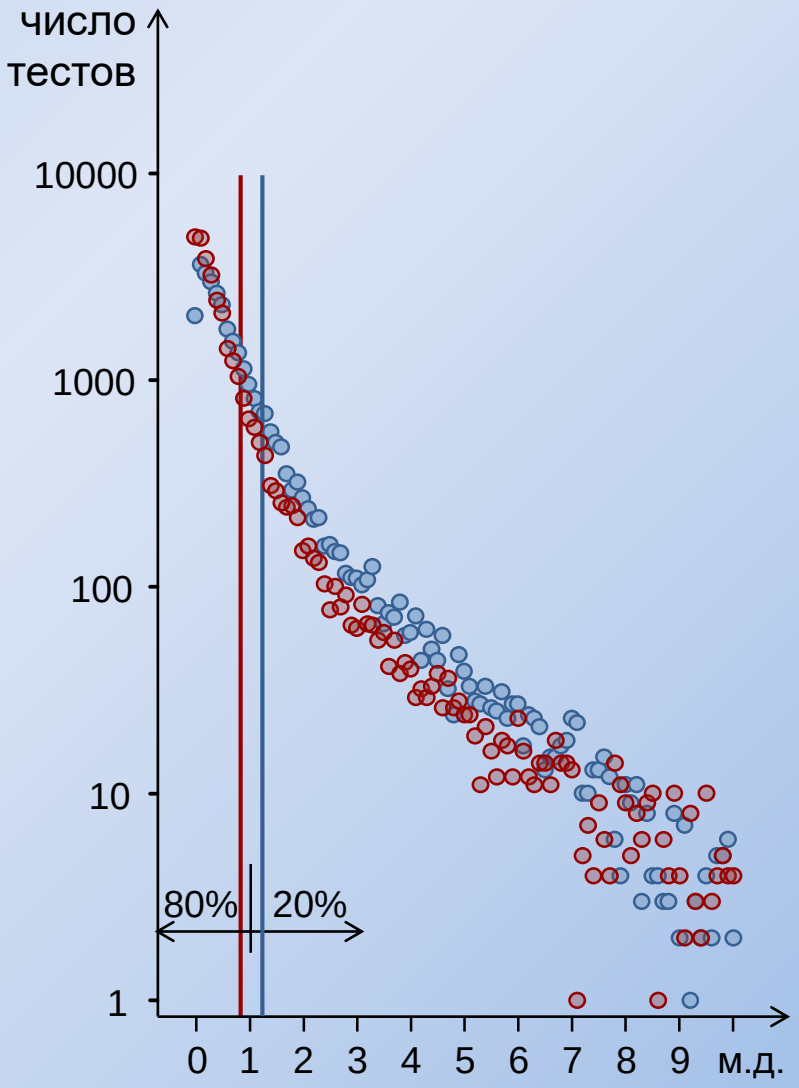
большой вес

каждый олигомерный фрагмент обобщается от малых весов к большим,
пока в базе не найдется достаточно обобщенных фрагментов,
затем данные из базы усредняются с учетом выбросов

Точность



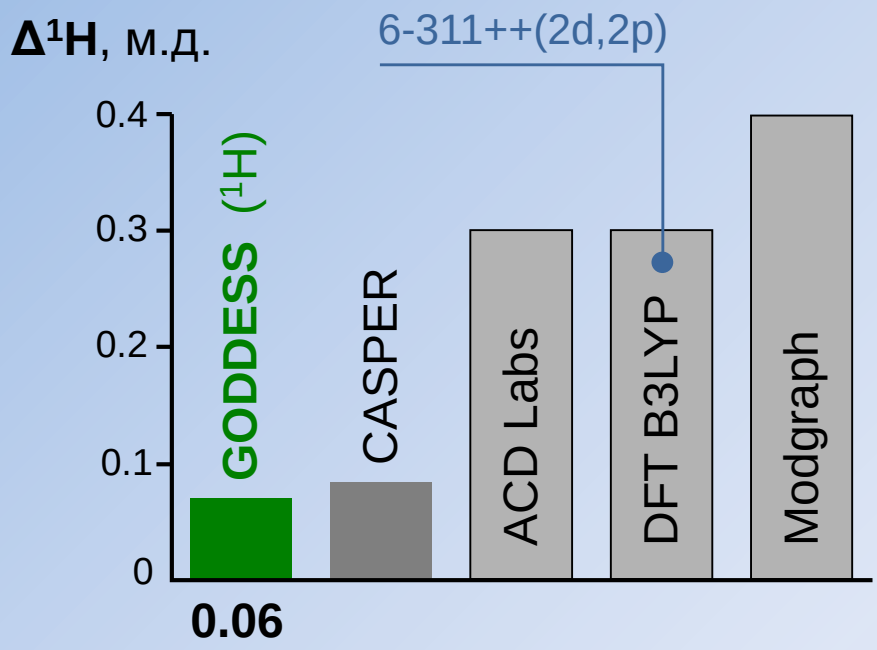
средняя точность по классам, м.д.

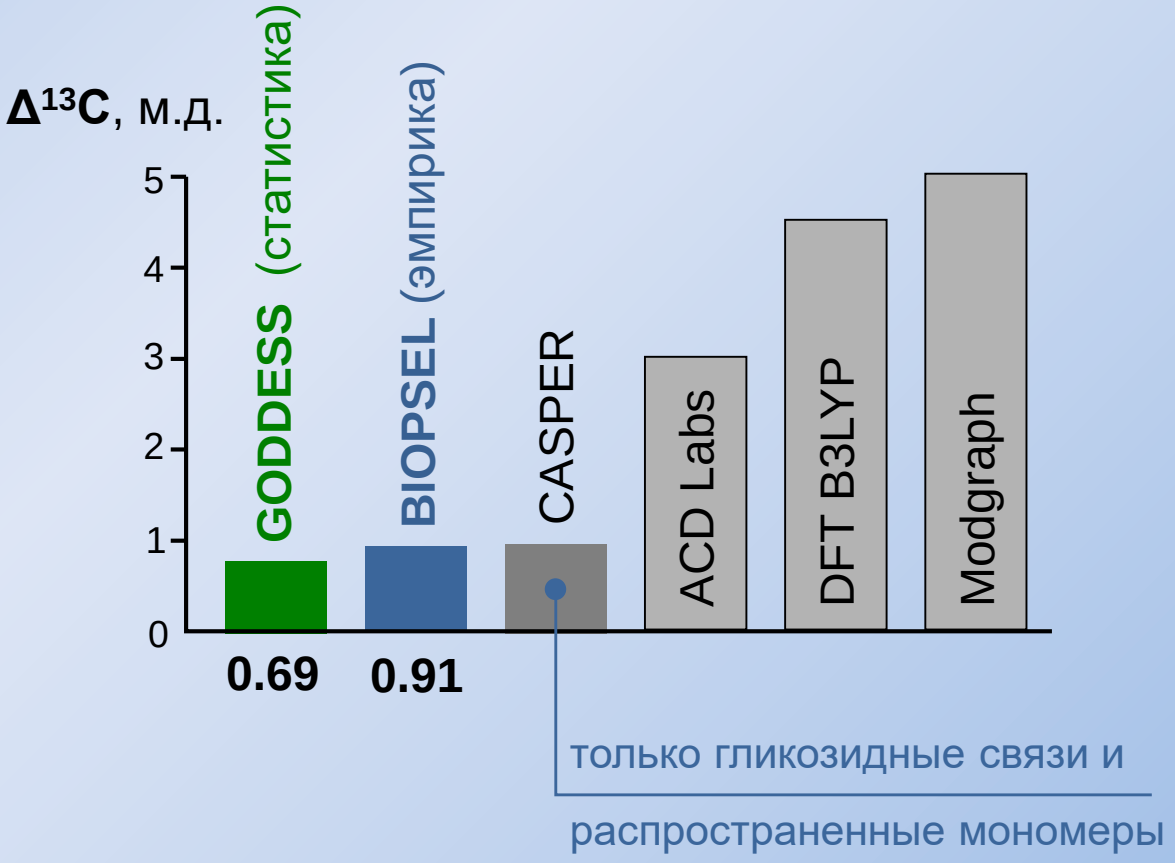


ошибка моделирования
химического сдвига

^{13}C - эмпирика
^{13}C - статистика

Сравнение подходов

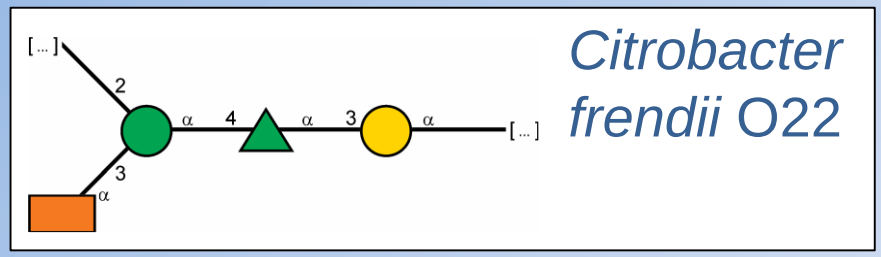




сравнение «предсказание-эксперимент» проведено на
~32000 химических сдвигов из всех классов биогликанов

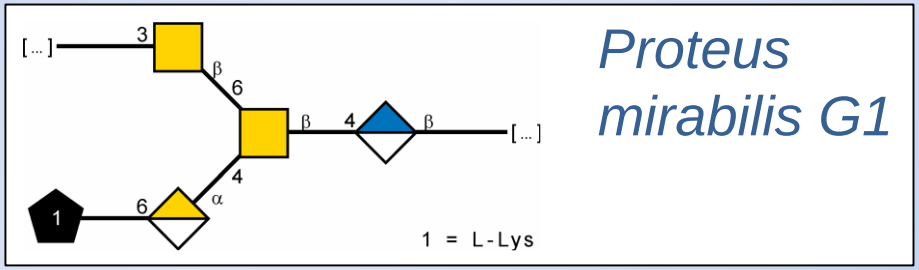
Kapaev R, Toukach Ph **Improved carbohydrate structure generalization scheme for ¹H and ¹³C NMR simulations** *Analyt Chem 2015*
Toukach F, Ananikov V **Computational predictions of NMR parameters for structure elucidation of carbohydrates** *Chem Soc Rev 2013*

Модель и эксперимент



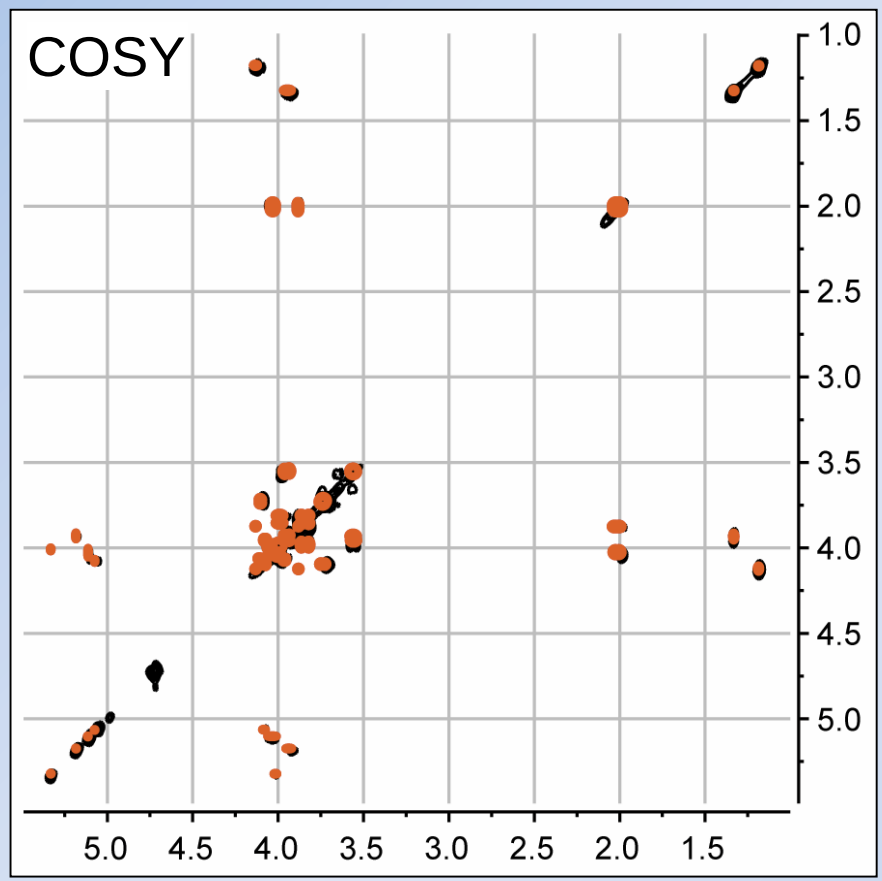
Citrobacter frendii O22

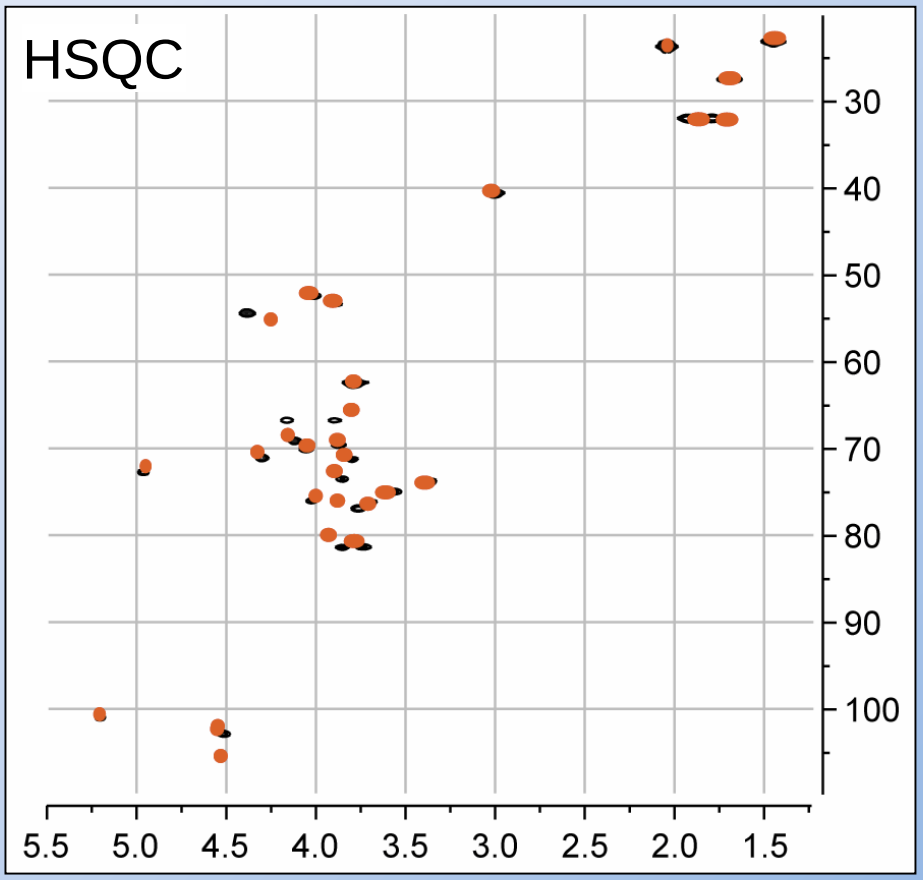
α–Abe-(1┐
↓3
→2)-α-D-Man*p*-(1→4)- α-L-Rha*p*-(1→3)- α-D-Gal*p*-(1→



Proteus mirabilis G1

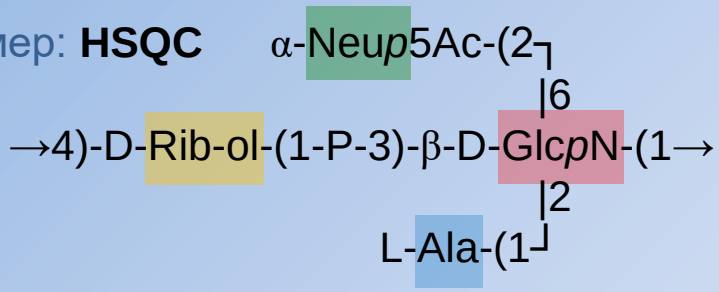
L-Lys-(2-6)-α-D-Gal*p*A-(1┐
↓4
→3)-β-D-Gal*p*NAc-(1→6)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→





Визуализация результатов

Пример: **HSQC**



Поддерживается:

1D: ^{13}C BB, таблицы отнесения ^{1}H и ^{13}C

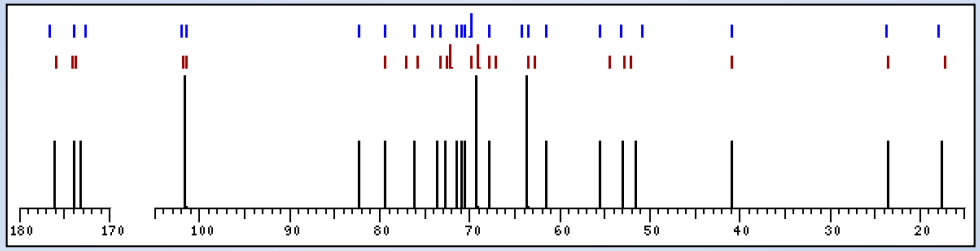
^{1}H-^{1}H: COSY, TOCSY, DQF COSY, COSY RCT

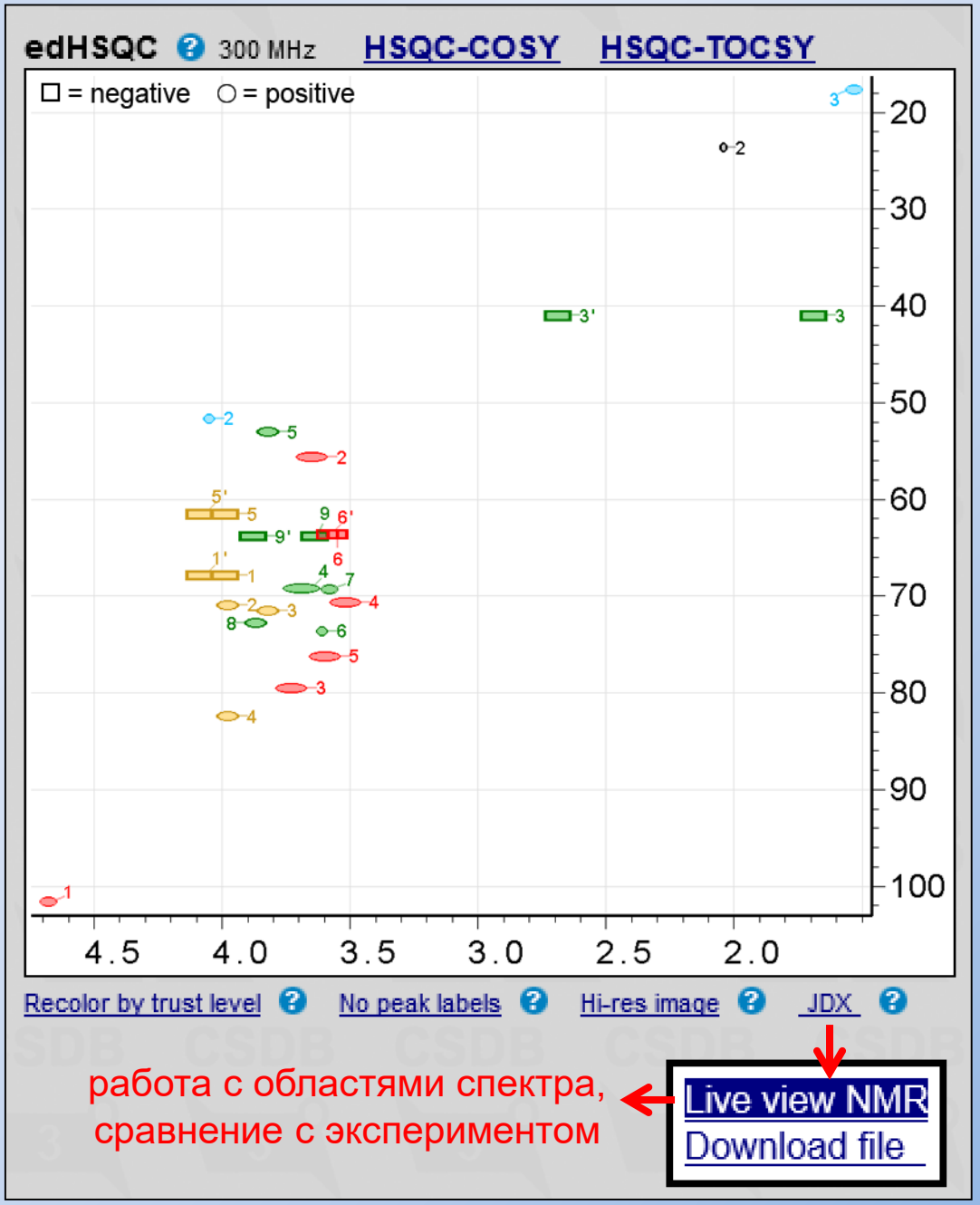
^{1}H-^{13}C: edHSQC, HMBC,
HSQC-COSY, HSQC-TOCSY

эмпирическая оценка КССВ
+ частота спектрометра → ширина кросс-пиков

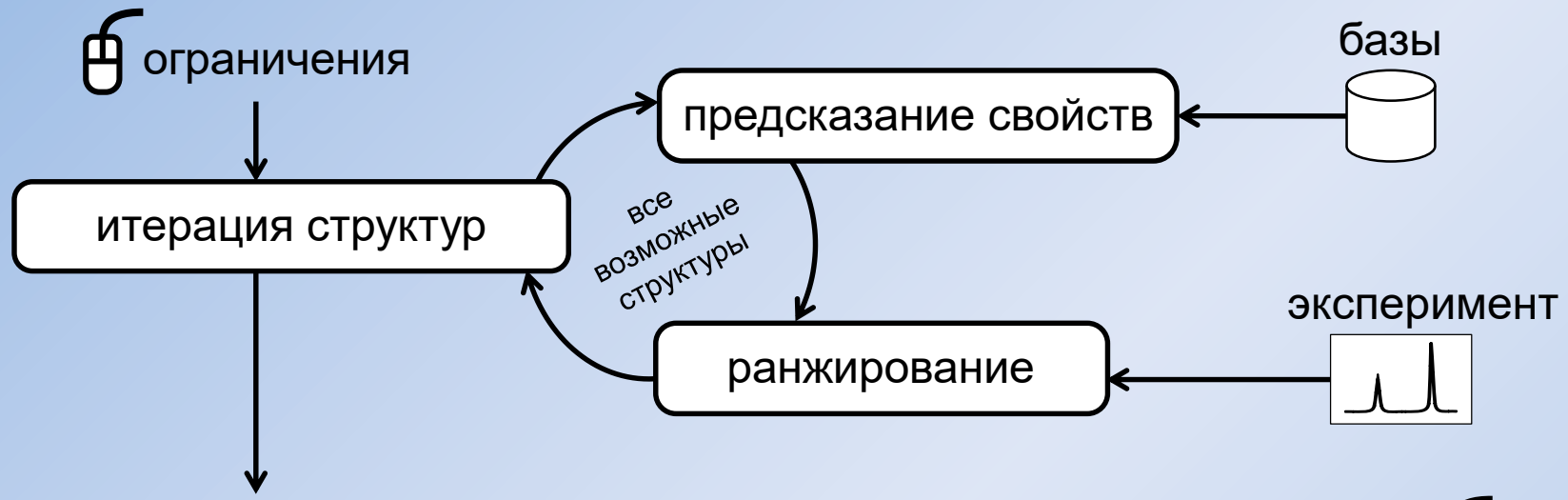
Скоро: NOESY / ROESY

На выходе: таблицы, изображения, «живая» работа в браузере, экспорт в CSV и Jcamp-DX





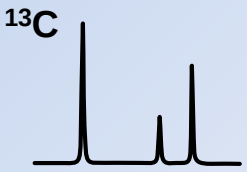
Итератор структур



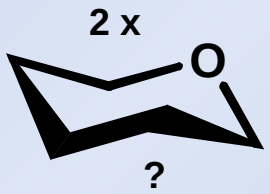
структуры и данные с наилучшим соответствием
данные прослеживаются до исходных баз и литературы

фильтры отбора

неотнесенный спектр ЯМР

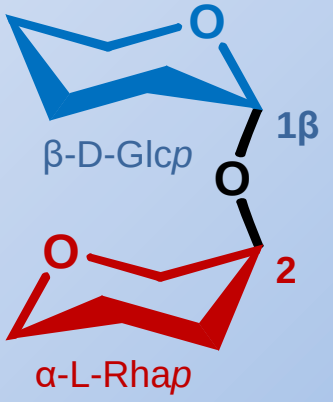


известные данные о структуре



алгоритм **GRASS**

Generation, Ranking and Assignment of Saccharide Structures



полная структура (оставшиеся неизвестные)

мономеры, конфигурации, модификации, позиции замещения, последовательность

Выбор структурных гипотез

мономерный состав,
частичная последовательность
(гидролиз и хроматография)

позиции замещения
(метилирование)

число β-аномеров, число CH_2-групп, N-ацетилирование (*разрешено, запрещено, возможно*), число фосфатов, присутствие фураноз и т.д.
(из 1D ЯМР)

ВВОД

абсолютные конфигурации
(гидролиз и модификация)

структурные ограничения
(**опционально**)

неотнесенный спектр ЯМР ^{13}C
(допустимы пропущенные / лишние сигналы)

число остатков (1..9)
в олигомере или
повторяющемся звене

для ~500 наилучших кандидатов:
точный статистический расчет
+
сравнение и ранжировка

перебор всех химически
возможных структур
+
быстрый эмпирический расчет

ОБРАБОТКА

рейтинг структурных гипотез
+
метрика соответствия

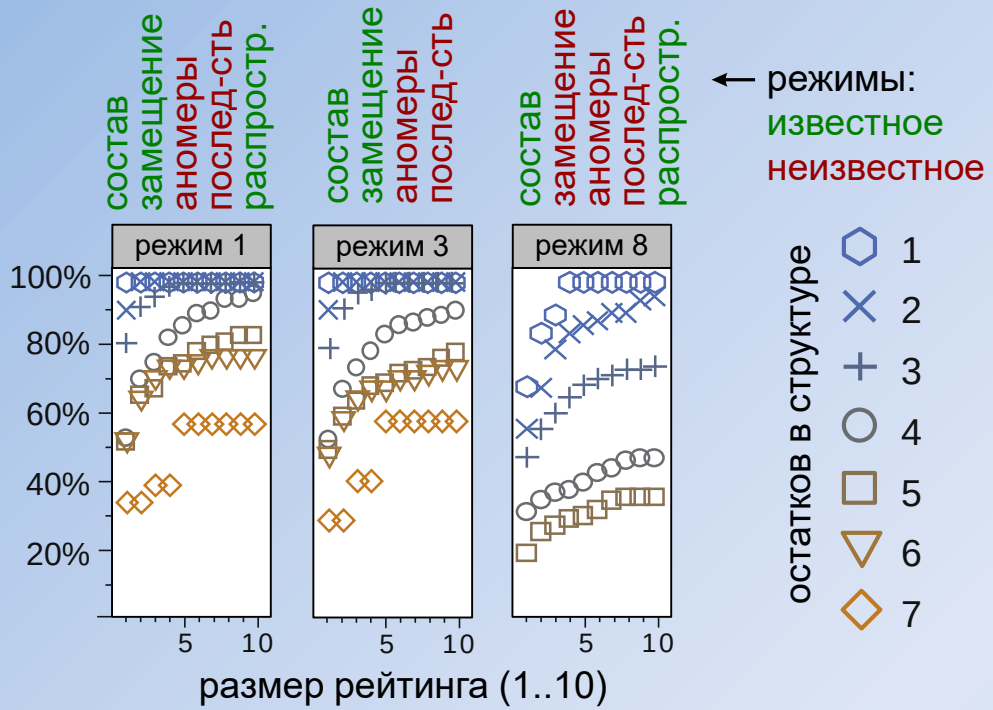
- первичная структура
- расчетные спектры ЯМР ^{1}H, ^{13}C, 2D
- экспорт и сравнение спектров
- атомные координаты

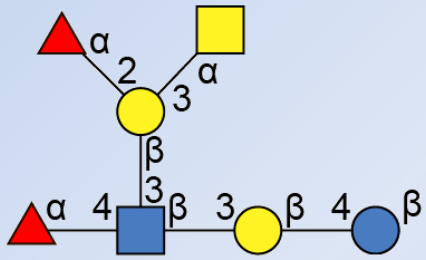
ВЫВОД

алгоритм **GRASS** Generation, Ranking and Assignment of Saccharide Structures

Результаты ранжирования

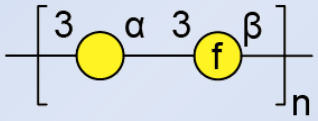
Статистика по 556 структурам





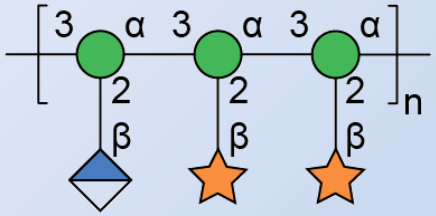
Антиген группы крови A

режим 1
→ **TOP-5** (→TOP-1 для пяти аномеров)



D-галактан I

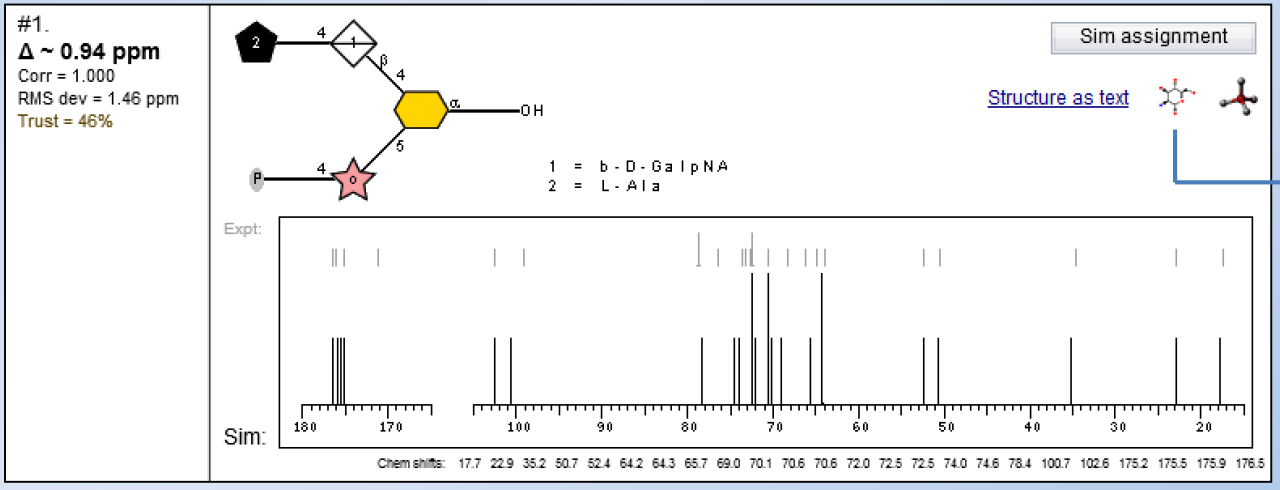
режим 8 *(ничего не известно)*
→ **TOP-1** (лучшая гипотеза)

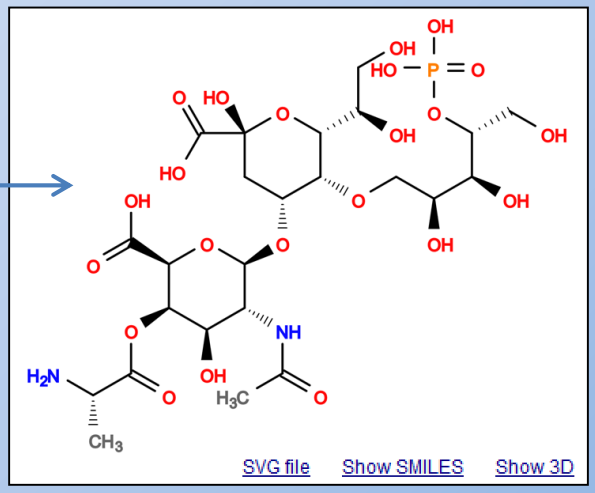


Ксиломаннан *C. neoformance* A

режим 3
→ **TOP-1** (лучшая гипотеза)

Метрики + вывод





Что сделано?

- **Цель:** оптимизация и автоматизация структурно-функциональных исследований углеводов; информатизация гликохимии и гликобиологии.
- **Средства:** разработка и апробация идеологии, моделей, стандартов, инструментов работы со структурами, спектрами, таксономией.

Что сделано?

- **Цель:** оптимизация и автоматизация структурно-функциональных исследований углеводов; информатизация гликохимии и гликобиологии.
- **Средства:** разработка и апробация идеологии, моделей, стандартов, инструментов работы со структурами, спектрами, таксономией.
- **Разработано и объединено в систему:**

правила обработки информации в гликохимии; — **подходы**

онтология в гликомике;

база данных и платформа CSDB; — **БД + софт**

углеводный язык и его связь с атомарными моделями; — **структуры + 3D**

быстрое получение молекулярной геометрии углеводов;

ЯМР-моделирование углеводов; — **спектры**

сравнение и предсказание структур по спектрам;

кластерный анализ химического разнообразия гликомов; — **таксоны**

интеграция с конкурентами; — **взаимодействие**

web-портал.

Что сделано?

- **Цель:** оптимизация и автоматизация структурно-функциональных исследований углеводов; информатизация гликохимии и гликобиологии.
- **Средства:** разработка и апробация идеологии, моделей, стандартов, инструментов работы со структурами, спектрами, таксономией.

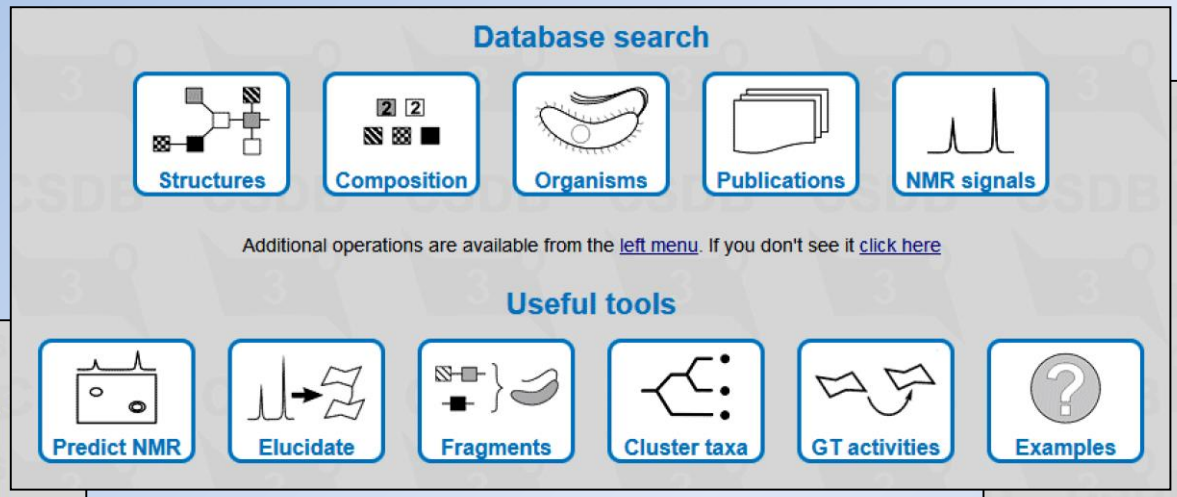
- **Итог:**

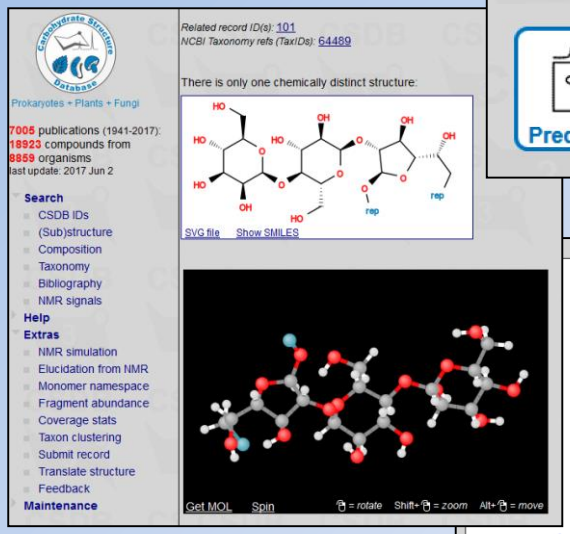
Новые гликохимические инструменты созданы, верифицированы на модельных системах и использованы для реальных исследований.

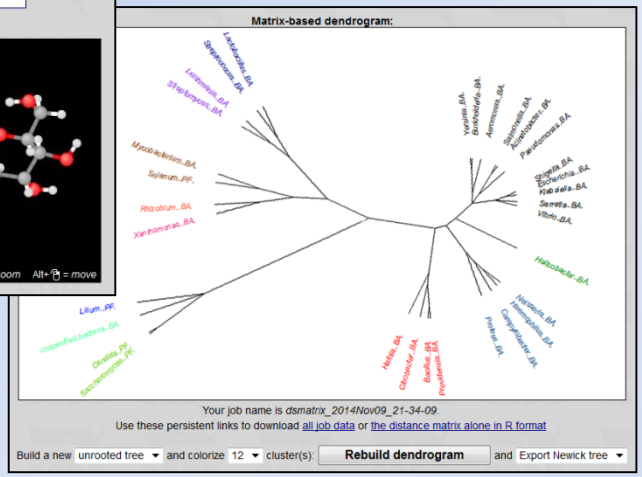
Заложен фундамент для статистических и прямых расчетов корреляции структура-свойство *(востребовано в направленном синтезе)*.

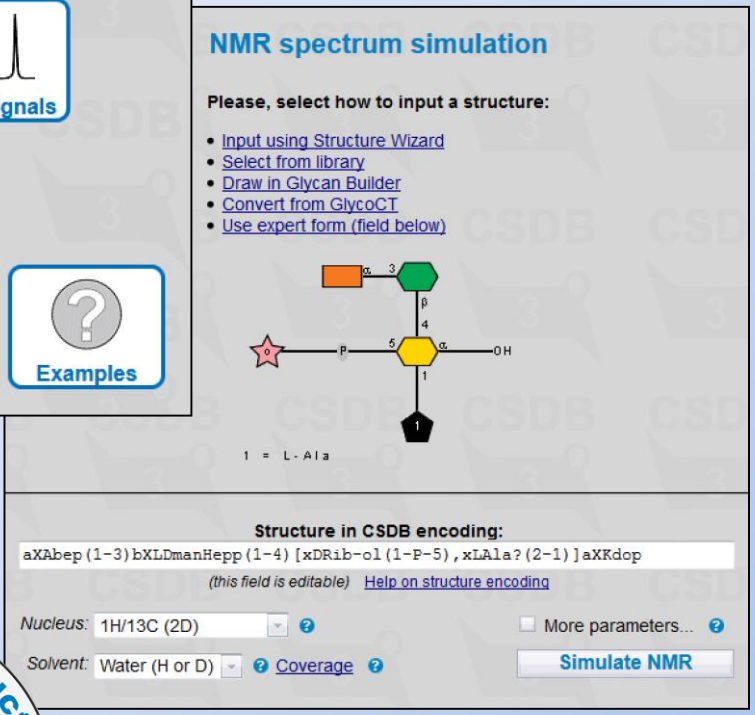
Преобразилась молодая область знания – гликоинформатика, задан и обеспечен мировой вектор ее развития.

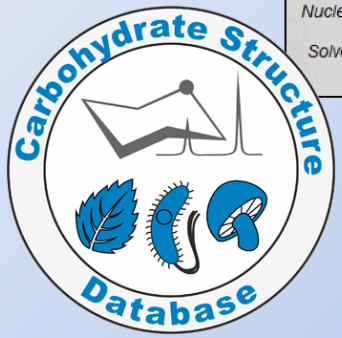
Сайт в Интернете











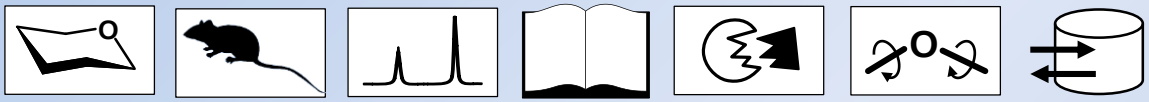
http://csdb.glycoscience.ru

- свободный доступ
- подробная документация
- примеры решения задач

Участники

Carbohydrate Structure Database

проверяемый контент, полное покрытие



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
Книрель (1 грант) Тоукач (3 гранта)	Книрель (1 грант)	Тоукач (1 грант)	Тоукач (4 гранта)	Тоукач (1 грант)	Тоукач (2 гранта)

Публикации

- По теме доклада: 34 статьи + 3 монографии (суммарный импакт-фактор: ~170)
- Использование наработок в исследованиях других авторов: ~600 цитирований
- Каждый год (кроме роботов): ~2500 поисковых запросов, ~1000 запросов на расчет
- Избранные публикации:

Toukach PhV, Joshi H, Ranzinger R, Knirel YuA, von der Lieth C-W **Sharing of worldwide distributed carbohydrate resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de** *Nucl Acid Res 2007, 35:D280-D286*

Toukach PhV **Bacterial Carbohydrate Structure Database 3: Principles and Realization** *J Chem Inf Model 2011, 51:159-170*

Egorova KS, Toukach PhV **Critical analysis of CCSD data quality** *J Chem Inf Model 2012, 52:2812-2814*

Toukach FV, Ananikov VP **Recent advances in computational predictions of NMR parameters for structure elucidation of carbohydrates: methods and limitations** *Chem Soc Rev 2013, 42: 8376-8415*

Kapaev RR, Egorova KS, Toukach PhV **Carbohydrate structure generalization scheme for database-driven simulation of experimental observables, such as NMR chemical shifts** *J Chem Inf Model 2014, 54:2594-2611*

Egorova KS, Kondakova AN, Toukach PhV **CSDB: tools for statistical analysis of bacterial, plant and fungal glycomes** *Database 2015, ID bav073*

Toukach PhV, Egorova KS **Bacterial, Plant, and Fungal CSDB: daily Usage**. в **Glycoinformatics**, ред. Lütteke T, Frank M. Springer, 2015, гл. 5: 55-85

Kapaev RR, Toukach PhV **Improved carbohydrate structure generalization scheme for ^{1}H and ^{13}C NMR simulations** *Anal Chem* 2015, 87:7006-7010

Ranzinger R *et al.* (13 авторов вкл. Toukach PhV) **GlycoRDF: an ontology to standardize Glycomics data in RDF** *Bioinformatics 2015, 31:919-925*

Toukach PhV, Egorova KS **Carbohydrate Structure Database merged from multiple taxonomic domains** *Nucl Acid Res 2016, 44:D1229-D1236*

Kapaev RR, Toukach PhV **Simulation of 2D NMR spectra of carbohydrates using GODESS Software** *J Chem Inf Model 2016, 56:1100-1104*

Egorova KS, Toukach PhV **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017, 27:285-290*

Kapaev RR, Toukach PhV **GRASS: semi-automated NMR-based structure elucidation of saccharides** *Bioinformatics 2018, 34:957-963*

Chernyshov IYu, Toukach PhV **REStLESS: automated translation of glycan sequences from residue-based notation to SMILES and atomic coordinates** *Bioinformatics 2018, 34:2679-2681*

Egorova KS, Toukach PhV **Glycoinformatics: bridging isolated islands in the sea of data** *Angewandte Chemie 2018, 57:14986-14990*

Toukach PhV, Egorova KS **New features of CSDB Linear, as compared to other carbohydrate notations** *J Chem Inf Model 2020, 60:1276-1289*

Stroylov VS, Panova MP, Toukach PhV **Comparison of methods for bulk simulation of glycosidic bond conformations** *Int J Mol Sci 2020, 21: ID 7626*

Scherbinina SI, Toukach PhV **Three-dimensional structures of carbohydrates and where to find them** *Int J Mol Sci 2020, 21: ID 7702*

Дополнения

(демонстрируются в рамках ответов на вопросы)

Сокращения на слайдах

BIOPSEL	BIOPolymer Structure ELucidation
CASPER	Computer Assisted SPectrum Evaluation of Regular polysaccharides
CFG	Consortium for Functional Glycomics
COSY	Correlation SpectroscopY
CSDB	Carbohydrate Structure DataBase
CSV	Comma-Separated Values
DFT B3LYP	Density Functional Theory: Becke 3-parameter Lee-Yang-Parr
DOI	Digital object Identifier
GODDESS	Glyco-Optimized Database-Driven Empirical Spectrum Simulation
GRASS	Generation, Ranking and Assignment of Saccharide Structures
GT	GlycosylTransferase
HMBC	Heteronuclear Multiple Bond Correlation
HOSE	Hierarchial Organization of Spherical Environment
HSQC	Heteronuclear Single Quantum Coherence
ICD	International Classification of Diseases
IUPAC	International Union of Pure and Applied Chemistry
MESH	Medical Subject Headings
MSDB	MonoSaccharide DataBase
NCBI	National Center for Biotechnology Information
OWL	Ontology Web Language
PDB	Protein Data Bank
PMID	PubMed IDentifier
RDF	Resource Description Framework
RESтLESS	ResiduEs as SMILES, LinkagEs as SMARTS
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular-Input Line-Entry System
SNFG	Symbolic Notation For Glycans
SPARQL	SPARQL Protocol and RDF Query Language
TOCSY	TOtal Correlation SpectroscopY
WSDL	Web Service Description Language
WURCS	Web-3 Unique Representation of Carbohydrate Structures

- стандартные
- новые

Выводы - кратко

- Аннотации ~10 тыс. публикаций (~20 тыс. структур). Свободно доступная база данных Carbohydrate Structure Database (CSDB), оснащённая многочисленными видами поиска, представления и анализа данных.
- Инструмент автоматического моделирования структуры биогликанов - фундамент для расчётов корреляции «структура – свойство» в химии углеводов.
- Аннотации ~9 тыс. спектров ЯМР биогликанов. Исследование взаимосвязи «структура – спектр». Метод обобщения атомного окружения в углеводных структурах и моделирования ХС с точностью 0.06 (^{1}H) и 0.69 (^{13}C) м.д.
- Программа генерирования и ранжирования структурных гипотез для установления первичной структуры по экспериментальным данным.
- Углеводный язык CSDB Linear. Семантическое описание углеводов, используемое в публикациях, связано с поатомным описанием, используемым в химических расчётах.
- Стандарты визуализации углеводных структур, их одно- и двумерных спектров ЯМР. SNFG рекомендована ведущими углеводными журналами.
- База данных активностей гликозилтрансфераз.
- Анализ распространённости структурных особенностей и характерных признаков углеводов в различных таксономических группах. Альтернативные углеводные «деревья жизни».
- Правила гликоинформатики и углеводная онтология GlycoRDF. Автоматическое взаимодействие между проектами - получение знаний, неявно содержащихся в нескольких базах разного типа.
- Выявление и исправление ошибок в ~340 опубликованных структурах / спектрах и ~2 тыс. ошибок в базах данных.

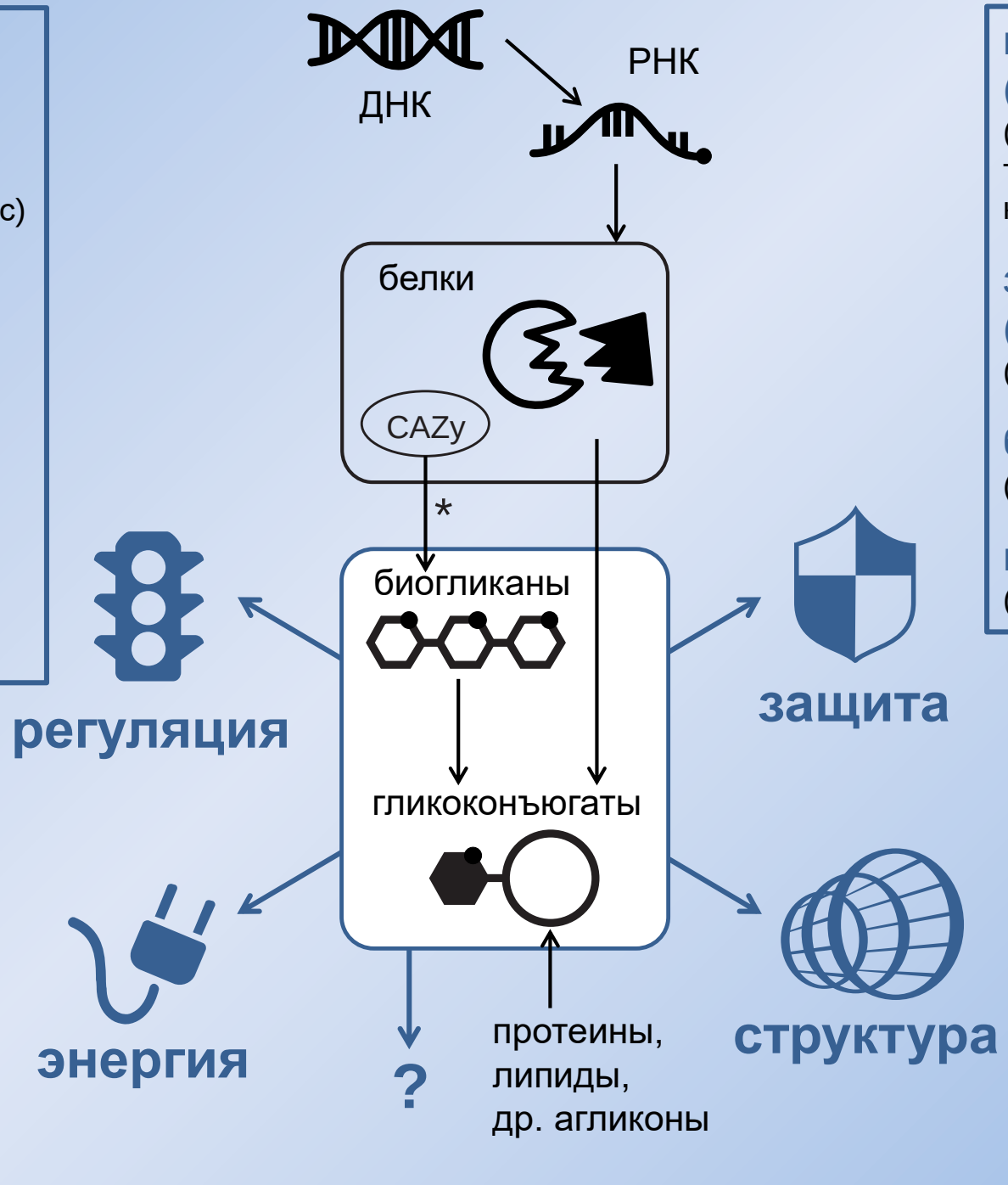
Роли углеводов

узнавание, адгезия, клеточные контакты
(лектины, ЭПС, глико-липиды, поверхностные гликопротеины, гликокаликс)

стабилизация кон-формации белков
(O- и N-гликаны)

барьерная функция, локальный pH
(O- и N-гликопротеины)

антикоагуляция
(производные гепарина)



иммунитет и мимикрия (контакт «свой-чужой»)
(ЛПС, ЭПС, O-антигены, тейхоевые к-ты, антигены крови и другие)

защита от протеолиза (контакт «свой-свой»)
(муцины, гликопротеины)

биоактивность
(растительные гликозиды)

механическая защита
(ГАГ, гиалуронаны, муцины)

биосинтез АТФ
(гликолиз, аэробный/анаэробный катаболизм, сахарфосфатный путь)

энергетический резерв
(крахмалы, гликоген)

нуклеиновые к-ты
(сахарфосфатный остов)

GPI-якоря
(фосфатидилинозитолы)

клеточная стенка
(муреин, хитин, целлюлоза, пектины, альгинаты, …)

Структурные базы

Все БД

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

GlycomeDB → **GLYTOUCAN** — 99
☺ мета-репозиторий
☹ неполные аннотации
☹ нет агликонов

CFG glycan — млекопитающие, **> 6**

SweetDB, SugaBase →
GLYCOSCIENCES.DE — **25 / 19**
PDB

$ **KEGG GLYCAN** — **11**

Eurocarb DB
☺ архитектура
☹ только модель

BCSDB — PFCSDB
13 / 5 (бактерии, археи)
7 / 2.5 (грибы, растения)
Carbohydrate Structure Database
ORIG
☺ полная по прокариотам
☺ курируемая

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

> 70
jCGGDB AIST
☺ коллекция баз
☹ слабо аннотированы

4 / 1
$ **UniCarbKB**
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
GTR
ORIG

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)

гликоэпитопы и антитела — GlycoEpitope — ~0.2 (0.6 ABs)

GlyTOUcan, репозиторий идентификаторов — ~99

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

Примеры вопросов к CSDB

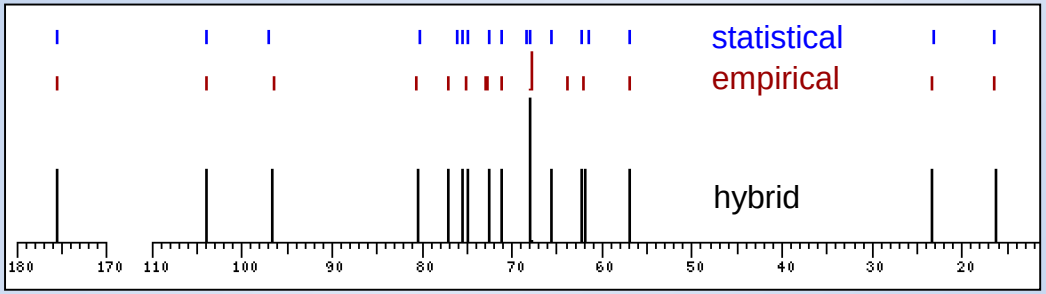
- Как введение аминогруппы влияет на химические сдвиги ЯМР в лактозном фрагменте?
- Какие структуры О-антигенов, содержащие галактуроновую кислоту и еще как минимум одну гексозу, были опубликованы после 2005-го года?
- Какие гликозиды, выделенные из растений рода паслёна, содержат агликон соланидин?
- 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?

Эмпирическое моделироваие

- Инкрементная схема со стерической коррекцией (BIOPSEL)
- Учитывает структурное окружение (9-13 дескрипторов)
- Использует специальные базы хим. сдвигов и эффектов

  (440 мономеров, 3300 димеров и тримеров, 300 эмпирических эффектов)
- Поддерживает большинство углеводных структур
- Оценивает степень достоверности предсказания

результаты смешиваются
со статистической симуляцией
с учетом достоверности



β-D-GlcpNAc-(1-3)-α-D-Fucp-(1-P-3)-D-Gro

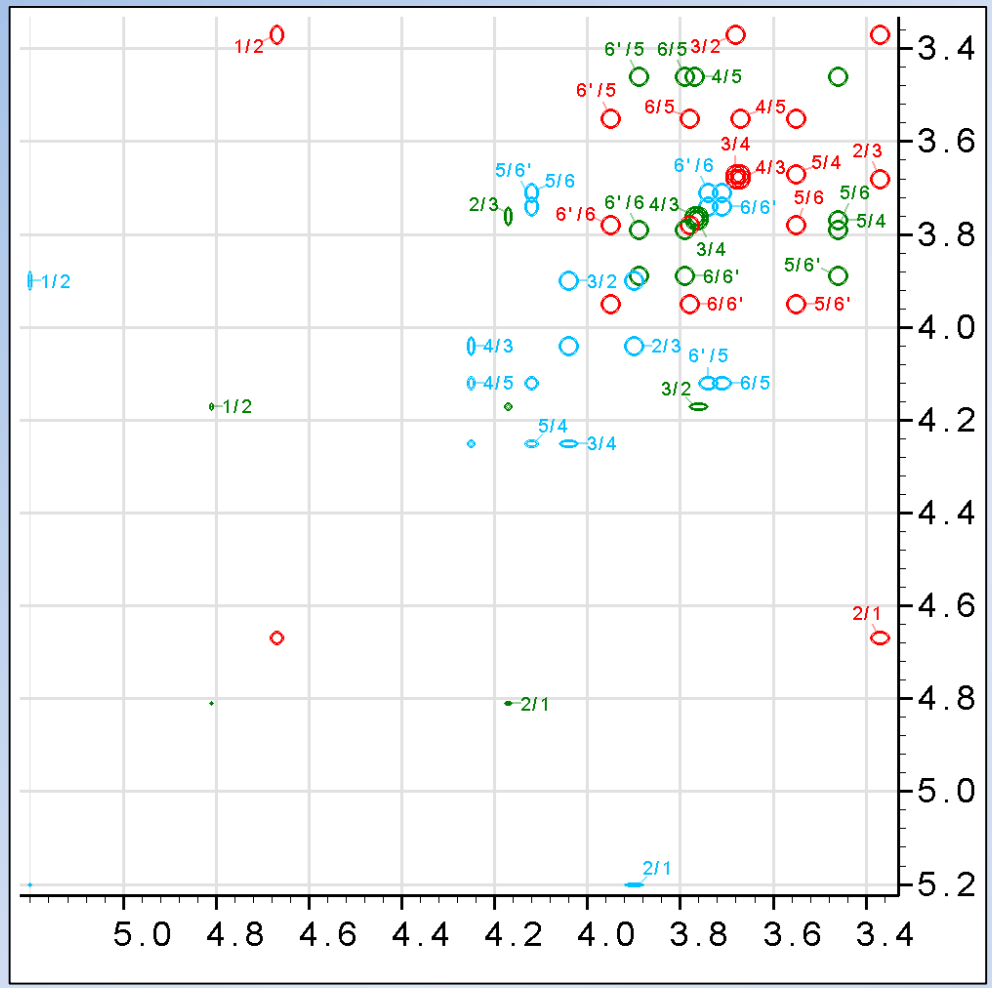
^{13}C NMR data (in D_2O):

Linkage	Residue	C1	C2	C3	C4	C5	C6	C7	C8	Accuracy
0	xDGro	64.0	73.0	68.0						*0.5*
	unsubstituted->	*64.0*	*73.5*	*64.0*						
	effect->		*-0.5*	*+4.0*						
3	xXP?									*2*
	unsubstituted->									
	effect->									
3,0	aDFucp	96.5	67.9	80.6	72.7	67.8	16.5			*4*
	unsubstituted->	*93.3*	*69.2*	*70.4*	*73.0*	*67.4*	*16.7*			
	effect->	*+3.2*	*-1.3*	*+10.2*	*-0.3*	*+0.4*	*-0.2*			
3,0,3	bDGlcpNAc	104.0	57.0	75.1	71.2	77.2	62.1	175.6	23.5	*4*
	unsubstituted->	*96.2*	*58.0*	*75.1*	*71.2*	*77.2*	*62.1*	*175.6*	*23.5*	
	effect->	*+7.8*	*-1.0*							

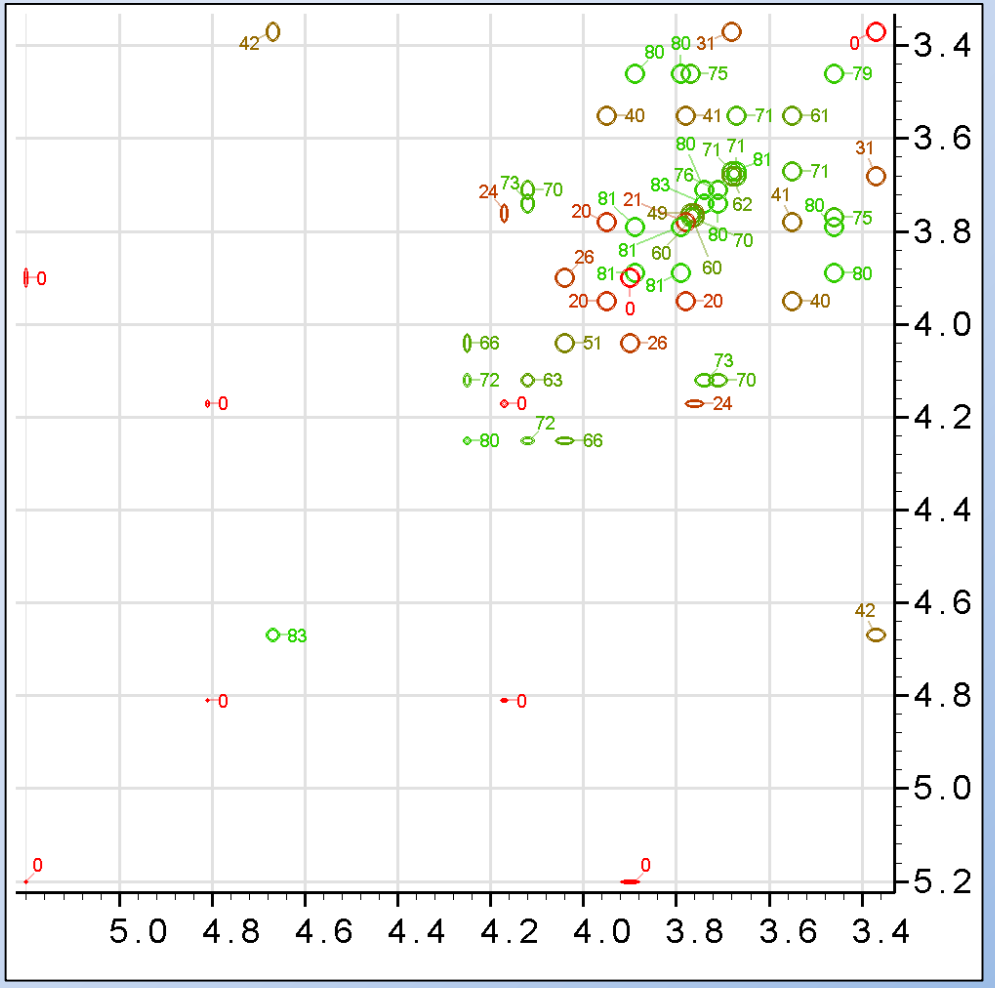
Цветовой код в 2D ЯМР

пример: ^{1}H-^{1}H COSY

→3)-α-D-Gal*p*-(1→3)-β-D-Man*p*-(1→4)-β-D-Glc*p*-(1→



принадлежность остатку



оценка достоверности

Оценка достоверности

0% ☹  100% ☺

- большой общий вес генерализаций
- в базе мало структур
- противоречивые данные ЯМР

- малый общий вес генерализаций
- в базе много структур
- данные ЯМР согласуются
- статистические и эмпирические предсказания близки

\+ ожидаемая погрешность каждого сигнала *(м.д.)* предсказывается из ХС и степени достоверности (регрессией)

Ссылки

http://glytoucan.org

Abrahams J.L. 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

JCGGDB AIST

http://jcggdb.jp/index_en.html

GLYCOSCIENCES.DE

http://glycosciences.de

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

KEGG GLYCAN

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

UniCarbKB

http://www.unicarbkb.org/

Ph. Toukach, K. Egorova **Carbohydrate Structure Database merged from bacterial, plant and fungal parts** (2016) *Nucl Acid Res* **44**, D1229–D1236. doi: 10.1093/nar/gkv840

http://csdb.glycoscience.ru

http://toukach.ru/rus/glycoinf.htm

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**

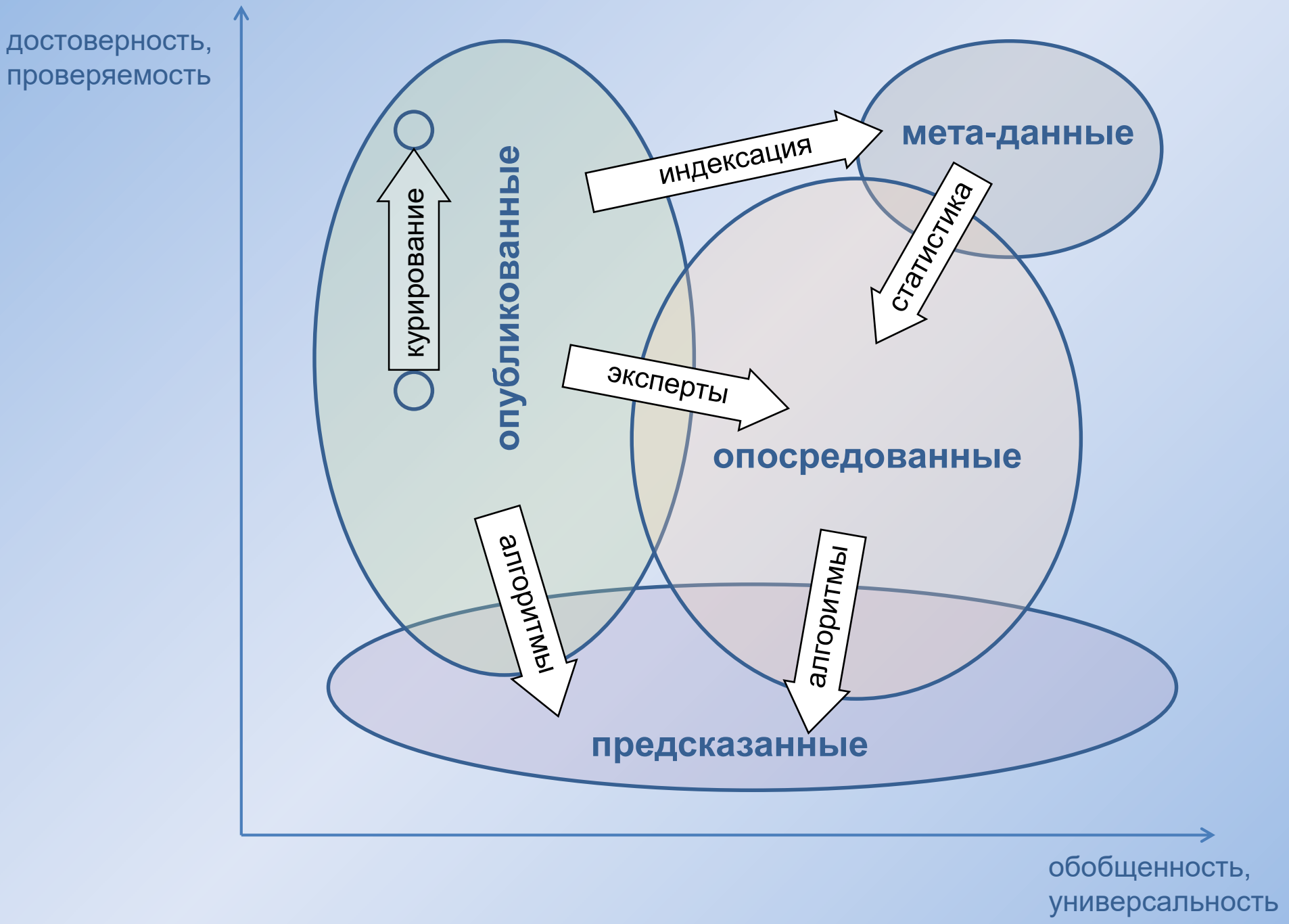
  (ввод-вывод структур, конформационные расчеты, предсказание свойств)
- **Стандартные индексы**

  (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?)

Уровни данных



Итерация структур в GRASS

фильтры:
«нет фураноз»,
widespread:
⛔ редкие остатки
(если есть обычные)

1. Наборы остатков

набор	α/β	D/L	имя	цикл
пример A	α	L	Fuc	p
B	?	D	Gal	?
C	β	D	Gal	?

A = α
B = α | α | β | β
C = β | β

фильтры:
поли/олиго
widespread:
⛔ суперразветвленные
⛔ большие боковые цепи

2. Топологии и Т-операторы

пример: 3 остатка, widespread

топо: **312**
топо: **311**
топо: **012**
топо: **011**

Т-операторы:
T1: смена узлов 2 и 3

для каждой топологии

3. Распределения наборов

T1 T1 T1

удаление дубликатов

для каждого распределения

фильтры:
разрешенные акцепторы
min/max заместителей
расположение в структуре
widespread:
⛔ суперразветвленные

фильтры:
widespread:
⛔ (1-1) связи и т.д.

5. Позиции замещения

набор	разрешено
пример A	2, 4
B	3, 4, 6
C	2, 6

T1

удаление дубликатов

для каждой структуры

4. Комбинации остатков

T1

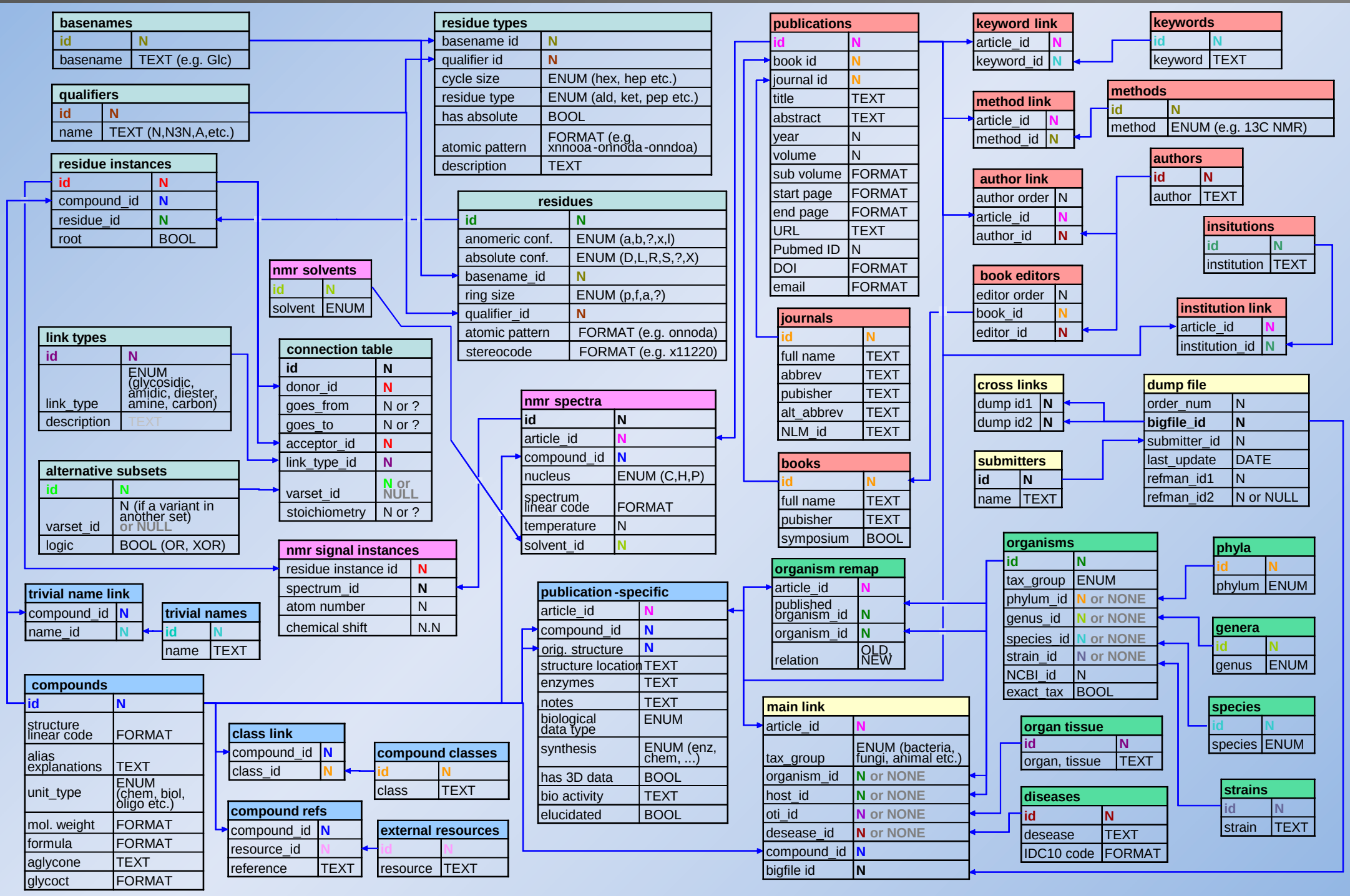
дубликат

для каждой комбинации

фильтры:
число сигналов в спектре
тип N-ацетилирования
число β-сахаров
число CH_2 групп
widespread:
⛔ суперразветвленные

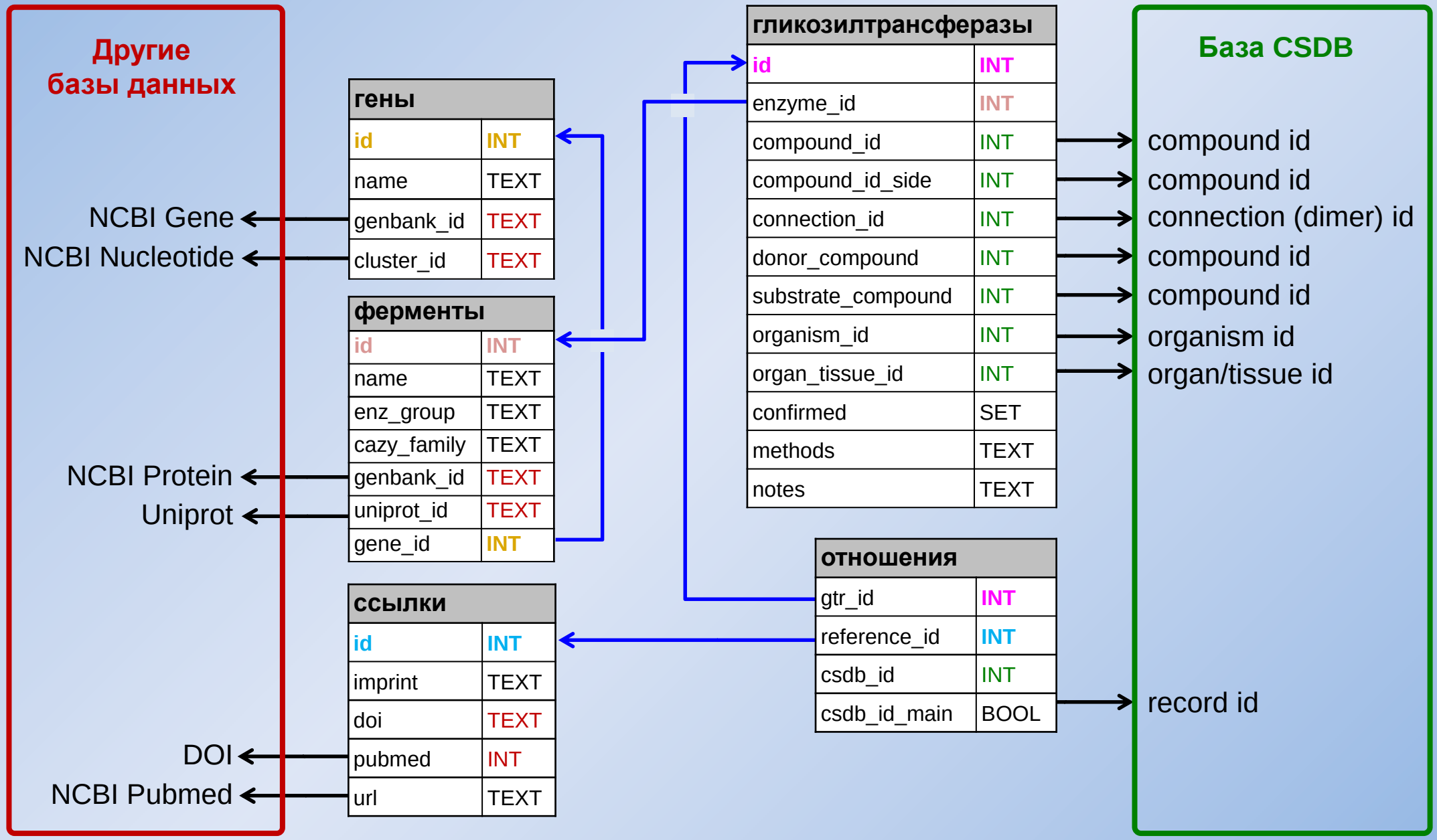
симуляция спектра ЯМР ^{13}C и ранжирование

Схема CSDB



http://csdb.glycoscience.ru/help/dbdocs.html

Схема CSDB GT



Экспорт в RDF

Структура

Ссылки (ресурс, ID, ссылка)
Записи структуры (GlycoCT, GlydeII, BCSDB, … , SweetDB)
Состав (ссылки на MSDB, non-glyco residues?)
Эксп. молекулярный вес (*+комментарии, напр. 1234 [M+]*)
Тип звена (*monosacharide or derivative, oligomer, chemical repeat, biological repeat, cyclic repeat, homopolymer repeat*)
Степень полимеризации
Тривиальное название
Структурный класс
Спектры ЯМР, отнесение, растворитель, температура
Атомные дескрипторы и координаты (*ссылка на файл*)
Изображения структуры (*ссылка на файл*)

Структура – Публикация

Методы (ссылки на MeSH)
Положение структуры в статье

Ассоциированные публикации

Ссылки (ресурс, ID, ссылка)
PMID, DOI, ISSN, ISBN, NLM ID, www-link
Название ресурса (*журнала или книги*),
Издательство, Язык
Тип ресурса (*журнал, симпозиум, глава*)
Год, том, выпуск, страницы, глава (Dublin Core)
Институты авторов
Ключевые слова

Ассоциированные организмы

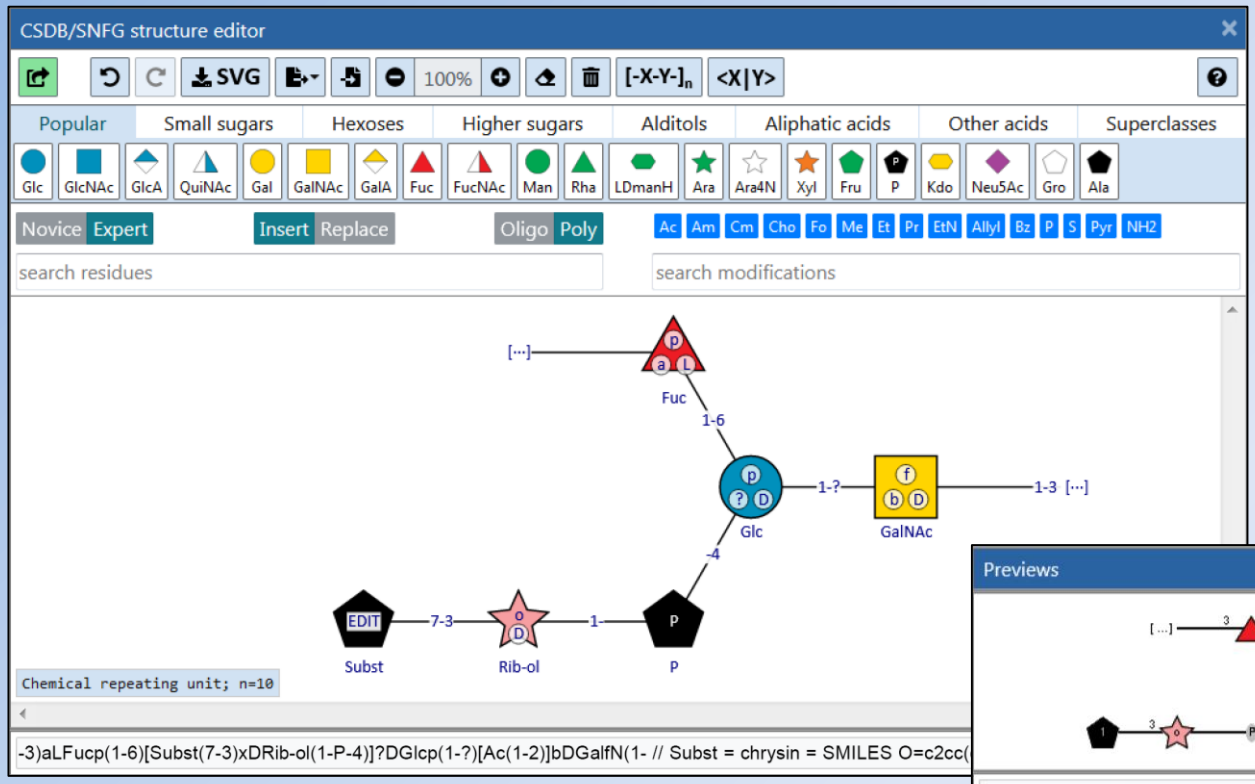
Ссылки (ресурс, ID, ссылка) (*UniProt и др.*)
Домен, Тип
Таксон (род, вид, штамм/серогруппа, ген. линия)
Хост-организм (*ссылка на запись RDF*)
Орган, ткань (*+ссылка на хост*)

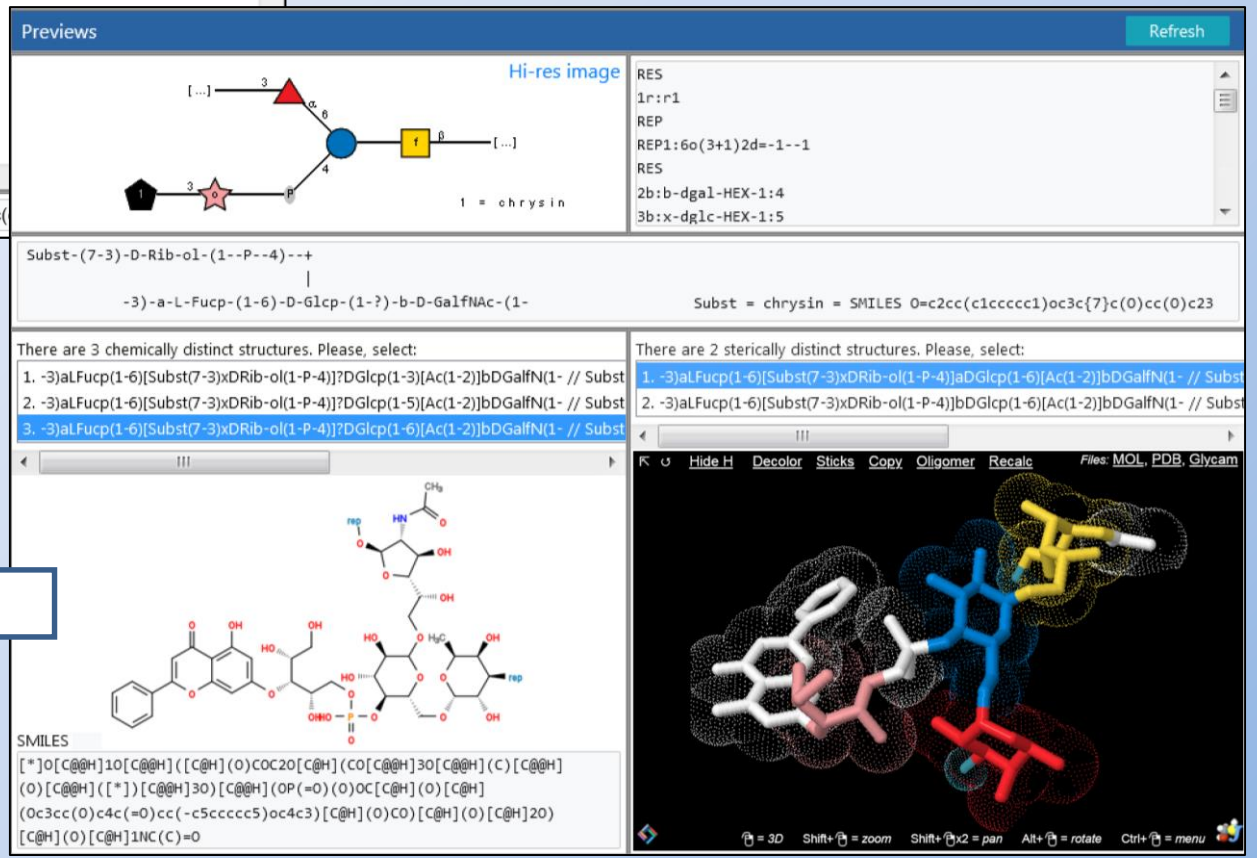
Структура – Организм

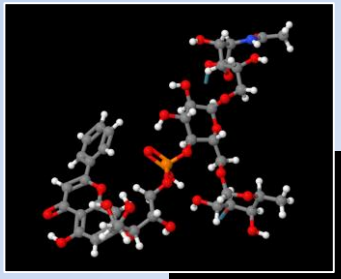
Болезнь (*IDC code*)
Стадия жизни
Вовлеченные белки (*ссылки*)
и их гены (*ссылки на GenBank*)

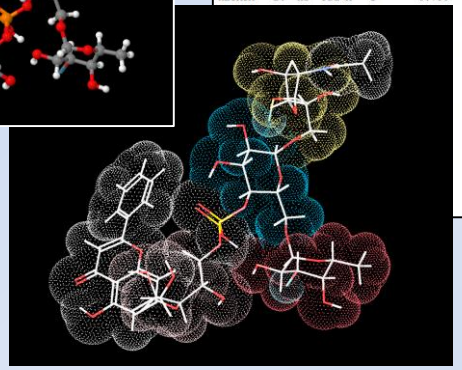
http://csdb.glycoscience.ru/integration/rdf.php
?id=<####> &mode=<structure, publication,organism,spectrum,relation,record>
&format=<turtle, rdfxml, rdfjson, ntriples> &clean=<0,1>

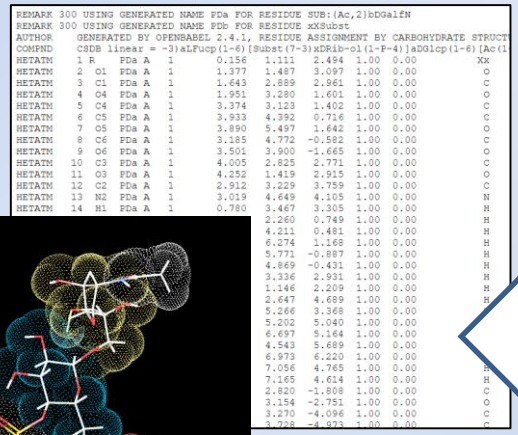
Ввод и вывод структуры









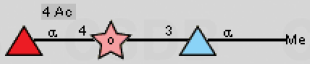


Поиск по фрагменту структуры

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

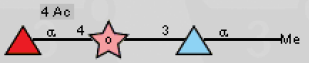
File Edit Structure View Help

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 составу

Prokaryotes + Plants + Fungi

7005 publications (1941-2017):
18923 compounds from
8859 organisms
last update: 2017 Feb 13

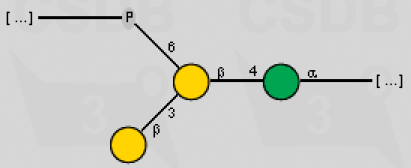
- **Search**
  - CSDB IDs
  - (Sub)structure
  - Composition
  - Taxonomy
  - Bibliography
  - NMR signals
- **Help**
  - About
  - Basic usage
  - Statistical tools
  - NMR tools
  - Usage examples
  - Advanced features
  - Structure encoding
  - Database docs
  - Credits
- **Extras**
  - NMR simulation
  - Elucidation from NMR
  - Monomer namespace
  - Fragment abundance
  - Coverage stats
  - Taxon clustering
  - Submit record
  - Translate structure
  - Feedback
- **Maintenance**

Found 6 structures. Displayed structures from 1 to 6

Expand all compounds Show all as text (SweetDB notation)

Requested composition: **2 Gal, 1 HEX, 1 P**

1. Compound ID: 6420



Show legend
Show as text

Structure type: polymer biological repeating unit ; n=~27
Aglycon: core part of the molecule
Trivial name: P-saccharide core
Compound class: lipophosphoglycan

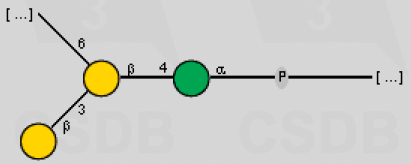
The structure is contained in the following publication(s):

- *Article ID:* 2648
  McConville MJ, Thomas-Oates JE, Ferguson MAJ, Homans SW
  "Structure of the lipophosphoglycan from Leishmania major" -
  Journal of Biological Chemistry **265** (1990) 19611-19623

  Leishmania major
  CSDB ID 235965 (all data & tools)

Expand this compound

2. Compound ID: 6361



Show legend
Show as text

Search for residue composition

Complete structural composition (3 units) :

1. phosphoric acid x 1 --> 1 x P
2. galactose x 2 --> 2 x Gal
3. any hexose x 1 --> 1 x HEX

Add unit (+) Remove unit (-)

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: 2 structures: 4104,8808

Search for molecule types: All molecule types
[x] Search for complete composition (not a fragment)
[ ] Restrict compound class: -select class-
[ ] Restrict taxonomical domain: All domains

Go! & display 30 records per page.

Home Help

Поиск по таксономии

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

Expand all organisms Show all as text (SweetDB notation)

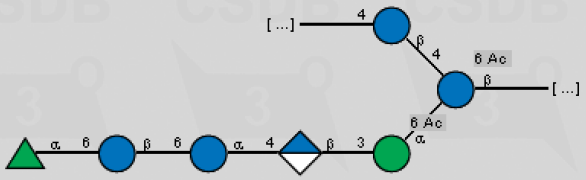
1. (*Organism ID:* 1005)

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:

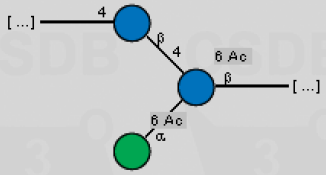
- *Compound ID:* 1717



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)

Previous results: 6 structures: <ID list>

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

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

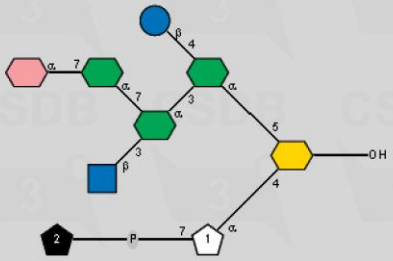
Carbohydrate Research **340(9)** (2005) 1625-1630

Following a report of variations in the lipopolysaccharide (LPS) structure c 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 c 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

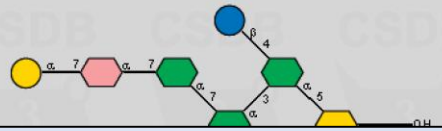


1 = a-Kop
2 = Et N

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.

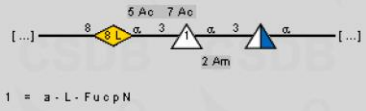
Close this window

Поиск сигналов ЯМР

Found **2** structures. Displayed structures from **1** to **2**

Expand all compounds Show all as text (SweetDB notation)

1. Compound ID: 150 *(similarity: 2.5)*



Show legend
Show as text

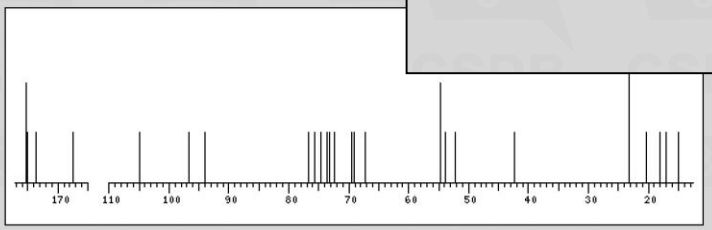
Structure type: polymer biological repeating unit

The average similarity of its ^{13}C NMR spectra with the search term (signals in bold) is **2.5** (help on similarity values)

^{13}C NMR spectra assigned to the structure:

- in Article ID 3480:
  NMR conditions: in D2O at 313 K
  ^{13}C NMR data:

Linkage	Residue	C1	C2	C3	C4	C5	C6	C7	C8	C9
3,3,5	Ac	**175.3**	23.1							
3,3,7	Ac	**175.3**	23.1							
3,3	aX8eLegp	173.7	**105.0**	**42.4**	69.2	53.8	73.6	54.7	73.2	14.9
3,2	Am	167.5	20.4							
3	aLFucpN	96.7	52.3	75.6	72.4	67.3	17.1			
2	Ac	**175.0**	23.3							
	aDQuipN	94.1	54.6	76.7	74.6	69.6	18.0			



The structure is contained in the following publication(s):

- *Article ID:* 31
  Bystrova OV, Lindner B, Moll H, Kocharova NA, Knirel YA, Zähringer U, Pier GB **"Structure of the lipopolysaccharide of Pseudomonas aeruginosa O-12 with a randomly O-acetylated core region"** - *Carbohydrate Research* **338(18)** (2003) 1895-1905

  Pseudomonas aeruginosa O12
  CSDB ID 1824 (all data & tools)

- *Article ID:* 3480
  King JD, Mulrooney EF, Vinogradov E, Kneidinger B, Mead K, Lam JS **"IfnA from Pseudomonas aeruginosa O12 and wbuX from Escherichia coli O145 encode membrane-associated proteins and are required for expression of 2,6-dideoxy-2-acetamidino-L-galactose in lipopolysaccharide O antigen"** - *Journal of Bacteriology* **190(5)** (2008) 1671-1679

  Pseudomonas aeruginosa O12
  CSDB ID 23007 (all data & tools)

Expand this compound

2. Compound ID: 11058 *(similarity: 1.364)*

Search for NMR signals

Nucleus: Carbon Proton

Threshold: 1.2 Threshold explanation

Chemical shifts:
105.15 42
176

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)

Signals within a single residue

Previous results: 2 structures: 150,11058

Go! & display 30 records per page.

Home Help

Поиск по ID и полная запись

2. (BCSDB ID: 50001)

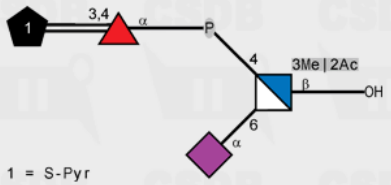
Egorova KS, Toukach FV

This is a test record for BCSDB 3 debug N2

Carbohydrate Research **489** (1944) 9726

```
                    a-Neup5(10%)Ac-(2-6)-+
                                         |
S-Pyr-(2-4:2-3)-a-Fucp-(1--P--4)--b-D-GlcpN
                                         |
                          /Variants 0/-+
```

```
/Variants 0/ is:
Me-3)-
OR (inclusively)
Ac-2)-
```



1 = S-Pyr

***Deliriumus trementii* O67 PCM2005**

(Ancestor NCBI TaxID 374, species name lookup)

Taxonomic group: bacteria / Pseudophylum, Newphylum *(Phylum: Pseudophylum, Newphylum)*
Host organism: Worra worra, Homo sapiens
Organ / tissue: brain, ass
Associated disease: googlomania

NCBI PubMed ID: 123456789
Journal NLM ID: 0043535
Publisher: Elsevier
Institutions: N.D. Zelinsky Institute of Organic Chemistry, Hirzfeld Institute

The structure of the O-specific polysaccharide of Deliriumus trementii O66 has been elucidated using 2d-NMR approach, including... The studies of biolo
the effect of this glycopolymer on sexual behaviour of Worra-worras bla-bla-bla...

structure, antigen, polysaccharide, worra, Deliriumus trementii

Structure type: oligomer ; 2000-3000
Location inside paper: HPLC fraction 8
Trivial name: khrenobiose
Compound class: X-antigen, Y-antigen

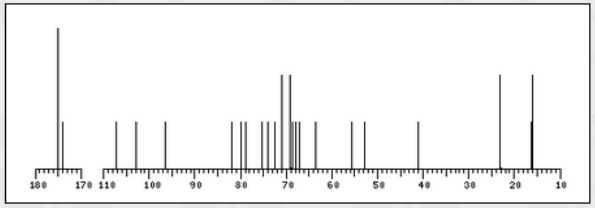
Methods: 1D-NMR, 2D-NMR, 13C-NMR
Biological activity: causes suppression of Worra-worra's natural instincts
Enzymes that release or process the structure: CoQ-III
Biosynthesis and genetic data: biochemical data
Comments, role: structure was revised (see RR 1234), absolute configurations are not determined
3D data: conformation data, computer modelling, dynamics

NCBI Taxonomy refs (TaxIDs): 6661313, 374
Reference(s) to other database(s): SpecDB:o1.9, CA:123, CA-RN:456, patent:USA#123, ProtDB:ABC456

NMR conditions: in D2O at 308 K

^{13}C NMR data:

Linkage	Residue	C1	C2	C3	C4	C5	C6	C7	C8	C9
6,5	10%Ac	175.0	23.0							
6	aXNeup	174.1	107.2	41.0	69.0	52.9	73.9	69.2	72.6	63.6
4,0,4	xSPyr?	175.1	80.0	16.0						
4,0	a?Fucp	96.5	68.6	71.0	81.9	67.8	15.9			
4	P									
0	bDGlcpN	103.0	55.6	78.9	70.9	75.4	67.1			
2	Ac	175.0	23.2							
3	Me	16.3								



Search for CSDB IDs

Scope: Record IDs | Structure IDs | Publication IDs | Organism IDs
RDF only: Source IDs | Spectrum IDs | Relation IDs

CSDB record IDs: 1,6,50001,200-400
You can use commas to separate IDs and hyphens to specify the range, e.g. 100-150,160-165,170

Go! & display 30 records per page.

Make RDF in Turtle | DCI XML | Home | Help

Found **88** records. Displayed records from **31** to **60**
Previous 30 record(s) Next 30 record(s)

Expand all records | Show all as text (SweetDB notation)

Report data error

< ПОЛНЫЕ ЗАПИСИ ЗДЕСЬ >

Previous 30 record(s) Next 30 record(s)

Resort records by: publication year

New query | Home | Help

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

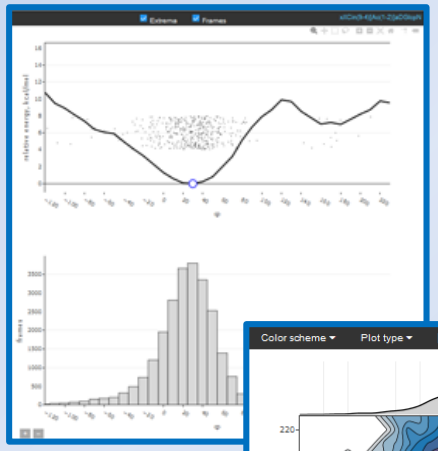
Force field: *Filter by MD method* — MM3-2000

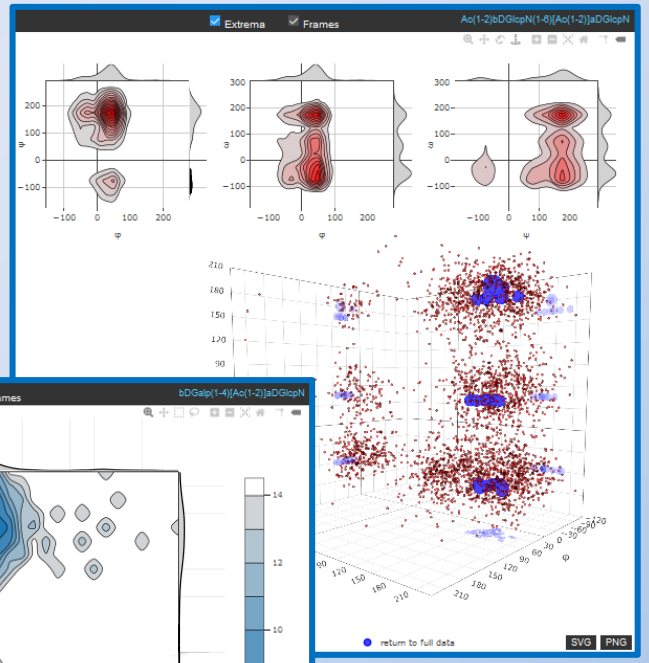
Temperature: *Filter by MD temperature* — 1000

Solvent model: *Filter by solvent model* — any

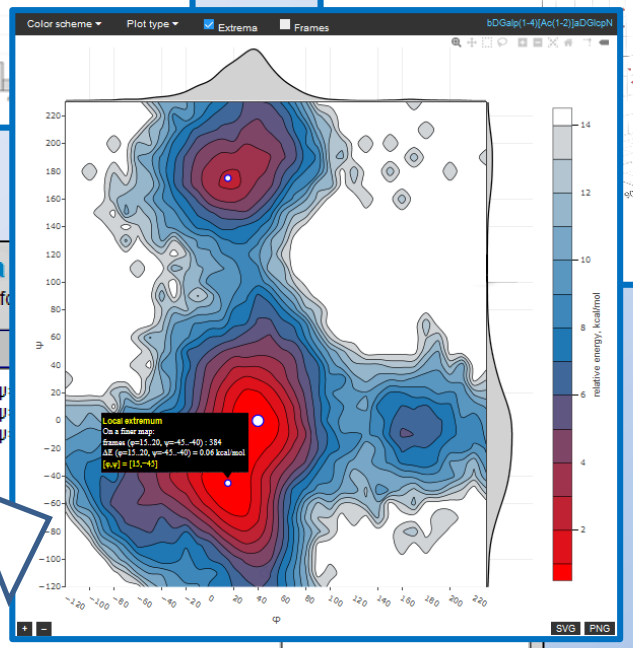
Home Help

1D





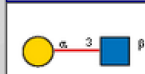
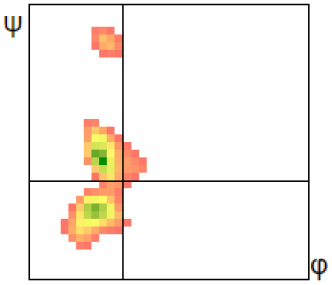
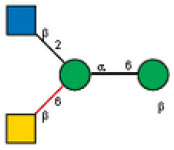
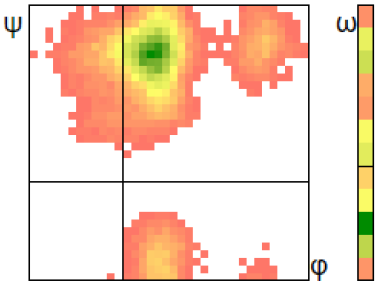
3D



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	
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

Фрагменты и таксоны

CSDB dimer abundance

The table lists **17** dimeric fragments present in **8** saccharides associated with **2** organisms from:
Herpetomonas muscarum, Herpetomonas samuelpessoai (species).

Anomeric forms were combined. Dimers containing monovalent constituents were excluded. Residues with undefined configurations or ringsizes are greyed.
Superclasses are in blue. To re-sort the list click the according column name.

Position	Donor	Linkage	Acceptor	Abundance	Compound IDs	Abundance in selected species
terminal	DManp	1-2	DManp	18 (17%)	5555, 5556, 5557, 5559, 5560, 5848	*Herpetomonas samuelpessoai*: 12 (67%) *Herpetomonas muscarum*: 6 (33%)
di-branched	DManp	1-3	DManp	15 (14%)	5345, 5555, 5556, 5557, 5558, 5559, 5560, 5848	*Herpetomonas samuelpessoai*: 11 (73%) *Herpetomonas*
inline linear	DManp	1-2				*muscarum*: 3 (43%)
			DManp	5 (4.7%)	5556, 5557, 5559, 5560	*Herpetomonas samuelpessoai*: 4 (80%) *Herpetomonas muscarum*: 1 (20%)
tri-branched, reducing end	DGlcpN	1-4	DGlcN	5 (4.7%)	5345, 5557, 5558, 5559, 5560	*Herpetomonas samuelpessoai*: 5 (100%)
di-branched, reducing end	DGlcpN	aglycone	diphosphodolichol	4 (3.7%)	5555, 5556	*Herpetomonas samuelpessoai*: 2 (50%)
tri-branched	DGlcpN	1-4	DGlcpN	4 (3.7%)	5555, 5556	
tri-branched	DManp	1-3	DGlcpN	2 (1.9%)	5848	
penta-branched	DGlcpN	1-4	DGlcpN	2 (1.9%)		
inline linear	DManp	1-3	DManp	1 (0.9%)	5557	
terminal	DGlcp	1-2	DManp	1 (0.9%)	5558	
Total				*107 (100%)*		

Export TSV Monomers Home

Monomer and dimer abundance

This page will generate a pool of monomers or dimers abundant in glycans from the selected taxonomic group(s).
First, please select a taxonomic rank of a desired group: species

Display groups: bacteria archaea protista algae fungi plants animals

Genus: (select only one)
Herbaspirillum
Herbidospora
Herpetomonas
Histophilus
Hyphomonas
Idiomarina
Inquilinus
Kaistella
Kineosporia
Kingella
Klebsiella
Kocuria

Species: (select multiple with CTRL key)
sp. (unassigned)
muscarum
samuelpessoai

Select all

- Combine anomeric forms
- Include undefined configs
- Include ONLY saccharides
- Include monovalent residues
- Include aglycons in oligomers
- Include aliases
- Explain 'Subst' aliases

- Distinguish inline / terminal / reducing
- Distinguish branching degree
- ...and ignore monovalent substituents

Display only those fragments that are unique for this species in all biota

Display monomers Display dimers

Monomer namespace Home Help

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:

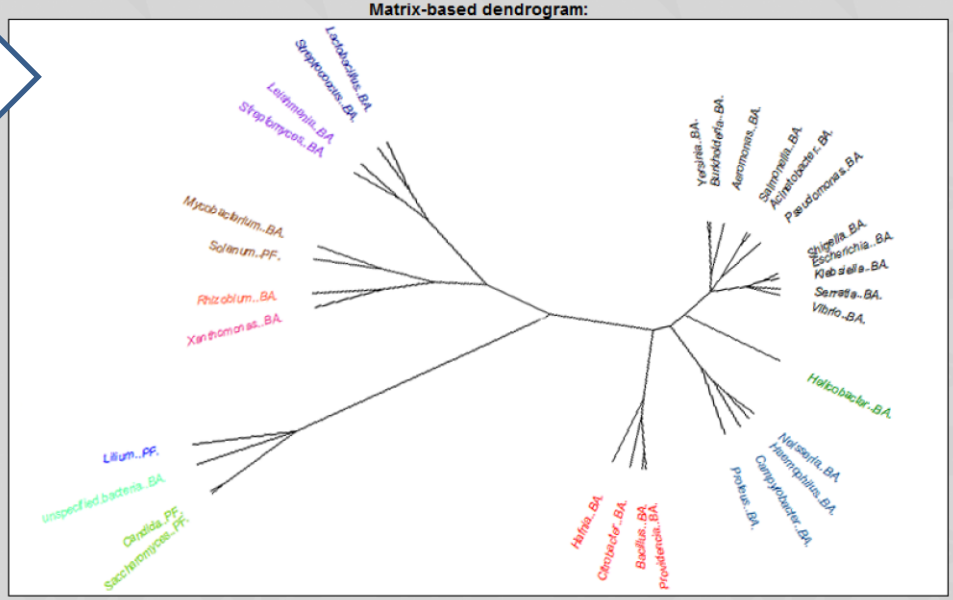
```
Hamming mode: YES
Fragment size: monomer
Fragment abundance filter: instance threshold: 550
Fragment abundance filter: structures threshold: 500
Fragment presence threshold: 2
                                    (atypical residues): NO
                structures of this type: any
Filter: differentiate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

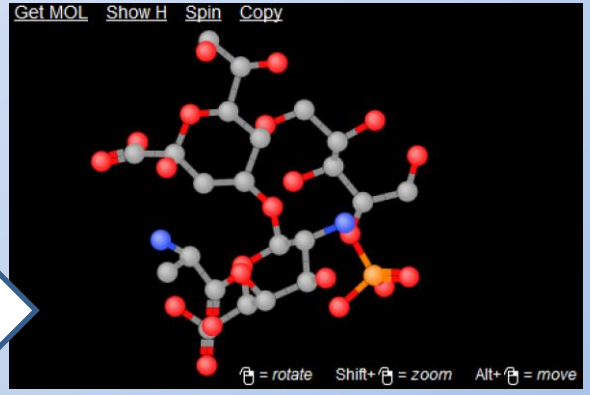
✓ Save generated structures

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



Соответствие специальности

Методологические компьютерные работы традиционно защищаются по специальности, для которой предназначены создаваемые инструменты.
Формулировки, охватываемые темой работы, в паспорте специальности **02.00.10** «биоорганическая химия» и их расшифровки в программе специальности :

- «Изучение структуры и функций биомолекул […] физико-химическими методами; структурно-функциональные исследования полисахаридов и смешанных биополимеров».
- «Моно-, олиго-, полисахариды: номенклатура, стереохимия, конформация. Методы изучения строения».
- «Соединения из микроорганизмов, грибов, водорослей, растений, […] липополисахариды бактерий».
- «Спектроскопия ЯМР, […] связь параметров спектров ЯМР с химической и пространственной структурой биомолекул, […] двумерная спектроскопия ЯМР».
- «Компьютерное моделирование молекулярной механики биомолекул».
- формулировки, связанные с компьютерным моделированием и расчётами геометрии *белков*. На момент составления паспорта специальности моделирование *гликанов* ещё не было популярно.

Кроме этого, работа расширяет методологию биоорганической химии на новые области, не указанные в описании специальности.