ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ
НАУКИ ИНСТИТУТ ОРГАНИЧЕСКОЙ ХИМИИ ИМ.
Н.Д. ЗЕЛИНСКОГО РОССИЙСКОЙ АКАДЕМИИ НАУК

---

*на правах рукописи*

ТОУКАЧ ФИЛИПП ВЛАДИМИРОВИЧ

# ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В СТРУКТУРНОЙ ГЛИКОХИМИИ И ГЛИКОБИОЛОГИИ

02.00.10 – Биоорганическая химия

ДИССЕРТАЦИЯ
на соискание учёной степени
доктора химических наук

Научный консультант:
д.х.н. проф. Ю.А. Книрель

Москва 2019

## Оглавление

# 1. Введение

## *1.1. Актуальность проблемы*

Углеводы – важные носители биологической информации, наряду с белками и нуклеиновыми кислотами. Углеводы выполняют разнообразные функции в клетках, в том числе структурные и энергетические; именно эти молекулы определяют ответ организма на заражение патогенами и участвуют в установлении иммунитета. Однако активные исследования роли углеводов в биологических процессах начались относительно недавно. Это одна из причин, по которым информационное обеспечение гликомики сильно отстаёт от геномики и протеомики, что затрудняет доступ учёных к накопленной информации и инструментам её обработки. Другая причина заключается в значительном химическом разнообразии углеводов и сложности их анализа. В результате учёные сталкиваются с нехваткой моделей и стандартов записи углеводных данных, отсутствием полных хранилищ данных и информационной изолированностью существующих проектов. Следует отметить отсутствие полных репозиториев экспериментальной информации о ферментативном аппарате, вовлечённом в биосинтез углеводных структур, которая востребована при разработке протоколов ферментативного синтеза ценных биологических продуктов.

Создание платформы, способной как хранить, так и перерабатывать данные и обеспечивающей доступ к структурным и биосинтетическим данным, устраняет отставание гликоинформатики от других компьютерных дисциплин, связанных с молекулярными носителями жизни, значительно облегчит исследования происхождения, строения и функций природных углеводов. Особенно это актуально для углеводов бактерий, растений и грибов. Несмотря на востребованность в химии, биологии и медицине, данные по этим доменам значительно хуже представлены в существующих базах, по сравнению с данными по гликанам животных (особенно млекопитающих), в том числе из-за большего разнообразия структур и сложностей с их формальным описанием.

Наличие инструментов обработки данных, привязанных к платформе базы данных, открывает доступ к неявно присутствующим в базе знаниям. Эти ин-

струменты позволяют неподготовленным в плане информатики учёным получать информацию, доступ к которой ранее требовал направленных компьютерных изысканий. Тем не менее, в химии углеводов наблюдается несоответствие огромного объёма накопленных структурных данных ограниченным возможностям их обобщения и прогнозирования свойств либо самой структуры. В частности, основной аналитический метод в структурной гликохимии (спектроскопия ЯМР) плохо обеспечен средствами автоматической интерпретации экспериментальных данных, что делает структурные исследования трудоёмкими и исключает их массовость. Предложенные в работе подходы позволили на порядок удешевить и ускорить установление первичной структуры природных углеводов. В свете того, что структура О-антигенов многих микроорганизмов не установлена, эти инструменты упростили поиск эпитопов взаимодействия «антиген – антитело», что важно для объяснения иммунного ответа на молекулярном уровне и для классификации патогенных микробов.

Стоит отдельно отметить инструмент прогнозирования молекулярной геометрии биогликанов и гликоконъюгатов. Геометрические и энергетические расчёты в гликохимии недоступны пользователям без специальной подготовки, а также требуют значительных вычислительных ресурсов. Из-за этого подбор структур с помощью потокового прогнозирования свойств, зависящих от вторичной структуры, проводится крайне редко, что тормозит поиск сахаридов с желаемыми свойствами. В первую очередь это касается взаимодействия с ферментами и биологической активности. Востребованность новых подходов к моделированию структуры основывается на том, что они позволяют автоматически предсказывать и хранить данные для десятков тысяч структур, характерных для биогликанов, в том числе идентичных уже описанным полисахаридам, гликозидам и гликоконъюгатам. Эти данные могут быть использованы для выявления кандидатов для детального анализа в скрининговых и статистических исследованиях.

Важность стандартизации углеводных данных была осознана лишь недавно благодаря росту популярности автоматической обработки данных в машиночитаемых форматах для поиска корреляций «структура - свойство» путём перебора и сравнения. Стандартизация позволяет связать «изолированные острова»

данных о биогликанах и получать разнотипные знания, распределённые по нескольким базам (в том числе фильтруя данные из одних баз по критериям, представленным в других базах). Предлагаемый в работе способ стандартизации с помощью модели Resource Description Framework и углеводной онтологии открывает путь к эффективной интеграции с существующими компьютерными ресурсами в химии и биологии углеводов.

Работа направлена на решение вышеописанных проблем как на фундаментальном, так и на методологическом уровне. Её материальное воплощение включает универсальную платформу гликоинформатики, объединяющую в себе базу данных природных углеводов бактериального, грибного и растительного происхождения (Carbohydrate Structure Database, CSDB), их производных и углевод-активных ферментов, участвующих в их синтезе, формальную углеводную онтологию, стандарты обмена информацией и форматы данных, инструменты предсказания свойств биогликанов (спектры, молекулярная геометрия и т.д.), инструменты ввода, визуализации и статистической обработки данных, специфичных для гликохимии и гликобиологии, и семантические связи с другими значимыми углеводными проектами. Соображения, изложенные в этом разделе, делают представленный междисциплинарный проект актуальным для всей науки об углеводах.

### *1.2. Цели работы*

Целью работы являлась оптимизация и автоматизация структурно-функциональных исследований углеводов, привнесение в гликомику уровня информационной обеспеченности, сравнимого с существующим в геномике и протеомике. Для достижения этой цели были сформулированы следующие задачи:

1. Проектирование, разработка, наполнение данными и поддержка базы данных природных углеводов, включающей информацию о структуре, таксономии, библиографии, спектрах ЯМР и другие данные, востребованные в изучении строения и свойств сахаридов и гликозидов. Эта база данных должна поддерживать множество видов поиска данных, быть недорогой в обслуживании с ростом числа записей, в перспективе иметь полное покрытие по всем природным углеводсодержащим молекулам и идеологически заменить собой CarbBank. Функции базы должны быть свободно доступны как химикам и биологам (через веб-портал), так и другим проектам глико- и хемоинформатики (через автоматические веб-сервисы).
2. Разработка алгоритмов, позволяющих получать данные о геометрии и конформации углеводов за разумное время и неподготовленными пользователями. Эта задача подразумевает создание промежуточной базы данных геометрии мономерных остатков, базы конформационных карт нежёстких фрагментов в сахаридах и полностью автоматических инструментов для молекулярно-динамических расчётов в молекулярно-механических силовых полях.
3. Изучение корреляции «структура - спектр», выявление структурных дескрипторов сахаридов, влияющих на спектральные параметры, и создание подходов к моделированию спектров ЯМР углеводов с точностью и скоростью, позволяющими химикам и биологам использовать эти данные в исследованиях структуры природных соединений. Эта задача также включает создание методологии оценки достоверности моделирования и статистическую валидацию моделей на большой выборке природных структур.
4. Разработка алгоритма сравнения спектров с учётом неопределённостей и инструмента предсказания первичной структуры олиго- и полисахаридов по легко получаемым экспериментальным данным, таким как одномерные спектры ЯМР, данные ГЖХ и экспериментов по метилированию.

5. Разработка языка описания структуры гликанов, гликополимеров и гликоконъюгатов, пригодного (в отличие от других языков) как для человеческой, так и для машинной интерпретации и обеспечивающего однозначное описание структуры любых углеводов и их производных, включая те, структурная информация для которых определена не полностью. Эта задача подразумевает также создание программ перевода информации на существующие углеводные (GlycoCT, Sweet-II, WURCS и т.д.) и общехимические (IUPAC, SMILES и т.д.) языки и с них.
6. Разработка интуитивно понятного способа визуализации углеводных структур в программах и публикациях, учитывающего все структурные особенности, характерные для биогликанов и обратно-совместимого с существующими публикациями и схемой визуализации, предложенной в 1980-х годах Консорциумом по функциональной гликомике.
7. Сбор данных о ферментах, вовлечённых в биосинтез углеводов, и создание базы данных, связывающей гены, ферменты, их активность, углеводные структуры и штаммы организмов, в которых эти структуры синтезируются.
8. Статистическое исследование особенностей химической структуры биогликанов, характерных для различных групп живых организмов (от царств до родов), и сравнительный анализ химического разнообразия гликомов в различных таксонах. Эта задача также включает построение альтернативных фенетических «деревьев жизни», основанных на сходствах и различиях гликомов, и их сравнение с классическими филогенетическими деревьями.
9. Разработка идеологии и правил обработки информации об углеводах, которые позволят сократить отставание гликомики от других наук о молекулярной основе жизни. Эта теоретическая основа гликоинформатики должна учитывать специфические для углеводов структурные и биологические особенности, а также исторически сложившиеся стандарты и ошибки прежних проектов. В эту цель входит и объединение мировых проектов гликоинформатики в единую информационную среду (в сотрудничестве с другими группами), включая прозрачную для пользователей интеграцию баз данных, создание формальной углеводной онтологии, стандартизацию используемых в гликоинформатике моделей данных, индексов и идентификаторов.

### *1.3. Результаты и их значимость*

В работе решена важная научно-прикладная проблема – устранён пробел в информационном обеспечении гликомики, связанный с отсутствием универсальных баз данных, объединяющих информацию по природным углеводам с компьютерными инструментами её анализа. Несмотря на существование отдельных баз по углеводам различных таксономических групп, ни одна из них не обеспечивала полного покрытия и не содержала исчерпывающей информации о ферментативном аппарате, вовлечённом в биосинтез углеводов. Более того, из-за отсутствия общепринятых форматов представления углеводных структур обмен информацией между этими базами данных был значительно затруднён, что ограничивало эффективность работы учёных. В результате представленной работы огромное количество накопленных данных об углеводах получило средства навигации в этом информационном поле.

Результатом проекта стала универсальная междисциплинарная платформа гликоинформатики на основе базы данных природных углеводов (CSDB), которая объединила данные по структурам исследованных природных углеводов бактериального, грибного и растительного происхождения и их производных, дополненные аналитической, таксономической, библиографической и другой информацией, с данными по ферментативному аппарату, участвующему в их биосинтезе (углевод-активные ферменты). Эта информация востребована в современной биологии, химии и медицине, особенно при разработке методов синтеза гликоконъюгатных продуктов (например, иммуностимуляторов и вакцин). Созданная платформа оснащена новыми для области инструментами ввода, проверки, визуализации и статистической обработки данных, специфических для гликохимии и гликобиологии. Идеология, модели, алгоритмы, собранные данные и новые веб-сервисы были интегрированы с проектами конкурентов с использованием разработанных стандартов.

Впервые представлены инструменты точного предсказания химических свойств биогликанов, пригодные для использования в потоковом режиме на больших множествах структур. Они включают алгоритмы моделирования вторичной структуры, основанные на новой базе конформационных карт углеводных фрагментов, алгоритмы моделирования спектров ЯМР, генерирования и

оценки достоверности структурных гипотез. Эти инструменты активно используются для структурных исследований природных гликополимеров бактерий и грибов и механизмов их взаимодействия с клеточными структурами в высших организмах.

Кроме того, в рамках проекта проведена стандартизация представления структурной информации по углевод-содержащим молекулам и создана специальная углеводная онтология, что впервые позволило наладить обмен данными между наиболее значимыми в настоящее время проектами, обеспечивающими химиков информацией о структурах и таксономии биогликанов (CSDB, Glytoucan, Glycosciences.de, UniCarbKB и Japan Consortium for Glycobiology and Glycotechnology DataBase).

Вышеперечисленные работы позволили снизить стоимость и качественно повысить эффективность как фундаментальной, так и прикладной работы учёных в широкой области знания – науке об углеводах. Разработки востребованы и уже использованы в многочисленных исследованиях структуры и функций биогликанов, проводимых другими группами биохимиков, молекулярных биологов, иммунологов и фармацевтов. Эти исследования включают установление строения новых природных антигенов бактерий, выявление молекулярных маркеров таксономических групп, выявление гликоэпитопов, вызывающих иммунный ответ на бактериальные инфекции в высших организмах, выяснение фундаментальной связи между химической структурой и наблюдаемыми спектральными параметрами углеводов, гликопротеинов и гликоконъюгатов, выяснение активности углевод-активных ферментов, хемотаксономическую классификацию патогенных микроорганизмов и многие другие.

Представленные компьютерные инструменты гликомики спроектированы на фундаментальной основе и реализованы в виде работающих прикладных инструментов, свободно доступных научной общественности. Они объединены в согласованную систему, верифицированы на модельных объектах и использованы для реальных исследований. За 13 лет развития платформа CSDB заняла ведущие позиции в мировой науке об углеводах и имеет перспективы стать единственной всеобъемлющей базой по углеводам (всеобъемлющих баз не существует с 1996 года, когда по причине просчётов в проектировании, приведших к экс-

поненциальному удорожанию обслуживания, была прекращена поддержка Carbbank). Заложен фундамент для статистических и прямых расчётов корреляции структура-свойство в химии и биологии углеводов. Преобразилась молодая область знания – гликоинформатика, задан и обеспечен мировой вектор её развития.

## 2. Литературный обзор

### *2.1. Роль углеводов и гликоинформатики в науках о жизни*

Углеводы считаются самыми распространёнными биологическими молекулами на Земле, а их окисление является главным способом получения энергии большинством нефотосинтезирующих клеток [1, 2]. Из универсальных характеристик всех живых клеток, гликом (совокупность углеводов вида или организма) демонстрирует наибольшую эволюционную изменчивость и разнообразие ролей [3-5] (Рис. 1). В то же время объем накопленных знаний о биологических функциях гликома отстает от такового в других областях. Несмотря на активное изучение и актуальность темы гликозилирования в медицинскимх исследованиях [6-10] многие детали функционирования биогликанов и связанных с ними патологий остаются невыясненными [11, 12].

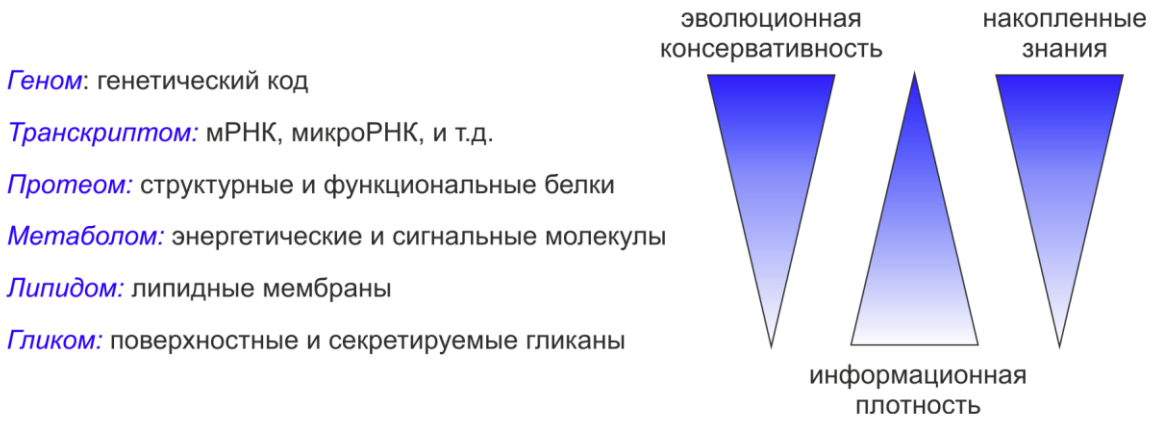



Рис. 1. Соотношение свойств универсальных составляющих живых клеток. Воспроизведено из [4].

Положение углеводов и их производных в центральной догме молекулярной биологии и основные роли биогликанов в живых организмах просуммированы на Рис. 2. Углеводные полимеры, олигомеры и конъюгаты выполняют структурную роль в клеточных стенках растений, бактерий и грибов и в соединительных тканях животных, участвуют в межклеточном взаимодействии и адгезии, играют ключевую роль в развитии и установлении иммунного ответа [13, 14]. Сложные сахариды, ковалентно связанные с липидами и белками, представляют собой сигнальные молекулы, определяющие внутриклеточную либо метаболическую «судьбу» таких гликоконъюгатов [15]. Тем не менее, в молекуляр-

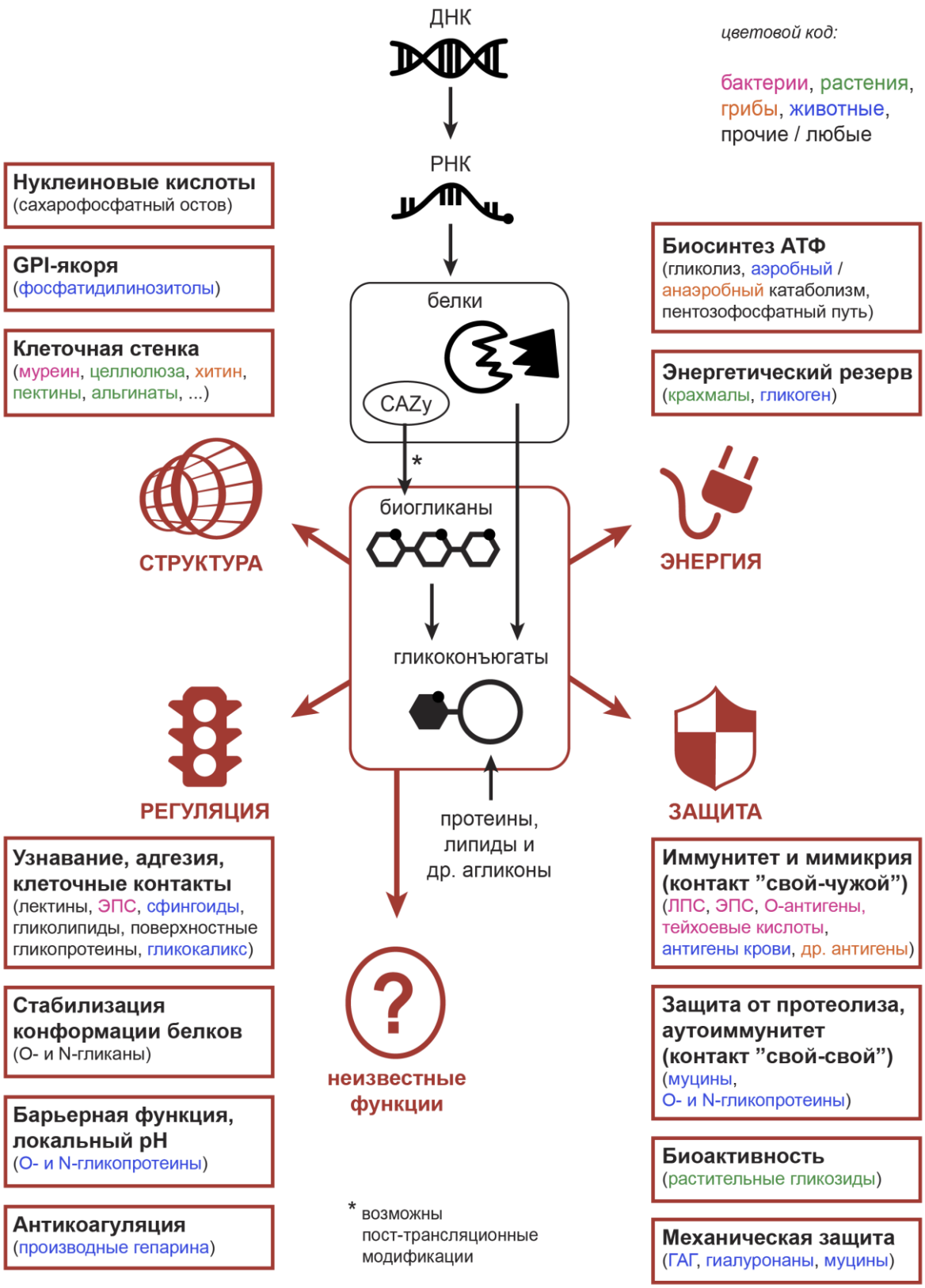



Рис. 2. Положение биогликанов в глобальной схеме биосинтеза компонентов клетки и их основные функции в живых организмах (обозначены красным). Цветовой код примеров в скобках отражает таксономический домен в соответствии с объяснением в верхнем правом углу. ЛПС – липополисахарид, ЭПС – экзополисахарид, ГАГ – гликозаминогликаны.

ной биологии углеводам лишь относительно недавно начали уделять не менее пристальное внимание, чем другим молекулярным носителям жизни – белкам и нуклеиновым кислотам. Открытие функциональной значимости гликозилирования белков способствовало интенсивному изучению структуры и функций природных углеводов [16-18], и их функциональной классификации на генетической основе [19]. В результате появились новые активно развивающиеся области знания – системная гликобиология [20] и гликоинформатика [21].

Углеводная клеточная стенка микроорганизмов - это динамичный барьер, одновременно защищающий клетки от окружающей среды и предоставляющий возможности взаимодействовать с ней [22-25]. Углеводы клеточной стенки бактерий отличаются структурным разнообразием моносахаридов и типов связей между ними [26], что делает их основными детерминантами вирулентности бактериальных штаммов и серологи-

ческой специфичности иммунного ответа организма-хозяина. Рецепторы клеток организма-хозяина взаимодействуют с углеводными антигенами патогенных микроорганизмов. Уникальность этих антигенов закладывает основу для хемотаксономической классификации бактерий и деления видов на серогруппы. Углеводы бактерий и грибов, сами по себе являющиеся слабыми иммуногенами, в составе липополисахаридов или гликопротеинов способны запускать углевод-специфичный Т-клеточный иммунный ответ [14, 27, 28]. Чтобы избежать иммунного ответа, микроорганизмы часто модифицируют свои клеточные стенки таким образом, что структуры углеводов на их поверхности напоминают структуры гликопротеинов и гликолипидов на поверхности клеток организма-хозяина [29]. Клеточная стенка патогенных бактерий также является одной из основных мишеней антибиотиков [24]. Соответственно, выживание патогенных микроорганизмов напрямую зависит от их способности синтезировать определённые углеводные структуры. Информация об этих структурах и аппарате их биосинтеза, в частности, об углевод-активных ферментах (carbohydrate-active enzymes, CAZy), находит широкое применение в современной биологии, медицине и биотехнологии и востребована при создании гликоконъюгатных вакцин [30].

Углеводы растительного происхождения также представляют значительный интерес. Помимо структурных углеводов, участвующих в построении кле-

точной стенки, и углеводов, выполняющих энергетические функции, многие растения продуцируют биологически активные низкомолекулярные соединения, которые содержат олигосахаридные фрагменты [31, 32].

Систематизация накопленных данных по структурам углеводов бактерий, грибов и растений, а также по ферментам, вовлечённым в их биосинтез (гликозилтрансферазы и др.), имеет первостепенное значение для многих направлений фундаментальной и прикладной науки, в частности, для медицинских скрининговых исследований и внедрения новых ферментативных реакций в биотехнологическое производство [33]. Однако информатизация гликомики значительно отстаёт от геномики и протеомики (Рис. 3). Это выражается в нехватке либо отсутствии общепринятых стандартов и моделей данных, протоколов обмена информацией, подходящих хранилищ данных и способов их визуализации, в информационной изолированности существующих проектов как друг от друга, так и от крупнейших проектов геномики, протеомики и медицины: Genbank [34][a], Uniprot [35][b], Pubmed [36][c], ICD [37][d] и других [38]. Ведущий разработчик биохимических баз данных (NCBI[e], NIH), продукты которого *de facto* стали стандартом в вопросах каталогизации объектов во всей биоорганической и медицинской химии, не поддерживает и не курирует углеводные базы.

Существующие способы компьютерного описания углеводов исторически разрабатывались для эффективного использования данных об углеводных частях гликопротеинов млекопитающих и не учитывали особенности углеводов из других доменов. С точки зрения наличия и доступности данных углеводы бактерий, архей и грибов представлены в базах значительно хуже, чем углеводы животных, в частности, по причине высокого химического разнообразия структур [39] и трудностей с их компьютерным описанием, а растительные гликозиды в существующих базах данных практически отсутствуют [40].

[a] https://www.ncbi.nlm.nih.gov/genbank/

[b] https://www.uniprot.org/help/about

[c] https://www.ncbi.nlm.nih.gov/pubmed/

[d] https://icd.who.int/browse11

[e] https://www.ncbi.nlm.nih.gov/

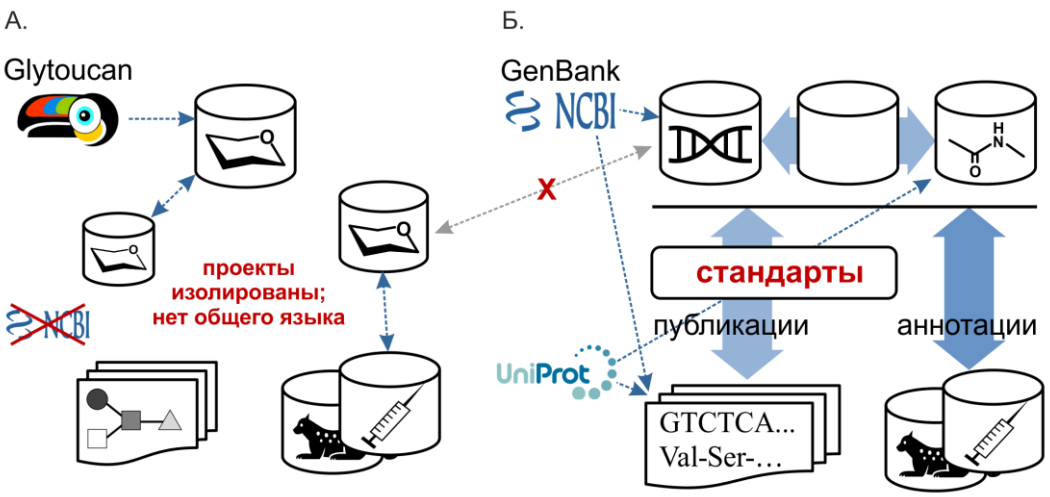



**Рис. 3**. Информатизация гликомики (А) в сравнении с геномикой и протеомикой (Б).

Также следует отметить проблему углевод-активных ферментов, которые вовлечены в синтез, сборку и процессинг углеводсодержащих структур у всех организмов. Использование ферментативных реакций в синтезе углеводных биологически активных продуктов и гликоконъюгатных вакцин – это перспективный способов решения проблем их химического синтеза либо выделения сложных углеводных структур из природных источников. Однако, несмотря на то, что множество известных генов бактерий, грибов и растений кодирует потенциальные гликозилтрансферазы, биохимическая характеризация их белковых продуктов заметно отстаёт от генетической, в том числе по причине отсутствия удобного хранилища накопленных данных по ферментативным активностям [31]. Так, в наиболее полной на данный момент базе данных углевод-активных ферментов CAZy активность подтверждена экспериментально менее, чем для 5% ферментов [41].

Многие существующие проекты гликомики отличает низкое качество данных, что обусловлено отсутствием формализации библиографических, таксономических, медицинских и других аннотаций структур, а также эффективных алгоритмов выявления и исправления ошибок [40, 42]. Первоочередная важность стандартизации углеводных данных была осознана в полной мере относительно недавно, благодаря росту популярности автоматической обработки данных в машиночитаемых форматах для поиска закономерностей и корреляций «структура - свойство» путём перебора и сравнения. Такая стандартизация позволяет связывать «изолированные острова» данных и получать разнотипные знания,

распределённые по нескольким базам. В этом направлении предпринят ряд шагов, однако ни один из существующих проектов гликоинформатики не поддерживает всех стандартов даже в тех областях, где стандарты уже выработаны. Способ стандартизации с помощью модели Resource Description Framework [43] и формальной углеводной онтологии является одним из наиболее перспективных с точки зрения интеграции существующих гликоресурсов [44].

Осознание учёными вышеперечисленных проблем привело к появлению в XXI веке новой научной дисциплины – гликоинформатики, среди основных задач которой можно назвать [40]:

- обеспечение учёных средствами доступа к накопленным данным по природным углеводам и средствами их статистической обработки;
- моделирование различных физико-химических, химических и биологических свойств углеводов;
- предсказание углеводных структур на основании их свойств;
- формализацию экспериментальных протоколов и систематизацию экспериментальных данных об углеводах.

Краеугольным камнем гликоинформатики является качество данных. Аннотирование уже опубликованных данных требует понимания контекста публикаций и, следовательно, крайне плохо автоматизируется. Неточности аннотирования представляют основной источник ошибок в базах данных, для выявления которых требуется привлекать квалифицированных специалистов в химии и биологии углеводов, что подразумевает высокие финансовые и временные затраты. По этой причине в биоинформатике в целом и в гликоинформатике в частности редко прибегают к экспертной проверке [40, 42].

Компьютерное представление углеводных данных, особенно первичных структур («углеводная нотация») кардинально отличается от традиционного, и при этом критически важно для распространения и использования баз данных. К настоящему времени разработано несколько нотаций, но в научной среде не достигнут консенсус мнений о том, какая из них наилучшим образом выполняет свои задачи. В этом контексте предприняты несколько шагов по разработке углеводных форматов [45-47] и онтологий [44], а также универсального идентифи-

катора углеводных структур (сходного с Genbank ID в геномике) [48]. Эти стандарты далеки от совершенства и поддерживаются не всеми современными проектами гликоинформатики. Стандартизация углеводных данных обязательно должна включать однозначную и функционально полную номенклатуру углеводов, поддающуюся как человеческой, так и компьютерной интерпретации [40].

Кроме российской группы CSDB, возглавляемой автором диссертации, информатизацией гликобиологии занимается несколько групп в США, Японии, Германии и других странах. Часть из них имеет собственные базы данных и языки описания, вокруг которых строятся остальные разработки. Конкретные проекты [20, 33, 49-52] и их сравнение с CSDB [53] описаны в обзорах, а генеральные линии развития области – в эссе автора диссертации [40]. Для основных направлений диссертации можно назвать следующие группы, занимающиеся схожими исследованиями (указаны область конкуренции, названия проектов, руководитель, преимущества и недостатки):

- Платформа гликоинформатики: GlycoSCIENCES.de (Томас Люттеке, Андреас Бонн, Германия) [54]. Структуры не курируются, нет адекватного инструментария работы со спектральными данными, есть инструментарий по работе с геометрией, есть формализация в RDF.
- Платформа гликоинформатики: Японский консорциум по гликобиологии и гликотехнологии, Japan Consortium for Glycobiology and Glycotechnology, JCGG (Кийоко Аоки-Киношита, Япония) [55]. Структуры не курируются, нет аналитических инструментов, есть репозиторий структур и другие близкие по тематике базы, есть формализация в RDF, нет собственного источника данных о структурах (данные импортируются из других баз).
- Структуры углеводов бактерий, архей, грибов, растений, простейших – значимых конкурентов нет.
- Структуры углеводов млекопитающих: UnicarbKB (Ники Пэкер, Мэттью Кэмпбелл, Австралия) [56]. Коммерческая, курируемая, неполная, ориентирована на O- и N-гликаны гликопротеинов, есть формализация в RDF.
- Структуры углеводов млекопитающих: GlycoSCIENCES.de (Томас Люттеке, Андреас Бонн, Германия) [54]. Некурируемая, неполная.

- Данные по гликозилтрансферазам: CAZy (Бернард Энриса, Франция) [41]. Огромный охват; большая часть активностей ГТ не подтверждена экспериментально, ограниченная функциональность открытой части базы, не хранятся данные о полных синтезируемых структурах, нет ссылок на оригинальные публикации.
- Данные по гликозилтрансферазам: ECODAB (Горан Видмальм, Швеция) [57]. В основном, *Escherichia coli*. Существуют и другие базы по отдельным организмам; эта приведена как наиболее востребованная.
- Теоретическая гликоинформатика, стандартизация – Центр исследования сложных углеводов, Complex Carbohydrate Research Center, CCRC (Вильям Йорк, Рене Ранцингер, США) [44]. Модельные исследования, разработка онтологии GlycoRDF.
- Стандартизация (отчасти) – Консорциум по функциональной гликомике, Consortium for Functional Glycomics, CFG (Ричард Каммингс, США). Генеральная линия деятельности и квалификация участников позволяют предположить разработку стандартов в области гликоинформатики.
- Симуляция данных ЯМР и предсказание структуры по спектрам: CASPER (Горан Видмальм, Швеция) [58]. Одномерные спектры и HSQC, ограниченная функциональность инструмента работы со структурами.
- Работа с молекулярной геометрией углеводов: GLYCAM (Роберт Вудс, США) [59]. Опирается на собственное силовое поле; ограничено небольшим числом остатков, характерных для гликанов млекопитающих.
- Работа с молекулярной геометрией углеводов: GlycoMaps / Sweet-II (Томас Люттеке и Мартин Франк, Германия) [60]. Используют информацию о предпочтительных торсионных углах в димерных фрагментах. Оптимизировано и валидировано для структурных компонентов, характерных для гликанов млекопитающих.
- Углеводная фенетика и гликотаксономия[a] – значимых конкурентов нет.

[a] Кластеризация и систематизация таксонов по их гликомам.

### *2.2. Информационные ресурсы в гликохимии и гликобиологии*

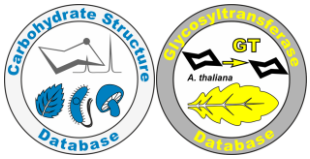
Гликоинформатика - это молодая и активно развивающаяся область знания, поэтому, несмотря на отдельные существующие базы данных и компьютерные инструменты, до сих пор не создано универсальной платформы, которая объединила бы данные по природным углеводам различного происхождения и аппарату их биосинтеза. Особенно заметно отставание гликоинформатики от других направлений биоинформатики в сфере автоматизации сервисов, позволяющей проектам и учёным обмениваться данными.

К настоящему времени накоплены огромные массивы данных по структурам и свойствам природных углеводов. Хранение и систематизация этих данных требует привлечения компьютерных ресурсов, поэтому основой гликоинформатики можно назвать специализированные электронные базы данных (БД). Наиболее значимые углеводные базы представлены в Табл. 1, а их взаимосвязи и основные особенности - на Рис. 4. Большинство из них в том или ином виде использует данные из БД Complex Carbohydrate Structure Database (CCSD, CarbBank) [61, 62], которая содержала около 23000 природных сахаридов и гликоконъюгатов, включавших более двух углеводных остатков. Финансирование этого исторически значимого проекта прекратилось в 1996 году; причины рассмотрены ниже.

За исключением немногочисленных узкоспециализированных БД (например, ECODAB [57], посвящённой *E. coli*, или новой базы экзополисахаридов EPS-DB [63]), большинство БД содержат гликаны млекопитающих, и лишь немногие включают информацию об углеводах организмов других доменов, в основном импортированную из CCSD. Самой полной на текущий момент базой данных по углеводам прокариот является база CSDB (Carbohydrate Structure Database, http://csdb.glycoscience.ru), разработанная под руководством автора диссертации [18, 64]. Помимо платформы CSDB, среди важных проектов гликоинформатики, связанных с базами данных, можно назвать: портал GLYCOSCIENCES.de (включает данные CCSD, а также данные ЯМР, теоретические и экспериментальные трёхмерные структуры, аналитические инструменты) [54]; UnicarbKB (включает углеводные структуры эукариот и углеводные части структур из CCSD, экспериментальные данные ВЭЖХ, МС и ЯМР) [56, 65, 66];

CFG (содержит углеводные структуры млекопитающих из CCSD и аннотированные структуры из базы Glycominds Ltd.) [67]; комплект мета-баз JCGG (объединяет базы данных гликопротеинов, аналитических данных и ассоциированных с гликомом заболеваний) [55]; KEGG Glycan (содержит структуры гликанов и биомедицинские и другие данные из Киотской энциклопедии генов и геномов (Kyoto Encyclopedia of Genes and Genomes)) [68]; вторичную базу GlyTouCan (представляет собой хранилище структур гликанов, разработанное для присвоения каждой углеводной структуре уникального идентификационного номера) [69] и некоторые другие [52, 70, 71]. Из прекращённых проектов, оказавших влияние на современные БД, помимо CarbBank следует отметить GlycomeDB (вторичная мета-база) [72] и EUROCarbDB (модельное исследование) [56, 73]. Полнота покрытия базы хотя бы в рамках определённого домена или класса соединений (а в идеале – по всем природным углеводам) – один из важнейших факторов, определяющий научную ценность базы, так как при полном покрытии даже отрицательный ответ на пользовательский поисковый запрос является значимой информацией, так как свидетельствует о том, что искомые данные не были опубликованы в научной литературе, и следовательно, обладают новизной. Ни одна из перечисленных БД не обеспечивает полного покрытия по природным углеводам, и лишь немногие являются курируемыми, причём проверка данных в них проводится для новых записей, но не для записей, импортированных из CCSD (исключение составляет полностью курируемая база CSDB). При этом CCSD содержит ошибки в 35% записей, и эти ошибки многократно воспроизводятся, переходя из одной базы в другую [42]. Экспертное курирование применяется лишь в нескольких биохимических базах данных (включая CSDB). Оно является необходимым этапом при импорте данных или ином расширении покрытия любой БД. Более подробно анализ существующих углеводных баз данных и инструментов анализа и моделирования сахаридов представлен в обзорах [20, 33, 49-52].

Табл. 1. Углеводные базы данных

| *База данных* | *Покрытие: основные данные (в скобках - приблизительное количество записей)* | *Годы существования и примечания* | *Литература и веб-адрес* |
|---|---|---|---|
| *1. Структурно-центрические* | | | |
| Carbohydrate Structure Database (CSDB)  (*проект автора диссертации*) | прокариоты, растения, грибы: **структуры*** (20000), таксономия (7000), библиография (9000), ЯМР (9000), ферменты (2000), 3D-структуры | 2005 - н.в. полное покрытие до 2017 по прокариотам, до 2010 по грибам | [18, 64, 74, 75] [a] |
| Complex Carbohydrate Structure Database, CarbBank | все природные углеводы: **структуры** (23000), таксономия, библиография | 1989-1996 полное покрытие до 1996 | [61, 62] |
| CFG Glycan Structure Database  | млекопитающие: структуры (>6000), таксономия, библиография, углеводные микрочипы | 2001-2011, основана на базе CCSD и Glycominds Ltd. | [67] [b] |
| KEGG Glycan  | млекопитающие: структуры (11000), библиография, ссылки | 2005 – н.в. | [68, 76] [c] |
| Glycosciences.DE  | млекопитающие: структуры (25000), таксономия, библиография (19000), ЯМР, 3D-структуры, ссылки | 1997 – н.в. | [54, 77, 78] [d] |

[a] http://csdb.glycoscience.ru/database

[b] http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp

[c] http://www.genome.jp/kegg/glycan/

[d] http://www.glycosciences.de/database/

| | | | |
|---|---|---|---|
| EurocarbDB  | модель: структуры (14000), таксономия, библиография, ЯМР (1300), МС | 2005-2010 модельное исследование, данные CCSD | [56, 73] [a] |
| GlycoSuite  | млекопитающие: **структуры** (10000), таксономия, библиография (1000), медицинские данные, сайты гликозилирования | 2001-2005, полное покрытие до 2005 | [65, 66] |
| GlycoBase/GlycoStore  | *O*- и *N*-гликаны: структуры (700), таксономия, библиография, ЖХ, ВЭЖХ, МС | 2008 - н.в. (Dublin) | [79, 80] [b] |
| GlycoBase  | животные: **структуры** (300), таксономия, ЯМР | 2007 - н.в. (Lille) | [c] |
| UniCarbKB  | млекопитающие: структуры (4000), таксономия, библиография (1000), ЖХ, МС | 2005 - н.в. основана на GlycoSuite и EurocarbDB | [56, 65, 66] [d] |
| Glycoconjugate  | гликоконъюгаты, все виды: структуры (44000), ссылки | 2008-2013 | [81] |
| GlycosideDB  | гликоконъюгаты и агликоны: структуры (>70000) | 2007 - н.в., в составе JCGGDB | [55] |

[a] http://relax.organ.su.se/eurocarb/

[b] http://glycobase.nibrt.ie/

[c] http://glycobase.univ-lille1.fr/base/

[d] http://www.unicarbkb.org

| | | | |
|---|---|---|---|
| Glycan Mass Spectrum  | *O*- и *N*-гликаны: структуры (3000), МС (3000) | 2006 - н.в., в составе JCGGDB | [55, 82] [a] |
| SugaBase | все виды: структуры, таксономия, ЯМР (500) | вошла в Glycosciences.DE | [83] |
| GlycomeDB  | все виды: структуры (43000), таксономия, ссылки | 2008-2016 | [72, 84, 85] [b] |
| *E. coli* O-antigen  | *E. coli*: **структуры** (300), таксономия, ЯМР, ферменты (300) | 2006 - н.в. | [57, 86] [c] |
| EPS-DB | бактерии: **структуры** (105) и функции экзополисахаридов и капсульных полисахаридов, таксономия, ссылки | 2018 - н.в. | [63] [d] |
| Гликозилпротеомные | | | |
| CAZy  | все виды: ферменты (340000), таксономия | 1999 – н.в. подтверждено 4% активностей | [41, 87] [e] |
| BRENDA  | все виды: ферменты (80000), таксономия | 1987 – н.в. | [88, 89] [f] |
| CFG GT | млекопитающие: ферменты, таксономия | | [g] |
| Glycogene  | человек и *C. elegans*: гены, ферменты (200) | 2004 – н.в., в составе JCGG | [90, 91] [a] |

[a] https://jcggdb.jp/rcmg/glycodb/Ms_ResultSearch

[b] http://www.glycome-db.org/

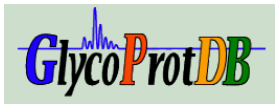
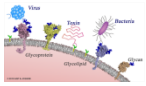
[c] http://www.casper.organ.su.se/ECODAB/

[d] http://www.epsdatabase.com

[e] http://www.cazy.org/

[f] https://www.brenda-enzymes.org/

[g] http://www.functionalglycomics.org/glycomics/molecule/jsp/glycoEnzyme/geMolecule.jsp

| | | | |
|---|---|---|---|
| GlycoEpitope  | гликоэпитопы (200), антитела (600) | 2006 – н.в. | [92, 93] [b] |
| KEGG Pathway | пути биосинтеза, ферменты, ссылки | 2006 – н.в. | [94, 95] [c] |
| Lectin Frontier  | углеводные микрочипы | в составе JCGG | [55, 96] [d] |
| GlycoProtDB  | *N*-гликопротеины мыши и *C. elegans*: структуры (3000), таксономия | 2011 – н.в.<br>в составе JCGG | [97] [e] |
| O-glycBase  | *O*- и *C*-гликопротеины (200) | 1996 - 2002 | [98] [f] |
| Pathogen Adherence to Carbohydrates  | медицинские данные, сайты связывания, библиография | 2010 – н.в.,<br>в составе JCGG | [55] [g] |
| SugarBind  | структуры (1000), таксономия, адгезия к патогенам | 2010 - н.в. | [99, 100] [h] |
| Специализированные | | | |
| MonosaccharideDB  | моносахариды (800) | 1997 - н.в. | [101, 102] [i] |
| GlycoMaps | 3D-структуры (3000) | | [60] [a] |

[a] https://acgg.asia/ggdb2/

[b] https://www.glycoepitope.jp/

[c] http://www.genome.jp/kegg/pathway.html

[d] https://acgg.asia/lfdb2/index

[e] https://acgg.asia/gpdb2/

[f] http://www.cbs.dtu.dk/databases/OGLYCBASE/

[g] http://jcggdb.jp/search/PACDB.cgi

[h] https://sugarbind.expasy.org/

[i] http://www.monosaccharidedb.org/

| | | | |
|---|---|---|---|
| GlyTouCan  | все виды: структуры (99000), ссылки | 2016 - н.в. репозиторий идентификаторов | [69] [b] |
| GlycoNAVI  | структуры (4000), химические реакции углеводов (3000) | 2007 - н.в., в составе JCGG | [55] [c] |
| GlycoPOD  | методики синтеза и анализа углеводов (200) | 2009 - н.в., в составе JCGG | [55] [d] |

* Слово **структуры** выделено жирным шрифтом, если база предоставляет структуры в основном на основании собственных усилий по аннотированию.

Следует отметить, что большинство БД углеводных структур, разработанных за последние 25 лет, содержат данные, импортированные из других БД, которые были разработаны химиками без учёта требований к системному управлению данными. В результате эти исходные БД оказывались немасштабируемыми и, по достижении определённого объёма данных, необслуживаемыми (Рис. 5). Так, самая первая и крупнейшая углеводная база CCSD просуществовала всего шесть лет. Согласно опыту нескольких научных групп, работающих в области гликоинформатики, успешный гликоинформатический проект должен представлять собой реляционную базу данных, использующую для большинства данных стандартные идентификаторы (doi, TaxID, PMID и т.п.) и контролируемые словари (мономеров и пр.), но не свободный текст, и содержащую структурные, таксономические и библиографические данные в виде отдельных записей. Подобные правила были названы «Десятью заповедями гликоинформатики» и учтены в модельном исследовании объединённой европейской группы Eurocarb [73]. Также первостепенную значимость для качества данных имеет наличие человекочитаемого дампа, который позволяет привлекать к курированию учёных, не обладающих специальными компьютерными навыками. База данных должна

[a] http://www.glycosciences.de/modeling/glycomapsdb/

[b] https://glytoucan.org/

[c] http://ws.glyconavi.org/

[d] https://jcggdb.jp/GlycoPOD

автоматически отправлять запросы в другие БД для получения дополнительной информации, и сама быть способной отвечать на аналогичные запросы на общепризнанном формальном языке. Немаловажное значение также имеет наличие подробного описания БД и дружественного интерфейса пользователя. [40, 73].

Подавляющее большинство существующих баз данных по углевод-активным ферментам (CAZy) посвящены отдельным таксонам (Табл. 1). Среди исключений можно назвать БД CAZy [41] и BRENDA [88]. CAZy[a] на данный момент является самой большой БД последовательностей ферментов CAZy и со-

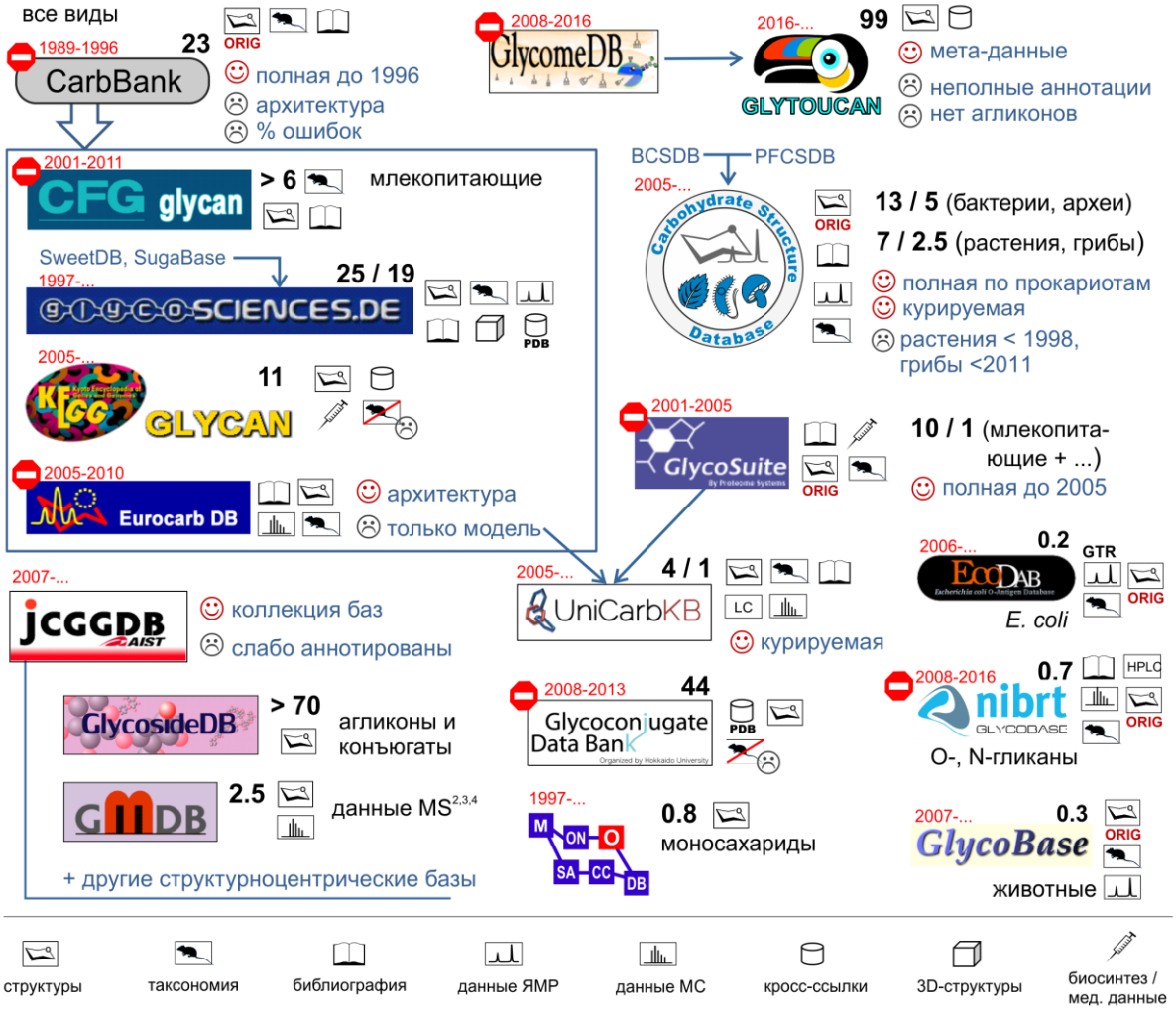



**Рис. 4**. Основные базы данных по углеводным структурам. Пиктограммы отражают типы сохраняемых данных, в соответствии с расшифровкой в нижнем ряду. Числа жирным шрифтом соответствуют числу структур / публикаций в тысячах. Важнейшие преимущества и недостатки перечислены синим. Схема включает несколько прекращённых, но исторически значимых проектов (обозначены символом ⛔).

[a] http://www.cazy.org

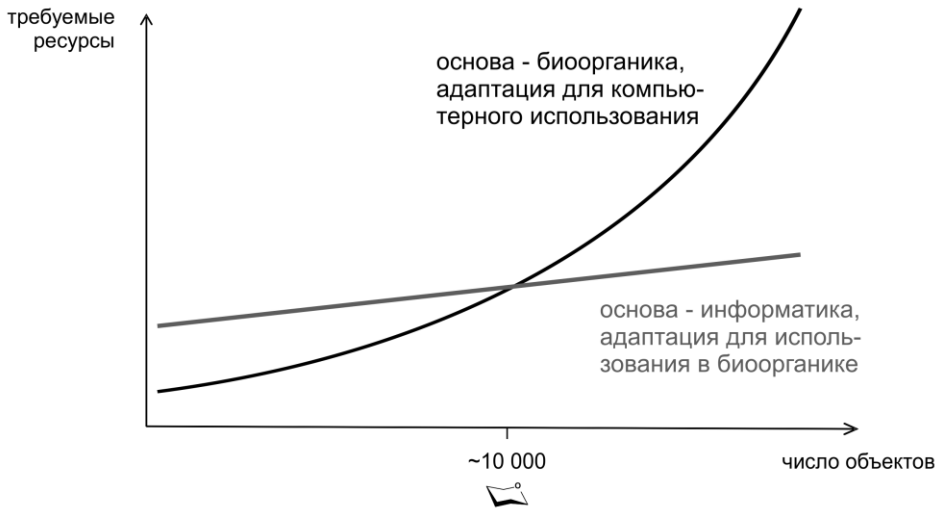



**Рис. 5**. Зависимость ресурсоёмкости проекта от количества структур в базе данных. Чёрный экспоненциальный график отражает подход, традиционный для химических баз предыдущего поколения. Серый линейный график отражает подход, основанный на правилах информатики.

держит информацию более чем о 340000 ферментов, активность 95% которых, однако, не охарактеризована экспериментально [41]. Кроме того, CAZy не содержит прямых ссылок на синтезируемые углеводные структуры и на оригинальные экспериментальные работы, поэтому на информацию из нее затруднительно корректно ссылаться в исследованиях. Некоторые неспециализированные БД также содержат информацию по ГТ отдельных организмов: Arabidopsis Information Portal (аннотированная геномная последовательность *Arabidopsis thaliana*) [103]; TAIR (The Arabidopsis Information Resource, хранилище различных данных по *A. thaliana*) [104]; PlantCyc (база данных по метаболическим путям нескольких видов растений; содержит информацию о некоторых ГТ, в том числе углеводные структуры и ссылки на литературу) [104, 105]; BRENDA (информационная система ферментов, включающая функциональные и структурные данные для ~80000 белков различных организмов) [88].

Из узкоспециализированных БД следует упомянуть ECODAB (О-антигены *E. coli*; содержит данные по ~600 предсказанным активностям гликозилтрансфераз (ГТ), 3% которых подтверждены экспериментально) [57]; Rice GT Database (база данных по ГТ риса, информация по предсказанным генам ГТ) [106]; GlycoGene Database (GGDB) и *Caenorhabditis elegans* GlycoGene Database (базы данных по генам белков аппарата синтеза углеводов у человека и *C. ele-*

*gans*) [90, 91]; базы ГТ Консорциума по функциональной гликомике (данные по ГТ, участвующим в биосинтезе некоторых гликанов человека и мыши).

За последние годы различными исследовательскими группами было разработано множество инструментов для облегчения интерпретации экспериментальных данных по природным углеводам. Швейцарский институт биоинформатики курирует коллекцию, объединяющую углеводные инструменты, доступные через Интернет неподготовленным пользователям (Glycomics@Expasy[a] [107]). В первую очередь компьютерные программы анализа данных оперируют спектрами ЯМР, МС и ВЭЖХ, которые наиболее часто используются для установления и подтверждения структуры гликанов [33]. Несмотря на то, что автоматическая интерпретация спектров ЯМР и многомерных масс-спектров остаётся трудной задачей, это направление активно развивается. На текущий момент больше всего инструментов разработано для работы с масс-спектрами углеводов. Основные инструменты компьютерного анализа гликанов приведены в Табл. 2 и подробно рассмотрены в обзорах [18, 20, 33, 51, 52].

Табл. 2. Компьютерные инструменты анализа углеводных экспериментальных данных

| *Название* | *Описание* | *Ссылка* |
|---|---|---|
| GlycoWorkbench | Отнесение масс-спектров гликанов | [108] |
| Glyco-Peakfinder | Определение состава гликанов по их масс-спектрам | [109] |
| GlycoPep ID | Идентификация пептидной составляющей гликопептидов с использованием неспецифичных ферментов | [110] |
| GlycoPeptideSearch | Масс-фингерпринтинг* (МС/МС) *N*-гликопептидов с использованием гликанов GlycomeDB | [111] |
| GlycoMiner | Анализ состава гликопептидов | [112] |
| GlycoMod | Предсказание структуры олигосахаридных частей гликопротеинов на основании масс- | [113] |

[a] https://www.expasy.org/glycomics

| | | |
|---|---|---|
| | спектров | |
| AutoGU | Анализ ВЭЖХ | [79] |
| GALAXY | Визуализация двумерных карт ВЭЖХ | [114] |
| ProspectND | Обработка спектров ЯМР | [73] |
| CCPN Tools | Отнесение спектров ЯМР | [115] |
| CASPER | Установление структуры олигосахаридов и регулярных полисахаридов по спектрам ЯМР | [58] |
| BIOPSEL | Установление структуры гликополимеров по спектрам ЯМР | [116] |
| GlyNest | Предсказание спектров ЯМР углеводов | [58] |
| CSDB NMR Tools | Набор инструментов для симуляции спектров ЯМР и предсказания структуры углеводов по спектрам ЯМР (см. Разд. 3.4.1 и 3.4.2) | [18, 117-120] |

* Потоковое сличение данных с эталоном (от fingerprint – отпечаток пальца).

### *2.3. Описание, идентификация и визуализация структур*

Стандартизация способов представления информации об углеводах, в первую очередь об их первичной структуре, крайне важна для упрощения использования данных, хранящихся в базах. Отсутствие общепринятых форматов и стандартов описания структур значительно затрудняет обмен данными между существующими углеводными БД. Углеводы – самый химический разнообразный класс биомакромолекул, демонстрирующий множество мономерных остатков и их конфигураций, химических модификаций, наличие точек ветвления и другие признаки вариативности первичной стрктуры [39]. Это разнообразие осложняет компьютерное описание углеводных последовательностей, делает существующие наработки из других наук о жизни малопригодными и требует особого подхода к разработке языка описания биогликанов. Согласно опыту, накопленному группой автора диссертации и другими научными коллективами, этот язык должен удовлетворять следующим критериям [64]:

- Способность описать все особенности, которые могут присутствовать в углеводных молекулах, включая неопределённые и не полностью установленные структурные компоненты. Все существующие на данный момент языки подходят для описания распространённых структур и различаются по числу особых случаев, с которыми они могут справиться. Полнота языка (число различных структурных особенностей гликанов, которые может описать язык) часто ограничена словарём мономерных остатков.
- Описание гликана и его структура должны взаимно-однозначно соответствовать друг другу, в том числе в аспекте сортировки цепей, старшинства и порядка следования заместителей. Для каждой структры, описание должно быть единственным, недвусмысленным и машиночитаемым. Языки на основе номенклатуры IUPAC не удовлетворяют этому критерию.
- Поскольку все исходные данные и все базы данных содержат ошибки, описание структуры должно быть человекочитаемым, что позволит избежать проблем с отслеживанием ошибок, которые в противном случае очень трудно обнаружить и исправить. Языки, основанные на таблицах

связности, обычно не удовлетворяют этому критерию и требуют разработки специальных процедур визуализации и проверки.

Таким образом, важнейшим направлением стандартизации является разработка однозначной и функционально полной номенклатуры углеводов, понятной как человеку, так и компьютеру. Среди основных претендентов на эту роль можно выделить языки GlycoCT [46], WURCS [47, 121] и разработанный в группе автора диссертации CSDB Linear [64, 70]. Каждый из них имеет свои преимущества и недостатки. В настоящий момент из-за неполной поддержки углеводно-языковых стандартов проектами гликоинформатики, а также из-за несовершенства самих стандартов, пользователи не имеют возможности получать неявно связанные данные из нескольких баз [44, 52].

Кроме универсальных языков кодирования химической структуры, для углеводов были разработаны специальные схемы кодировки (нотации), упрощающие работу со структурой с помощью компьютерных программ. Сравнение этих нотаций, включая разработанные в рамках диссертации SNFG и CSDB Linear, приведено в Табл. 3. Поддержка важнейших свойств применительно к углеводам качественно обозначена четырьмя градациями: - (отсутствует), -/+ (слабая), +/- (удовлетворительная), + (полная). *Курсивом* показаны языки, поддержка которых в углеводных базах данных в настоящее время прекращена.

Табл. 3. Сравнение существующих общехимических и углеводных нотаций.

| *Группа* | *Язык [лит. ссылка]* | *Проект* | *Подход* | *Свойства* | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | *Полнота* | *Однозначность* | *Человекочитаемость* | *Машиночитаемость* | *Поддержка неоднозначностей* |
| Общехимические | *IUPAC* [122] | повсеместно в статьях | свободный текст* | –/+ | –/+ | +/– | – | –/+ |
| | IUPAC extended [123] | CarbBank; SweetDB | псевдографика | –/+ | –/+ | + | – | –/+ |
| | MOL [124] | повсеместно в хемоинформатике | атомарный | + | + | – | + | – |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SMILES [125] | химические редакторы и БД органических молекул | атомарный* | + | –/+ | – | + | –/+ |
| | InChi [126] | | атомарный* | +/– | + | – | + | –/+ |
| Специализированные | *Glyde* [127] | - | граф*,** | +/– | +/– | – | + | –/+ |
| | *CabosML* [128] | JCGGDB | граф** | | +/– | – | + | – |
| | Linear Code [129] | CFG | граф*,** | – | +/– | –/+ | + | +/– |
| | KCF [76] | KEGG Glycan, RINGS | таблица связности | –/+ | + | – | + | –/+ |
| | LinUCS [130] | Glyco-sciences.DE | граф*,** | – | + | +/– | + | +/– |
| | GLYCAM [59] | GLYCAM | граф*,** | – | + | + | + | – |
| | GlycoCT [46] | GlycomeDB, EurocarbDB, GlyTouCan и др. | таблица связности | +/– | + | – | + | +/– |
| | Glyde II [131] | EurocarbDB, GlyTouCan, GlycoWorkbench и др. | таблица связности | расширение GlycoCT и партономия | | | | |
| | WURCS [47, 121] | JCGGDB; ChEBI; PDB | таблица связности* | +/– | + | – | + | +/– |
| | UOXF [132] | Oxford Glycobiology Institute | графическая визуализация | – | + | + | – | – |
| | CFG (v.2) [133] SNFG (v.3) [134] | повсеместно в статьях | графическая визуализация | +/– | + | + | – | + |
| | CSDB Linear [64, 70] | CSDB | граф*,** | +/– | + | +/– | + | +/– |

* Может быть записан в виде одной строки и передан в составе URL.

** Во всех приведённых примерах граф имеет тип «направленное дерево». Это исторически сложившееся представление интуитивно сопоставимо с углеводными структурами [135].

Другие особенности кодировки углеводов в разных нотациях систематизированы в обзорах немецкого коллектива гликоинформатиков [135, 136].

Ещё одним важным вопросом является обязательное присвоение каждой опубликованной углеводной структуре уникального идентификационного номера, аналогично зарегистрированным последовательностям аминокислот или нуклеотидов (Genbank ID, Uniprot ID). Использование таких идентификаторов значительно упростит взаимную интеграцию существующих проектов гликоинформатики. Использовать идентификаторы из какой-либо существующей углеводной БД не представляется рациональным по причине отсутствия полных БД по всем природным углеводным последовательностям и сложностей, связанных с проверкой и исправлением ошибок при добавлении в базу информации, не подтверждённой опубликованным источником. Поэтому специально для этой цели был разработан репозиторий GlyTouCan, который в настоящий момент содержит углеводные части природных гликанов [48]. В отличие от базы данных, репозиторий только предоставляет связку «идентификатор - структура», не постулируя валидность структуры или наличие литературной ссылки. Это снимает проблему качества данных, но требует отслеживания соответствий между идентификаторами GlyTouCan и идентификаторами записей в углеводных БД. Для успешной реализации этой инициативы все крупные научные издательства должны требовать помещения каждой новой углеводной структуры в соответствующее хранилище и присвоения ей идентификатора при отправке статьи на рассмотрение к публикации [40].

Классификация знаний и взаимосвязи между ними в любой предметной области задаются онтологией. В то время как фактическая онтология – усреднённая картина мира, существующая в головах учёных – не приспособлена к автоматическому анализу, формальная онтология позволяет систематизировать оцифрованные знания и задаёт идеологию любых создаваемых в области проектов, позволяющую легко устанавливать связи между ними. К настоящему моменту описано несколько формальных углеводных онтологий (GlycO [137], GlycoRDF [44], GlycoCoO[a]). Наиболее универсальной из них представляется Gly-

[a] http://bioportal.bioontology.org/ontologies/GLYCOCOO

coRDF[a], которая позволяет формализовать накопленные данные об углеводах в модели Resource Description Framework (RDF) [43]. Из её фундаментальных недостатков следует отметить необходимость наличия внешнего репозитория RDF-триплетов, а из практических - отсутствие формализации биосинтеза углеводов и связи с ферментативным аппаратом и необходимость в конвертере структур в формат GlycoCT для их интеграции в доступный пользователю объем знаний.

Инструменты ввода [50, 138, 139] и визуализации [134, 140, 141] углеводных структур динамично развиваются, однако из-за прекращения поддержки Интернет-браузерами Java-апплетов в 2016 году наиболее популярный из них - GlycanBuilder - требует замены. GlycanBuilder [138] был разработан в составе проекта EurocarbDB и поддерживал основные символьные нотации гликанов, а также предоставлял возможность текстового ввода с поддержкой различных форматов [139]. Как и другие существующие инструменты, GlycanBuilder плохо приспособлен к разнообразию структур, характерному для биогликанов прокариот.

Проблема визуализации углеводных структур в публикациях и компьютерных интерфейсах чрезвычайно важна с точки зрения взаимопонимания между авторами из разных областей науки. Наряду с представлением углеводных структур в виде последовательностей текста (condensed IUPAC [122]) и псевдографики (extended IUPAC[123], SweetDB [142]) большую популярность имеют пиктографические (символьные) способы визуализации. В 1978 г. Корнфельд с коллегами разработал простую систему представления гликанов позвоночных животных, которая впоследствии была стандартизована как «Нотация CFG» [143] и широко использовалась для описания структур и биосинтеза гликанов млекопитающих [144, 145]. Благодаря объединённым усилиям нескольких научных коллективов, включая коллектив автора диссертации, эта система легла в основу символьной номенклатуры гликанов (Symbol Nomenclature for Glycans, SNFG [134]), лучше приспособленной для описания структурных особенностей

[a] Онлайн-визуализация: http://www.visualdataweb.de/webvowl/#iri=http://data.bioontology.org/ontologies/GLYCORDF/submissions/56/download?apikey=8b5b7825-538d-40e0-9e9e-5ab9274a9aeb

гликанов в различных биологических системах (см. раздел 3.3.2). SNFG совместима с обозначением гликозидных связей, принятым в популярных системах графического представления структур (Рис. 6, нотация CFG [143], нотация Oxford [132]).

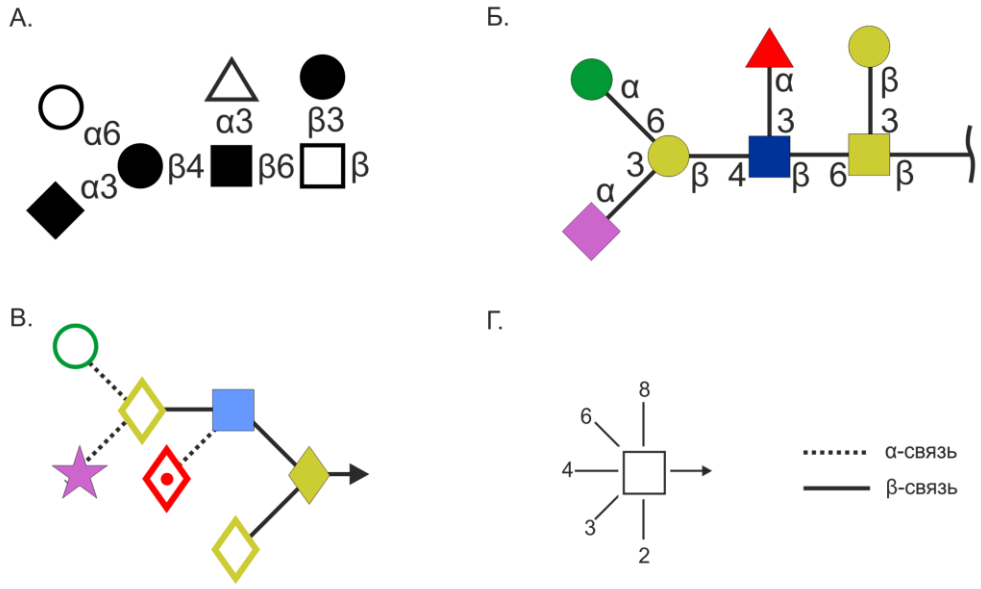



Рис. 6. Графическое представление структуры одного и того же гликана Neu5Ac-[Man-]Gal-[Fuc-]GlcNAc-[Gal-]GalNAc согласно схемам, принятым в различных нотациях. А. Нотации CFG, Б. Расширение CFG с использованием пиктограмм из SNFG; В. Нотация Oxford; Г. Принцип отображения положения (на примере глюкопиранозы) и конфигурации связи в нотации Oxford.

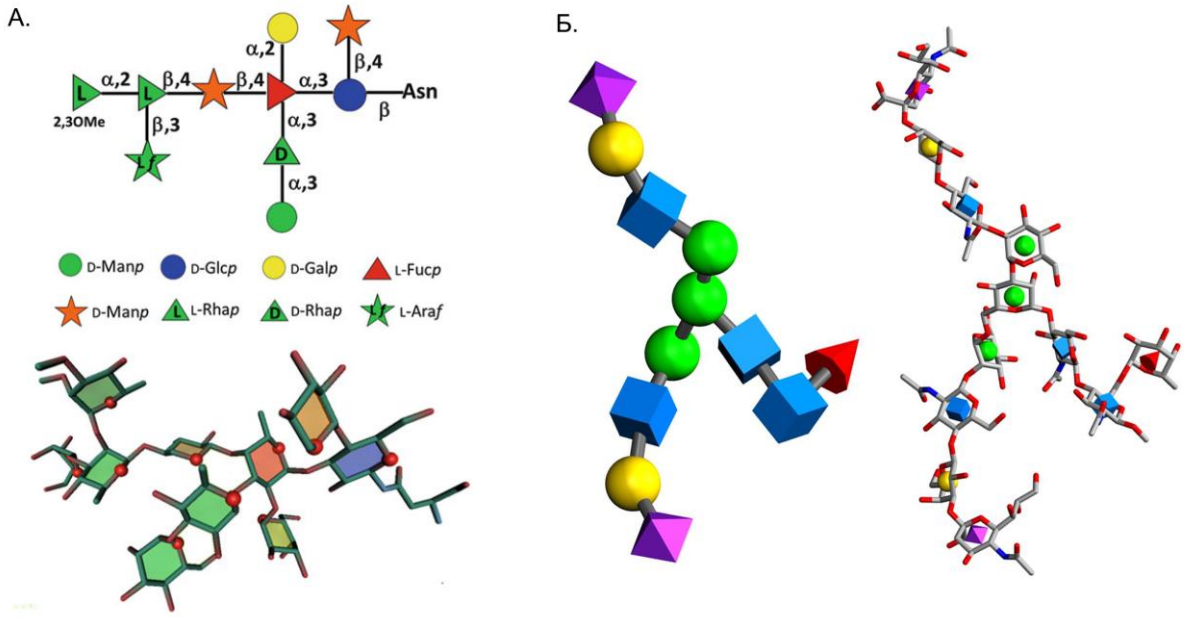



Рис. 7. Визуализация состава и связности гликанов в сочетании с трёхмерной структурой в SweetUnityMol (А) и в 3D-SNFG (Б).

В большинстве существующих углеводных БД используются семантические нотации, основанные на отдельных «строительных блоках» (остатках). Следует отметить непригодность такого описания для структурных, конформационных и энергетических расчётов, где требуется оперировать характеристиками отдельных атомов. Альтернативой является язык WURCS (Web3 Unique Representation of Carbohydrate Structures), который сочетает подходы, основанные как на остатках, так и на индивидуальных атомах [47]. Однако WURCS позволяет описать не все структурные особенности гликоконъюгатов и не поддерживается значимыми программными пакетами. Решением данной проблемы могут быть трансляторы с углеводных языков на общехимические. Одним из наиболее подходящих языков для перевода углеводных нотаций, основанных на остатках, в нотации, основанные на атомах, является SMILES, стандартный язык описания первичной структуры в органической химии [146]. Для получения трёхмерной визуализации не только первичной структуры, но и конформации биогликанов, они должны быть описаны на атомарном языке, включающем атомные координаты, например, MOL [124]. Однако координаты атомов и полученные из них молекулярные модели малопригодны для публикаций, так как сотни атомов сложно сопоставить с моносахаридными остатками и их взаимным расположением даже в интерактивном режиме, не говоря уже о статическом рисунке в статье. В последнее время с развитием SNFG появилось несколько инициатив, призванных решить эту проблему. Так, группа проф. Переза предложила ограничивать визуализируемую молекулу скелетными атомами и раскрашивать условную плоскость цикла в соответствии с цветовым кодом, принятым в SNFG (инструмент SweetUnityMol [147], Рис. 7А), а группа проф. Вудса предложила совмещать структурные элементы молекулы с трёхмерными пиктограммами SNFG (инструмент 3D-SNFG [148], реализованный на платформе Visual Molecular Dynamics [149], Рис. 7Б). Начиная с конца 2018 года, окраска остатков в трёхмерной структуре также поддерживается в Glycosciences.de [78].

### *2.4. Моделирование структуры углеводов*

Моделирование трёхмерной структуры углеводов, востребованных в экспериментах по молекулярному докингу и разработке новых лекарственных препаратов - перспективное направление гликомики. Сейчас такие структуры есть лишь в некоторых БД, таких как Белковый банк данных (Protein Data Bank, PDB[a] [150]) и для небольших молекул - Кембриджская структурная база данных (Cambridge Structural Database, CSD[b] [151]). Следует отметить, что углеводные структуры, содержащиеся в PDB, обычно ковалентно связаны с белком либо формируют с ним комплекс и часто представляют собой фрагменты исходных гликанов. Более того, это преимущественно гликопротеины млекопитающих, причём большинство данных получены на основании кристаллографических экспериментов (следовательно, могут отличаться от конформаций гликанов в растворе) или не подтверждены экспериментально. Также отмечалось, что PDB содержит большое количество ошибок [152]. В PDB нет удобного поиска гликанов [33, 51], однако поиск содержащихся в PDB трёхмерных углеводных структур возможен через портал GLYCOSCIENCES.de [153]. Опубликованы базы данных, извлекающие данные о молекулярной геометрии биогликанов из PDB: GlyTorsion [154] и GlycanFragmentDB [155]. Некоторые инструменты предсказания и анализа трёхмерных углеводных структур приведены в Табл. 4.

Табл. 4. Инструменты предсказания и анализа трёхмерной структуры углеводов и гликопротеинов.

| *Название* | *Описание* | *Ссылки* |
|---|---|---|
| BALLDock/SLICK | Докинг белков и углеводов | [156, 157] |
| CARP | Анализ двугранных углов гликозидных связей, сходный с картами Рамачандрана | [154] |
| CAT | Инструмент конформационного анализа молекулярно-динамических траекторий | [c] |
| GLYCAM | Генерация моделей гликанов и гликозилирование | [158] [a] |

[a] https://www.rcsb.org/

[b] https://www.ccdc.cam.ac.uk/

[c] http://www.md-simulations.de/CAT

| | | |
|---|---|---|
| Biomolecules Builder | белков *in silico*, подготовка входных файлов для AMBER | |
| Glycan Reader | Поиск углеводов в файлах PDB, подготовка входных файлов для CHARMM | [159, 160] |
| glyProt | Гликозилирование белков *in silico* | [161] |
| glyTorsion | Статистика двугранных углов в углеводных структурах в PDB | [154] |
| glyVicinity | Пространственно-близкие углеводам аминокислоты в PDB | [154] |
| pdb2linucs | Поиск углеводов в файлах PDB | [162] |
| pdb-care | Проверка файлов с трёхмерными углеводными структурами | [163] |
| Sweet-II | Предсказание трёхмерных углеводных структур | [164] |
| GlycoMapsDB | База данных рассчитанных конформационных карт олигосахаридов, присутствующих в *N*- и *O*-гликанах | [60] |

Предсказание геометрии сахаридов в автоматическом режиме (без специальной параметризации под конкретные остатки) поддерживается в проектах GLYCOSCIENCES.de [77], GLYCAM[b] [59] и CarbBuilder [165]. GLYCOSCIENCES.de использует упрощённый алгоритм Sweet-II [164] и данные для гликопротеинов из PDB [150], GLYCAM – молекулярно-механическое силовое поле собственной разработки [59], CarbBuilder – конформационные данные по заранее описанным наиболее распространённым остаткам. Sweet-II не использует данные о конформациях гликозидных мостиков, несмотря на то, что БД конформационных карт дисахаридов, заполненная расчётными данными для фрагментов, характерных для *O*- и *N*-гликанов млекопитающих (GlycoMapsDB [60]), разработана тем же коллективом, что и GLYCOSCIENCES.de. Все вышеописанные сервисы поддерживают ограниченное число структурных особенностей углеводов по причине жёстко заданного набора остатков, отсутствия под-

[a] http://glycam.org/

[b] http://www.glycam.org/cb

держки неуглеводных компонентов и требования, чтобы структуры не содержали атомов или фрагментов с неопределёнными конфигурациями.

В настоящее время для моделирования трёхмерных структур углеводов применяются следующие теоретические модели и методы [166]:

- молекулярная механика (MM) и молекулярная динамика (MD) в молекулярно-механических силовых полях;
- полуэмпирические методы;
- моделирование *ab initio* с применением теории функционала плотности;
- гибридные подходы QM/MM, QM/QM и ONIOM, в которых основная молекула обсчитывается методами молекулярной или квантовой механики на более низком уровне теории, чем группа атомов, для которых нужно получить геометрические параметры.

Методы MM используют классическую механику для моделирования молекулярных систем и вычисляют их потенциальную энергию на основе набора атомистических параметров («силовых полей»), полученных для небольших модельных соединений. Некоторые из этих силовых полей (CHARMM, GLYCAM) оптимизированы для углеводов [51]. MM3 является одним из наиболее популярных универсальных силовых полей для оптимизации структур олигосахаридов [166]. Однако для корректного описания гибкости и степеней свободы, присущих углеводным структурам, часто приходится прибегать к молекулярно-динамическим расчётам [167]. Обычно MM и MD используют одинаковые классические силовые поля, но, в отличие от MM, MD может прибегать к квантово-химическим уровням теории. Также для конформационных исследований углеводов привлекают молекулярно-динамические методы на основе обмена копиями (replica-exchange molecular dynamics, REMD) [168]. Основанные на MM и MD процедуры минимизации энергии широко используются в специализированных расчётных программных пакетах (Wavefunction Inc. Spartan [169], MOSCITO [170], COSMOS [171] и др.) и универсальных программах (Gaussian Inc. Gaussian [172], GAMESS [173] и др.).

Полуэмпирические методы привлекают наборы параметров, полученных на основе экспериментальных данных, для упрощения аппроксимации уравне-

ния Шрёдингера. Поэтому вычисления не требуют высоких затрат компьютерных мощностей и могут применяться для крупных молекул или для получения стартовой точки для последующих расчётов *ab initio*. Во многих случаях полуэмпирические методы плохо работают с молекулами, содержащими водородные связи, с переходными структурами и молекулами, содержащими атомы, для которых эти методы слабо параметризованы [166]. Для моделирования трёхмерной структуры углеводов применяют методы AM1, PM3 и MNDO, а также PM5 и PM6 [174, 175].

Квантовохимическое моделирование подразумевает сочетание теоретического подхода (уровня теории) с базисным набором. В последнее время все большую популярность в моделировании биомолекулярных систем приобретает теория функционала плотности (density functional theory, DFT [176, 177]), которая позволяет проводить расчёты с хорошей точностью при умеренных затратах компьютерных мощностей [166]. Обычно для структурных исследований углеводов применяют гибридные функционалы B3LYP и B3PW91 и их модификации [178, 179].

Гибридные подходы QM/MM, QM/QM и ONIOM (our own *n*-layer integrated molecular orbital and molecular mechanics approach) [180, 181] позволяют разбивать большие молекулярные системы на несколько подсистем (слоёв) и работать с ними на различных уровнях. При гибридных расчётах самая важная и маленькая часть молекулы (высший уровень) обсчитывается на более точных уровнях квантово-механической теории, в то время как остальные части молекулы - на менее затратных с точки зрения компьютерной мощности уровнях квантовой механики или с помощью молекулярной механики. Применение гибридных подходов позволяет ускорить расчёты и справиться с ограничением, накладываемым на размеры молекул. В идеале гибридный подход сочетает точность квантово-механических расчётов высшего уровня и скорость относительно быстрых низкоуровневых методов (ММ и других) [166].

При моделировании трёхмерных структур также нельзя забывать о способности углеводов, особенно полисахаридов, принимать различные динамические конформации в растворе, что во многом определяет их биологические функции. Поэтому при расчётах геометрии углеводных структур необходимо

учитывать их взаимодействие с растворителем, в качестве которого обычно выступает вода [166, 182]. В отличие от жёстких и неполярных молекул, углеводы склонны формировать с растворителем сильные водородные связи и обладают конформационными степенями свободы, что затрудняет моделирование динамического поведения углеводных структур в растворах. Одним из самых распространённых подходов к решению этой проблемы является применение молекулярно-динамических расчётов для растворенного вещества, окружённого молекулами растворителя, с последующим получением состояний («кадров») из файла с траекторией [166]. Также разработано несколько моделей растворителя для расчётов *ab initio* и DFT. Например, в модели поляризуемого континуума (polarizable continuum model, PCM) растворитель представлен в виде континуума, а в её модификациях DPCM и CPCM – в виде диэлектрика и проводника, соответственно [183, 184]. Также существуют модели, основанные на квантово-механической плотности заряда растворенного вещества и параметризованные для различных органических растворителей [185]. Модель COSMO (conductor-like screening model) рассчитывает поляризацию растворителя на основании распределения электрического заряда растворенного вещества и является более точной для растворителей с высокой диэлектрической проницаемостью, таких как вода [186].

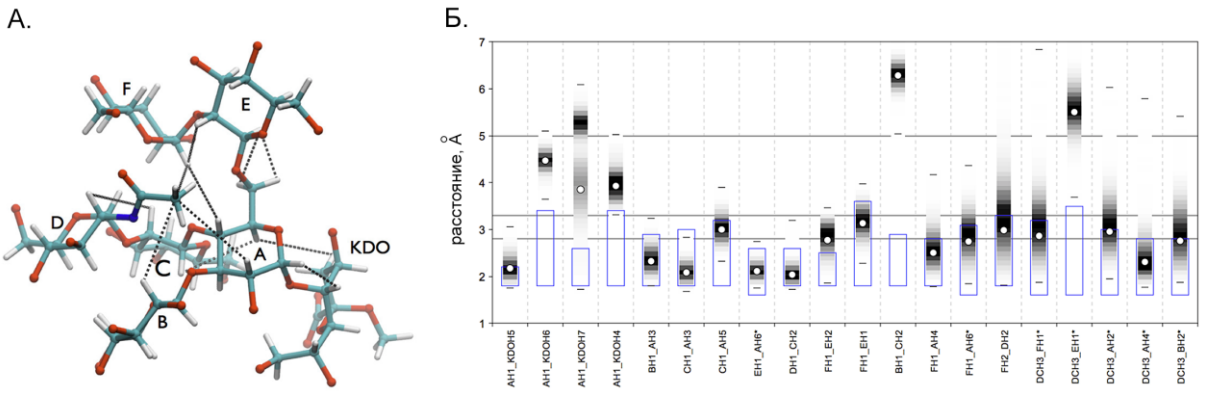



Рис. 8. А. Характеристические межатомные контакты в преимущественной конформации олигосахарида *Moraxella catarrhalis* lgt2Δ. Б. Отбор MD-состояний (оттенки серого, в соответствии с распределением; среднее значение показано белой точкой) на основании соответствия экспериментальному ЯЭО (синие прямоугольни). На горизонтальной оси обозначены наиболее заселённые конформации. Воспроизведено из [187].

Основными методами валидации методов моделирования геометрии являются РСА (для кристаллов) и сравнение усредненнных межатомных расстоний с экспериментально наблюдаемым ядерным эффектом Оверхаузера (ЯЭО) в растворе. Репрезентативным примером может являться работа Франка и соавторов по фильтрации конформаций, полученных молекулярной динамикой с явным учетом растворителя, на основании наблюдаемых ЯЭО [187]. Рис. 8 демонстрирует протон-протонные пространственныйе контакты и критерии отбра конформаций на примере гептасахарида lgt2Δ бактерии *Moraxella catarrhalis*, обладающего необычной конформацией.

### *2.5. Моделирование спектров ЯМР углеводов*

Спектроскопия ядерного магнитного резонанса (ЯМР) является одним из самых распространённых методов установления первичной структуры углеводов. Кроме того, она позволяет напрямую установить трёхмерную структуру в водном растворе, где протекают многие значимые биологические и химические процессы. Добиться этого можно посредством измерения ключевых экспериментальных параметров: химических сдвигов, констант спин-спинового взаимодействия, скоростей релаксации и величин ядерного эффекта Оверхаузера (ЯЭО, NOE). В настоящее время высокочувствительные одномерные и двумерные эксперименты ЯМР широко вошли в повседневную исследовательскую практику и применяются в биологических и химических лабораториях различного профиля [166, 182].

Однако интерпретация параметров ЯМР и установление взаимосвязи между ними и молекулярной структурой вещества по-прежнему остаётся сложной задачей, особенно в случае такого химически разнообразного класса биомакромолекул, как углеводы. В отличие от белков и нуклеиновых кислот, большая часть исследований углеводов ограничивается ядрами $^{1}H$ и $^{13}C$, а изотопное мечение, широко используемое в спектроскопии ЯМР белков, находит лишь ограниченное применение в ЯМР углеводов. Кроме того, несмотря на большое разнообразие углеводных строительных блоков, их химические сдвиги расположены в значительно более узкой области спектра ЯМР, по сравнению с белками и нуклеиновыми кислотами. В силу этого правильная интерпретация спектров ЯМР углеводов требует теоретического анализа, в частности, моделирования параметров ЯМР для полных структур или характеристичных структурных фрагментов [33, 166].

Исторически первый класс подходов к решению этой проблемы включал эмпирические методы, основанные на связности атомов или остатков в молекуле [116, 188, 189]. Эти методы не требуют определения атомных координат, за исключением грубой оценки взаиморасположения фрагментов, следующей из стереоконфигурации атомов. На их основании было разработано несколько простых в применении и эффективных инструментов, которые до сих пор используются в структурных исследованиях. Для моделирования параметров ЯМР с привлече-

нием методов *ab initio* и функционала плотности необходимо предварительно провести расчёт молекулярной геометрии, или набора геометрий преимущественных конформаций молекулы. Быстрое развитие методов ЯМР-моделирования позволило разработать новые мощные инструменты установления структуры [166].

Среди эмпирических методов предсказания химических сдвигов можно выделить:

- подходы, основанные на базах данных;
- нейронные сети;
- регрессионные подходы;
- подход CHARGE;
- инкрементный подход на уровне остатков.

Простейшие эмпирические методы подразумевают наличие небольшой справочной базы данных, набора правил сложения и инкрементов, параметризованных для каждого класса соединений. Для углеводов это химические сдвиги свободных моносахаридов и измеренные эффекты гликозилирования в зависимости от положения связей, конфигурации и природы связанных остатков. Отклонения от аддитивности, вызванные стерическим влиянием заместителей в соседних положениях, и прочие факторы компенсируются дополнительными базами данных химических сдвигов в ди- и трисахаридных фрагментах [116]. Вместе с нейронными сетями эмпирические методы позволяют проводить самые быстрые и полностью автоматизированные расчёты с точностью 1.6-1.8 м.д. Программы, использующие статистическую обработку химических сдвигов, хранящихся в справочных базах данных, обеспечивают схожую или более высокую точность при бóльших, но все ещё приемлемых затратах ресурсов. Каждому молекулярному фрагменту присваиваются дескрипторы, которые соответствуют основным структурным особенностям данного фрагмента. На основании этих дескрипторов из базы данных отбираются схожие структуры, что позволяет получить средневзвешенные значения экспериментальных параметров ЯМР, соответствующих этим структурам [166, 190].

Следует отметить, что полнота баз данных накладывает ограничения на подобные предсказания. В результате эмпирические методы находят ограниченное применение в установлении вторичной структуры, поскольку неспособны предсказать неусреднённые свойства молекул в определённой конформации или в условиях, отличающихся от хранящихся в базе данных. Кроме того, эти методы не принимают в расчёт различия в условиях регистрации спектров [166].

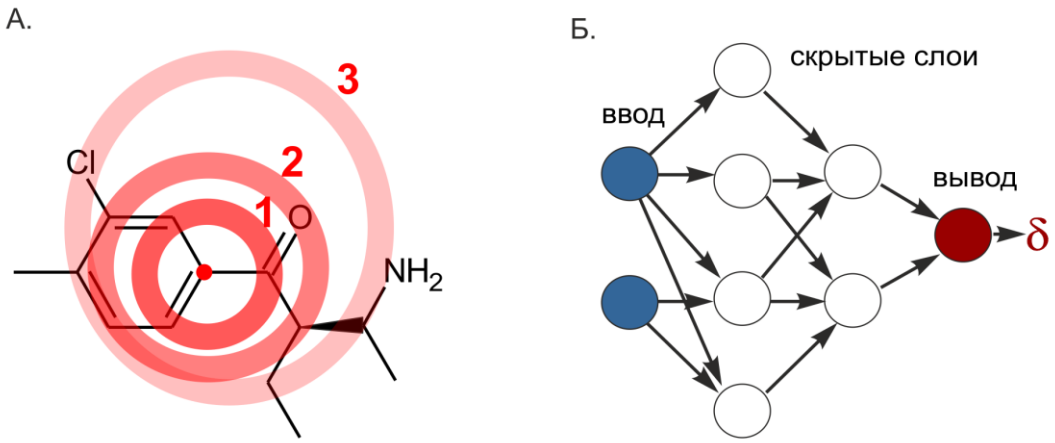



Рис. 9. А. Принцип метода HOSE. В данном примере предсказываются свойства атома, отмеченного красной точкой; сферы пронумерованы. Б. Схематическая архитектура частично-связанной нейронной сети с двумя скрытыми слоями. Синим обозначены входные параметры атома (химический элемент, состояние гибридизации и т.д.) и его соседей в других HOSE-сферах, красным – выходной параметр (предсказываемый химический сдвиг).

Первый подход к предсказанию химических сдвигов, основанный на базах данных, был предложен в 1978 году Бремзером и получил название HOSE (hierarchical organization of spherical environment, иерархическая организация сферического окружения) [191]. Он остаётся одним из самых популярных алгоритмов описания структур при предсказании параметров ЯМР на основе баз данных. HOSE начинает рассмотрение с атома, параметры которого предсказываются, отходит от него на одну связь («1-я сфера») и пробует найти подобное окружение в базе данных. При успешном результате поиска алгоритм отходит от целевого атома на две связи («2-я сфера»), проводит поиск в базе данных и так далее (Рис. 9А). HOSE даёт хорошие результаты для структур, фрагменты которых широко представлены в базе данных. Если анализируемый атом удаётся предсказать с использованием трёх или более сфер, предсказание считается надёж-

ным. В современном варианте HOSE также может учитывать стереохимию (3D HOSE) [192], придавая структурам, имеющим одинаковые конфигурации связей с анализируемой структурой больший вес при усреднении данных. HOSE применяется в химическом программном обеспечении ACD/NMR, Modgraph, MestreLabs NMRPredict, PerkinElmer ChemBioOffice и др. Среди прочих баз данных по химическим сдвигам можно назвать CSEARCH, WINDAT, NMRshiftDB и другие [166, 193, 194].

Нейронные сети представляют собой математические конструкции, позволяющие оптимизировать нелинейные зависимости между дескрипторами на входе и значениями на выходе. Такие сети состоят из искусственных нейронов, организованных в слои, где каждый нейрон представляет собой функцию, которая превращает входную величину в выходную величину. Первый «входной» слой собирает численные атомарные дескрипторы и не проводит с ними вычислений. Во входной слой поступают структурные параметры, которые превращаются в числа с использованием HOSE, инкрементов либо других схем описания структур. При предсказании химических сдвигов последний «выходной» слой состоит из одного нейрона, который выдаёт предсказываемое значение. Выходное значение каждого нейрона в скрытых промежуточных слоях является входным для нейрона следующего слоя (Рис. 9Б). Различные нейронные связи имеют различные «веса», отражающие вклад выхода нейрона в следующий слой. Основными преимуществами предсказаний на основе нейронных сетей является их самообучаемость и способность моделировать свойства соединений без необходимости понимания лежащих в их основе явлений, что особенно востребовано в случае нелинейных взаимосвязей, характерных для инструментальной аналитической химии. Для использования нейронных сетей в моделировании параметров ЯМР необходимо обучить их на выборке известных экспериментальных данных, чтобы оптимизировать веса нейронных связей [166, 195]. К настоящему моменту нейронные сети неоднократно применяли для предсказания химических сдвигов ЯМР, особенно $^{13}$C, для различных классов органических и некоторых биомолекулярных соединений, в том числе углеводов [196-198].

Поиск математической взаимосвязи между структурными дескрипторами и химическими сдвигами (особенно для атомов углерода) и получение весовых

факторов представляли собой сложную задачу на протяжении нескольких десятилетий. В 1987 году Макинтайр и Смолл разработали методологию моделирования спектров ЯМР $^{13}C$ моносахаридов [199]. Использовав собственные и литературные экспериментальные данные, авторы построили модели, соотносившие наблюдаемые химические сдвиги с несколькими числовыми параметрами, кодировавшими особенности химического окружения атомов (функциями расстояний, ван-дер-Ваальсовых энергий и т.п.). Эти параметры описывали влияние кислородных атомов, окружающих атом углерода. С помощью линейного регрессионного анализа были получены модели, независимо предсказывавшие химические сдвиги для пяти видов углеродных атомов в остатках пираноз [199]. Это исследование положило начало развитию регрессионных методов компьютерного анализа параметров ЯМР углеводов [200, 201].

Подход CHARGE представляет собой полуэмпирическую инкрементную схему, основанную на электронных, стерических и других эффектах, параметризованных для разных функциональных групп [202]. Данная схема подразумевает вычисление частичных атомарных зарядов и, на их основании, – химических сдвигов. CHARGE не включает оптимизацию подвижных частей структуры, поэтому геометрия молекулы должна быть заранее известна. Этот подход использован в алгоритме Modgraph[a]. Он не был параметризован для углеводов, однако параметризация для полиатомных спиртов, в том числе инозитола, позволила добиться приемлемого схождения с экспериментальными данными [166, 202].

Универсальные компьютерные инструменты, основанные на инкрементном подходе и нейронных сетях, не обеспечивают точности, достаточной для автоматической расшифровки спектров природных гликанов. В отличие от подходов к фрагментированию на атомарном уровне, алгоритмы, разбивающие структуру на уровне остатков, намного лучше параметризованы для углеводов. Такое разбиение подразумевает приложение эффектов замещения к спектрам моносахаридов или других небольших структурных фрагментов. Чем больше структурных особенностей заместителей учитывается при предсказании, тем более точной оказывается модель, поэтому точность предсказания химических сдвигов в

[a] http://www.modgraph.co.uk/product_nmr.htm

значительной степени зависит от полноты спектроскопических баз данных для определённого класса моносахаридов [166].

Инкрементная схема предсказания химических сдвигов ЯМР $^{13}$C [188] была использована в разработанном с участием автора диссертации инструменте BIOPSEL, который предсказывал химические сдвиги регулярных гликополимеров в водных растворах со средним отклонением 0.13-0.45 м.д. для биогликанов, построенных из распространённых остатков [116]. В настоящий момент эта программа доработана и интегрирована в платформу Carbohydrate Structure Database для моделирования химических сдвигов и эффектов гликозилирования олигомерных и полимерных гликанов, в том числе содержащих редкие моносахаридные остатки и неуглеводные заместители [18].

В основе программы CASPER, предназначенной для расшифровки структур олиго- и полисахаридов на основании данных ЯМР $^{1}$H и $^{13}$C, также лежит алгоритм, использующий инкрементный подход [203, 204]. Для симуляции спектров ЯМР используется три категории данных: химические сдвиги моносахаридов, сдвиги гликозилирования в дисахаридах и корректирующие наборы, представляющие собой разницу между наблюдаемыми химическими сдвигами и химическими сдвигами, рассчитанными с применением аддитивного подхода [204] Среднее отклонение между рассчитанными и экспериментальными химическими сдвигами составило 0.54 и 0.06 м.д. для ядер $^{13}$C и $^{1}$H, соответственно [205].

Для моделирования спектров ЯМР применяют и другие методы предсказания. Теория поляризации связей (bond polarization theory, BPT), полуэмпирический подход, который ищет линейные зависимости между энергиями поляризации связей и атомными зарядами и химическими сдвигами, был использован для предсказания параметров ЯМР твёрдого тела [206]. Применение BPT для расчёта тензоров химических сдвигов $^{13}$C [207] привело к улучшению подхода, в частности, к разработке силового поля COSMOS [171].

Моделирование параметров ЯМР с помощью теории функционала плотности (density functional theory, DFT) обычно проводится в две стадии: сначала оптимизируется геометрия молекулы с целью получения её трёхмерной структуры, а затем для определённой геометрии рассчитываются параметры ЯМР. Часто на этих стадиях применяют разные уровни теории: в большинстве случаев рас-

чёт параметров ЯМР требует более сложного уровня. Выбор правильного сочетания уровней теории имеет большое значение. Для предсказания параметров ЯМР разработано несколько подходов, пригодных для углеводных молекул: Gauge-Independent Atomic Orbitals (GIAO) [208], Individual Gauge Localized Orbital (IGLO) [209, 210], Localized orbital/local origin (LORG) [211], Gauge-Including Projector Augmented-Wave (GIPAW) [212]. Низкую производительность GIAO в значительной степени компенсирует развитие компьютерных технологий, поэтому данный подход в сочетании DFT часто используют для быстрого предсказания параметров ЯМР органических и биомолекулярных систем [213]. На основании развёрнутого анализа предсказательной силы различных базисных наборов минимально достаточным уровнем теории для предсказания геометрии и химических сдвигов ЯМР в сахаридах признан B3LYP/6-311++G(2d,2p) [214, 215]. Другие уровни теории и базисные наборы, пригодные для ЯМР моделирования в органической химии, перечислены в анализе достоверности квантовохимического ЯМР-моделирования, сделанного Панкратьевым и коллегами [216].

Применительно к углеводам типичная точность модели в зависимости от выбранного метода проиллюстрирована на Рис. 10 для распространённого модельного объекта, содержащего как консервативный пиранозный, так и конформационно-лабильный фуранозный остатки – сахарозы (α-D-Glc*p*-(1-2)-β-D-Fru*f*). Несмотря на активное развитие универсальных общих моделей *ab initio*, частная углеводная эмпирическая модель все ещё демонстрирует лучшую точность для характерных представителей сахаридов. Время счета[a] варьировалось от 0.1 секунды для эмпирического расчёта до 68 часов для COSMO/GIAO. Учёт растворителя привёл к пятикратному увеличению ресурсоёмкости, а использование реализации уровня теории PBE в программе PRIRODA [217] позволило получить относительно точную модель за 29 минут.

Примеры применения квантовомеханических подходов для предсказания химических сдвигов моно- и олигосахаридов приведены в Табл. 5. Более подробно эти и другие исследования квантовомеханических ЯМР-моделей углеводов описаны в обзоре автора диссертации [166]. Следует отметить, что в экспе

[a] На одном ядре 3.0 ГГц процессора персонального компьютера [данные автора].

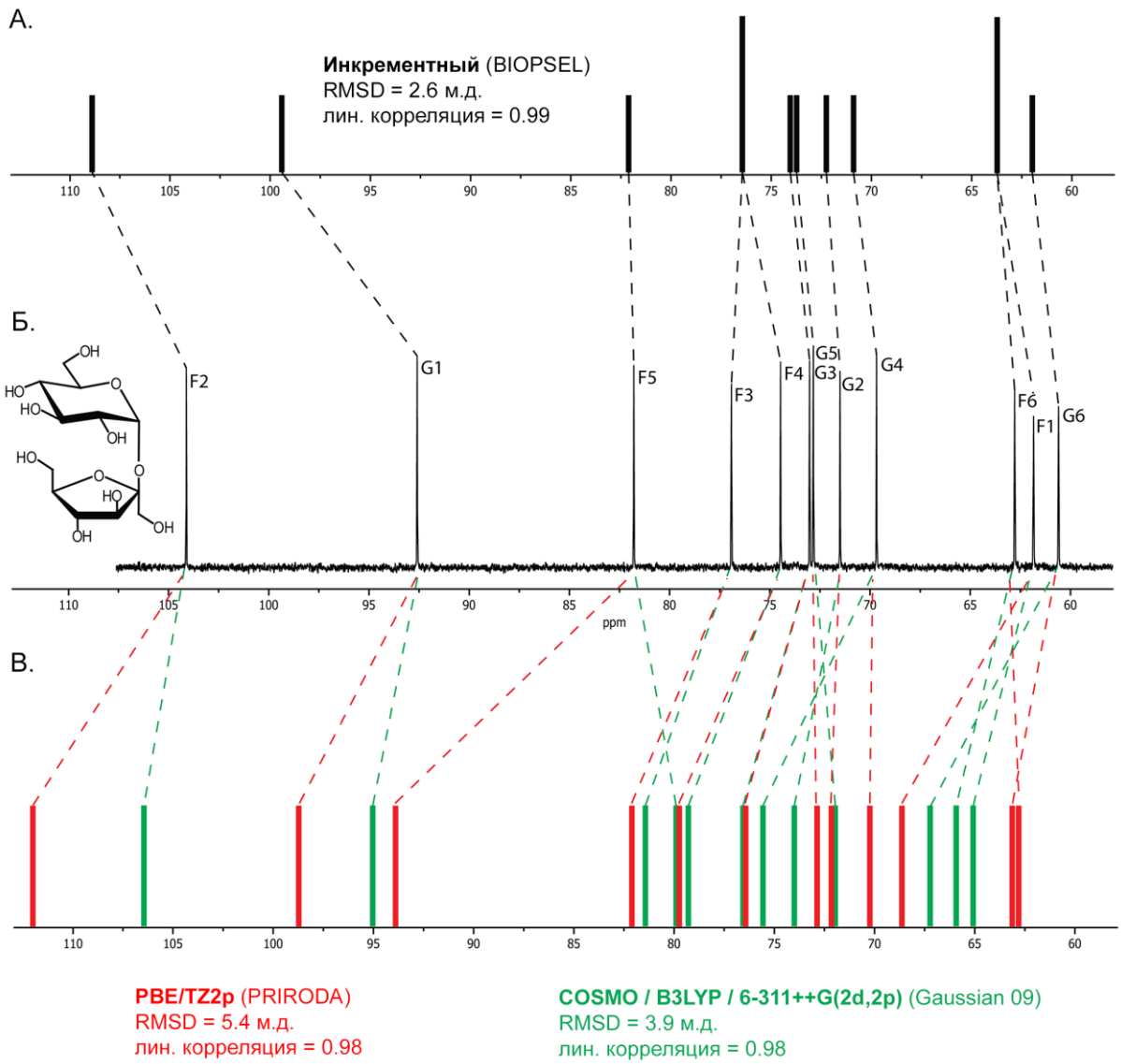



Рис. 10. Сравнение $^{13}$C ЯМР-моделей сахарозы, полученных эмпирически (А) и на двух квантовомеханических уровнях теории (В) с экспериментальным спектром в водном растворе при 25°C (Б, записан автором). Красные линии и текст соответствуют PBE/TZ2p, зелёные – B3LYP/6-311++G(2d,2p), модель растворителя COSMO. Пунктир отражает корреляции между сигналами, латинские буквы – отнесение сигналов к остатку глюкозы (G) или фруктозы (F). Воспроизведено из [166].

рименте ЯМР регистрируется химический сдвиг, усреднённый для равновесных конформаций, существующих в растворе гликана, в то время как квантовомеханическая модель химического сдвига соответствует конкретной молекулярной геометрии. Предсказание химических сдвигов в зависимости от торсионных углов гликозидного мостика, как структурно значимого, но при этом конформационно-лабильного фрагмента в углеводах наиболее изучено для модельных дисахаридов. В качестве характерного примера можно привести расчёт геометриче-

ских параметров [218] и химических сдвигов [219] трансгликозидных атомов углерода в α-мальтозе (α-D-Glc*p*(1→4)α-D-Glc*p*, Рис. 11). Применение этой методологии к предсказанию усреднённых химических сдвигов всех атомов с учётом распределения конформаций по энергии требует несоразмерно больших вычислительных ресурсов.

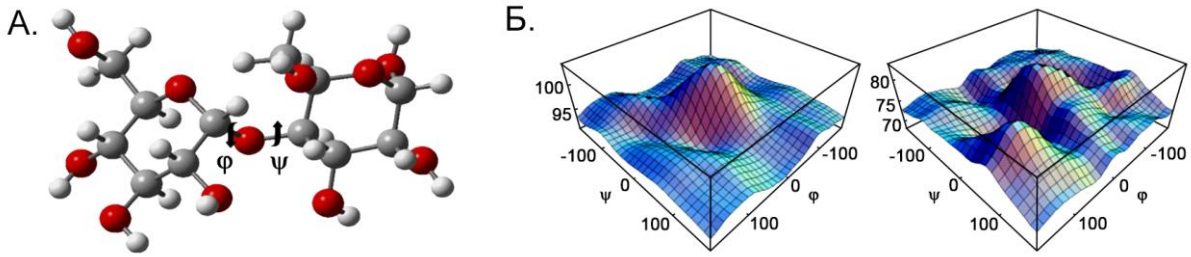



Рис. 11. Структура мальтозы, полученная методом ONIOM DFT:HF (А) и зависимость химических сдвигов, предсказанных на уровне HF/321G и нормированных в 6-311G**, от торсионных углов гликозидной связи (Б). Скомпоновано из [218, 219].

Табл. 5. Примеры использования методов теории функционала плотности для предсказания химических сдвигов сахаридов.

| *Объект (молекула) [ссылка]* | *Параметр: ядро** | *Метод расчёта* | | *Применение; программа* |
|---|---|---|---|---|
| | | *Геометрия* | *Экранирование* | |
| α-D-Glc*p*, β-D-Glc*p* (равновесные конформеры в водном растворе) [220] | ХС: $^{1}$H, $^{13}$C | B3LYP/6-31G(d,p) Энергия сольватации: B3LYP/6-311++G(2d,2p) | B3LYP/pcJ | Анализ экспериментальных данных. *Gaussian 03* |
| α-D-Glc*p*N (мономер хитозана) [221] | ТХС: $^{1}$H, $^{15}$N, $^{17}$O (твёрдое тело) | PCA; B3LYP/6-31++G(d,p), только для протонов | B3LYP/6-311++G(d,p), B3LYP/6-31++G(d,p) | Исследование влияния водородных связей на тензоры ХС. *Gaussian 98* |
| α-D-Glc*p* (газовая фаза) [222] | ХС: $^{1}$H, $^{13}$C | B3LYP/6-31G(d,p), B3LYP/6-31+G(d,p) | B3LYP/cc-pVTZ; B3LYP/aug-cc-pVTZ | Исследование эффектов растворителя и сравнение методов расчёта. |

| | | | | |
|---|---|---|---|---|
| | | | | *Gaussian 03 (QM), MOSCITO (MM)* |
| β-D-Glc*p*-(1-4)-β-D-Glc*p*, α-D-Glc*p*-(1-4)-α-D-Glc*p* [223] | XC: $^{13}$C (C1) (твёрдое тело) | B3LYP/6-311+g(d,p) | B3LYP/6-311+g(d,p) | Исследование молекулярного окружения в хиральных полостях коммерческих полисахаридных сорбентов. *Gaussian 03* |
| β-D-Glc*p*-(1-4)-β-D-Glc*p* (целлобиоза) [224] | XC: $^{13}$C | Данные РСА | B3LYP/6-311++G(2d,p) (процедура GAIOCHF) | Теоретическое исследование влияния конформации и водородных связей на изотропные химические сдвиги $^{13}$C. *Gaussian 03* |
| α-D-Glc*p*-(1-4)-α-D-Glc*p*, α-D-Glc*p*-(1-4)-β-D-Glc*p* [225] | XC: $^{1}$H, $^{13}$C (твёрдое тело) | PBE / planewave | GIPAW PBE / planewave | Исследование слабых водородных связей. *CASTEP (геометрия), PARATEC (ЯМР)* |
| α-D-Glc*p*-(1-2)-β-D-Fru*f* (сахароза) [226] | ТХС: $^{13}$C (твёрдое тело) | Данные нейтронографии | RHF, HFB, HFS, BLYP, B3LYP, B3P86, BVWN, SVWN, MPW1PW91 / cc-pVDZ / cc-pVTZ | Сравнение функционалов DFT и HF. *Gaussian 03* |

* ХС, химический сдвиг, ТХС, тензор химического сдвига.

Неэмпирические подходы, в отличие от остальных, моделируют не только химические сдвиги, но и константы спин-спинового взаимодействия (как в пределах остатков, так и трансгликозидные), времена релаксации и другие параметры, наблюдаемые в экспериментах ЯМР. В частности, для предсказания КССВ

широко применяют квантово-химические подходы [227], включая расчёт нерелятивистских составляющих КССВ на основе уравнений Рамсея [228]. Молекулярную динамику используют для моделирования времён релаксации и ядерных эффектов Оверхаузера (NOE) [166].

Разработка инструментов полностью автоматического установления структуры углеводов по спектрам ЯМР в настоящее время находится в зачаточном состоянии. Кроме созданной под руководством автора диссертации платформы CSDB, тщательно параметризованные для углеводов инструменты моделирования спектров ЯМР и предсказания структуры по спектрам реализованы только в рамках проекта CASPER [203], однако их возможности ограниченны малым набором мономерных остатков и структурных особенностей [119, 120]. По данным направленного сравнительного исследования существующие способы предсказания данных ЯМР (включая квантово-механические, статистические, эмпирические и нейросетевые методы) плохо оптимизированы для углеводов и не пригодны для решения структурных задач гликобиологии [166]. Это касается и реализованного в GLYCOSCIENCES.de инструмента получения химических сдвигов углеводов, усреднённых для атомов в остатках без учёта влияния соседних остатков.

### *2.6. Статистический и кластерный анализ гликомов*

Статистический анализ содержимого углеводных баз данных, особенно претендующих на полноту покрытия, представляет значительный интерес для исследователей и позволяет выявить уникальные структурные особенности сахаридов, характерные для конкретных таксономических групп. Эта информация востребована в иммунологических исследованиях и серотипировании [26, 31, 229].

Прокариоты отличаются большим структурным разнообразием клеточных стенок [230]. Клеточные стенки Грам-положительных и Грам-отрицательных бактерий построены на основе пептидогликана – полимера, в котором полисахаридные цепи перекрёстно сшиты короткими пептидными цепями. Грам-отрицательные бактерии имеют дополнительную наружную мембрану, которая состоит из комплекса белков с липополисахаридами, состоящими, в свою очередь из липида А, олигосахаридного кора и полимерного О-антигена. У Грам-положительных бактерий наружная мембрана отсутствует, но пептидогликановая стенка толще (>30 нм, в сравнении с 10 нм) и содержит полисахариды с тейхоевыми кислотами [230]. Оба типа бактерий продуцируют внеклеточные полисахариды, которые представляют собой капсулу, присоединённую к клеточной оболочке, либо слизь, слабо связанную с поверхностью клетки. Эти гликоконъюгаты и полисахариды содержат антигенные детерминанты, ответственные за запуск иммунного ответа в организме-хозяине, и являются сайтами узнавания патогенов [231], в том числе, бактериофагами [26].

Полимерные структуры углеводных антигенов уникальны и часто построены из множества повторяющихся звеньев. Информация о строении бактериальных поверхностных углеводов и их модификациях необходима для понимания механизмов клеточного узнавания, адгезии и развития иммунного ответа на молекулярном уровне, что, в свою очередь, является основой разработки синтетических углеводных вакцин, диагностических инструментов и иммуностимуляторов [26, 232].

Известно, что углеводы бактерий проявляют намного большее структурное разнообразие, чем углеводы млекопитающих [233, 234], однако до сих пор было предпринято лишь несколько попыток статистического анализа бактери-

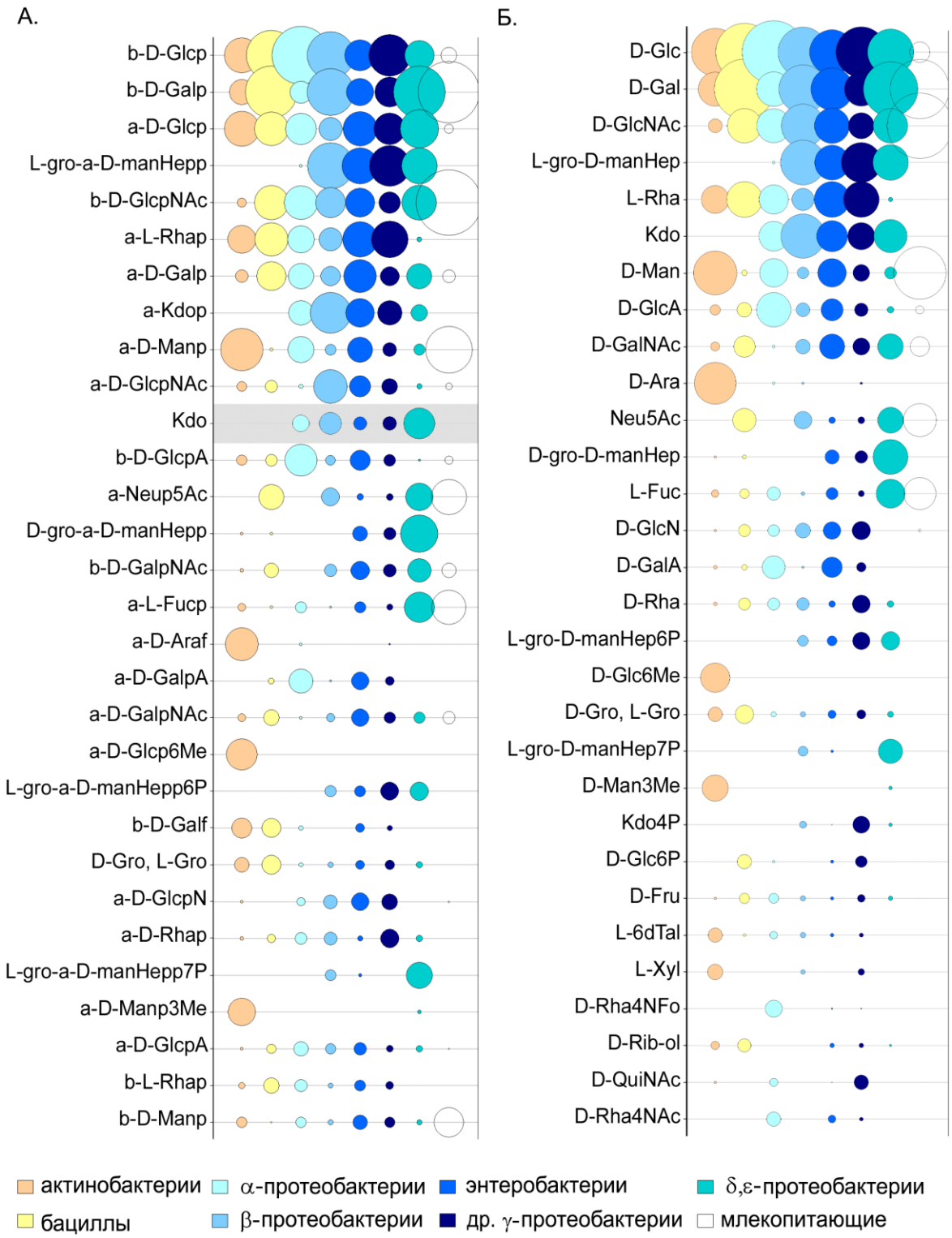



Рис. 12. Моносахариды, наиболее распространённые в бактериальных гликанах. Площади кругов соответствуют относительной частоте встречаемости 30 наиболее распространённых остатков с учётом (А) или без учёта (Б) аномерной конфигурации в структурах, хранящихся в BCSDB. Остаток 3-дезокси-D-*манно*-окт-2-улозоновой кислоты (Kdo) без аномерной конфигурации является аналитическим артефактом и выделен серым цветом. Воспроизведено из [26].

альных структур, собранных в базах данных. В исследовании Хергета и коллег с участием автора диссертации проведена оценка моносахаридных остатков (Рис. 12), модификаций и связей, присутствующих в углеводных структурах бактерий, хранящихся в базе данных BCSDB, по сравнению с гликанами млекопитающих [26]. Целью этого исследования было определение минимального набора строительных блоков для автоматического твердофазного синтеза произвольного гликана [235-237] из заданной таксономической группы. Впоследствии аналогичный статистический анализ был проведён под руководством автора диссертации с более широким таксономическим охватом с использованием базы данных CSDB, которая включает в себя бактериальную часть на основе BCSDB [31].

Классификация живых организмов является сложной задачей биологии [238]. Согласно классификации, основанной на последовательностях рибосомальной РНК, выделяют три домена жизни: эукариоты, бактерии (эубактерии) и археи (архебактерии) [239]. Внедрение секвенирования геномов в рутинную лабораторную практику позволило построить обширные генетические библиотеки для различных видов и привело к расцвету геномного филогенетического подхода [240, 241]. Помимо классической филогении по консервативной субъединице 16S рРНК, также используют анализ, не связанный напрямую со сходствами последовательностей, например, присутствие/отсутствие белковых семейств, видоспецифичное использование кодонов, аминокислотный состав, распределение типов белковой укладки, присутствие консервативных пар генов и сравнительный порядок генов-ортологов [240, 242-246]. Результаты большинства таких исследований согласуются между собой, хотя тщательное изучение различных филогенетических деревьев позволяет выявить новые взаимосвязи между видами и новые эволюционные феномены, такие как горизонтальный перенос генов [247].

Однако поиск эволюционных взаимосвязей возможен не только на уровне геномов. Активность генов обычно проявляется не изолированно, а в составе сетей и путей различной сложности. Наглядный пример такого проявления – метаболизм, который зависит от согласованного действия ферментов, объединённых в специфические цепи [248, 249]. Перетасовка ферментативных активностей позволяет обеспечить метаболическую пластичность, необходимую для успешной адаптации организма к различным экологическим нишам. Распределение

ферментативных функций у различных видов можно сопоставить с их физиологическими особенностями, давлением отбора, которому подверглись эти виды в ходе эволюции, и такими свойствами, как патогенность. Агиляр и коллеги провели анализ фенетических деревьев, построенных на основе различных метаболических путей, для 27 организмов, принадлежащих доменам эукариот, бактерий и архей, и сравнили их с классическими деревьями на основе рРНК [247]. Их результаты свидетельствуют о том, что филогенетически близкие организмы могут быть метаболически далеки друг от друга, и наоборот. Метаболические связи между организмами в значительной степени коррелируют с экологическими факторами, более того, была предпринята попытка сопоставить уникальность углеводного метаболизма с группами в пределах таксона бактерий (Рис. 13).

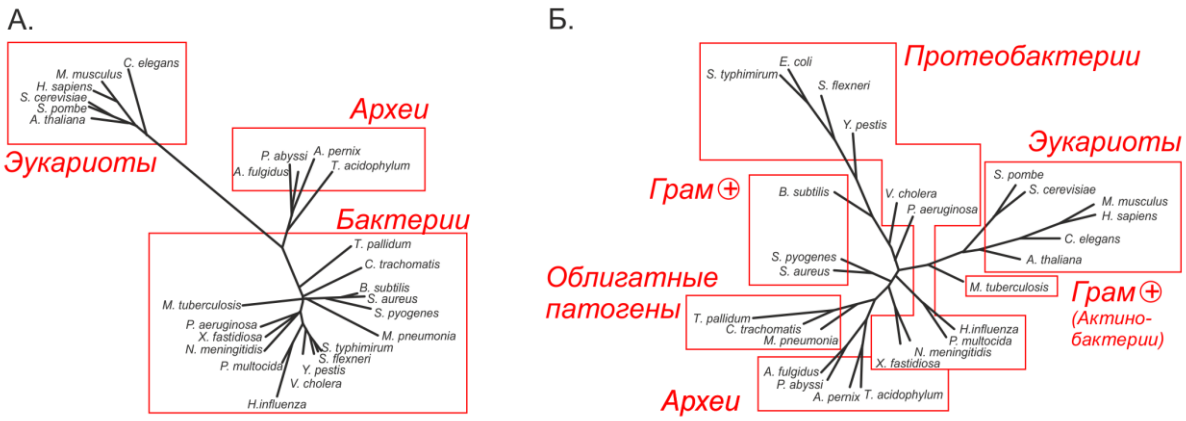
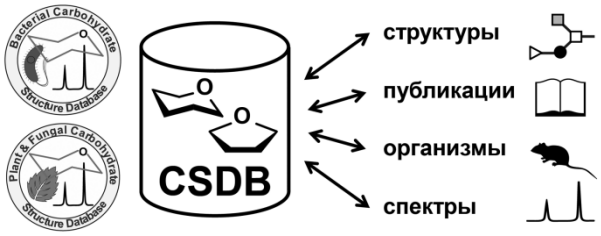



Рис. 13. Кластеризация и филогенетические взаимосвязи отдельных таксонов, построенные на основании последовательностей рРНК (А) и распределения ферментов метаболизма углеводов (Б). Модифицировано из [247].

Аналогичным образом, кластерный анализ гликомов и построение фенетических деревьев на основе углеводных структур могут быть востребованы в установлении и подтверждении ферментативных активностей гликозилтрансфераз, экспериментальная характеризация которых значительно отстаёт от теоретической. Два организма, обладающих сходными углеводными структурами, должны также обладать ферментами CAZy со сходными активностями. Следовательно, выявление углеводных «родственных» связей между таксонами сможет ускорить изучение углевод-активных ферментов с неустановленными функциями [31]. В настоящий момент подобные исследования ведутся только в группе автора диссертации.

## 3. База данных природных углеводов CSDB как платформа гликоинформатики (обсуждение результатов в контексте работы)

Новая курируемая реляционная база данных (Carbohydrate Structure Database, CSDB[a]) была спроектирована, разработана, заполнена данными, снабжена интерфейсом и представлена в сети Интернет. Этот процесс начался в 2005 году [75], а к 2010 году проект обрёл современную реализацию, основанную на новых правилах компьютерного описания углеводов [64]. Предназначением CSDB является представление опубликованных данных по природным гликанам, гликополимерам, гликоконъюгатам и другим углеводсодержащим соединениям. Четыре основных типа информации, содержащейся в базе, отражены на схеме:




В ходе своего развития CSDB превратилась в платформу для углеводных сервисов, которая стала одним из ведущих мировых ресурсов гликоинформатики, не имеющим аналогов как по покрытию, так и по идеологии (обзоры других авторов: [33, 51, 52, 250-253]; статьи коллектива разработчиков: [18, 64, 254, 255]). В настоящее время она предоставляет не только данные [18, 74, 256, 257], но и инструменты их выборки [18, 64, 70, 254, 255], верификации [42], визуализации [119, 134, 146], анализа [26, 31], предсказания [117-120, 146, 255] и другие сервисы углеводной тематики [44]. Все наработки свободно доступны для практического использования с помощью веб-интерфейса. Главный экран сайта CSDB (Рис. 14) дает представление об основных функциях и наиболее востребованных инструментах. Подробности приведены в последующих разделах. Основные вехи развития пользовательских инструментов CSDB представлены в разделе «История обновлений» на сайте проекта[b].

[a] http://csdb.glycoscience.ru/

[b] http://csdb.glycoscience.ru/help/about.html

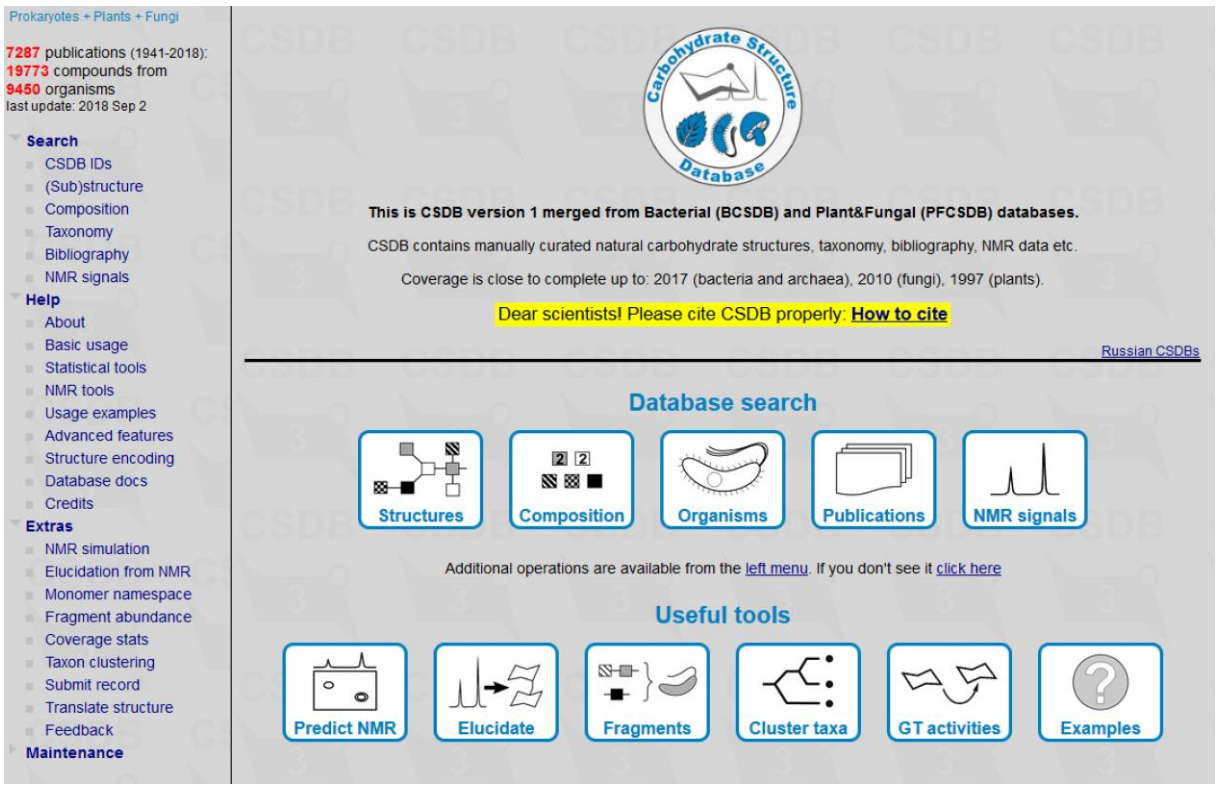



Рис. 14. Входная веб-страница базы CSDB: главное меню, логотип, основные инструменты.

## *3.1.Данные CSDB*

### *3.1.1. Типы данных CSDB*

Типы данных CSDB с различным уровнем верифицируемости соответствуют традиционной для биоинформатики классификации: курируемые первичные данные, полученные анализом публикаций; мета-данные (данные о данных), полученные вручную и автоматически; опосредованные данные, полученные с помощью дальнейшего анализа первичных данных; и предсказанные данные, получение которых подразумевает использование теоретических моделей. Расположение этих типов на шкалах проверяемости и универсальности, а также пути переходов между ними, приведено на Рис. 15. Не вся информация, предоставляемая платформой CSDB, хранится непосредственно в базе – некоторые ее виды генерируются непосредственно по запросам пользователей в соответствии с моделями, прошедшими валидацию в рамках проекта CSDB.

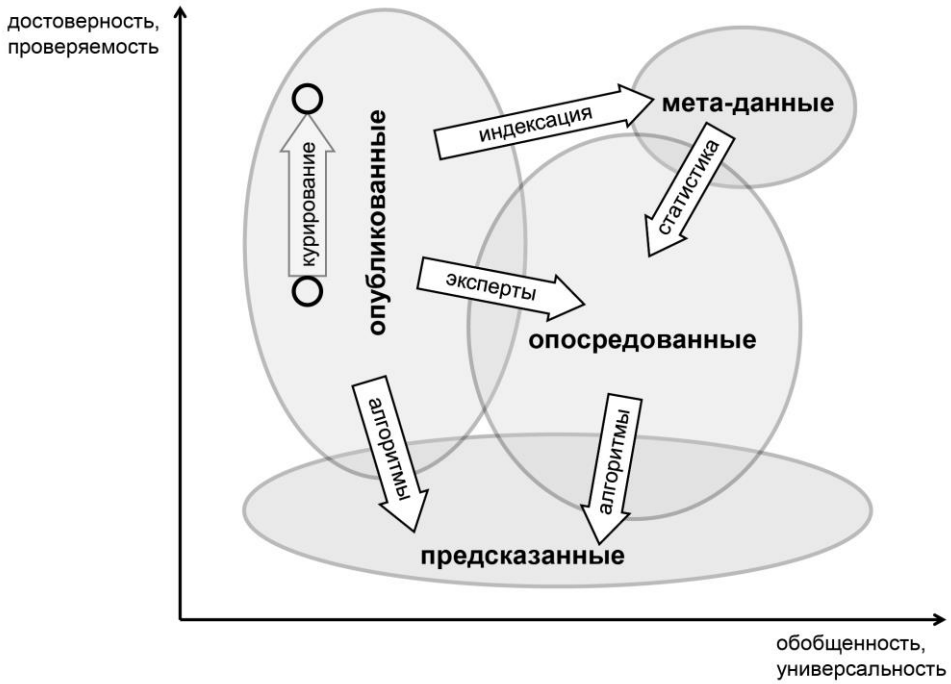



Рис. 15. Взаимодействие глобальных типов данных в CSDB.

Опубликованная и мета-информация, содержащаяся в CSDB, суммирована в Табл. 6. **Жирным шрифтом** показаны обязательные данные, присутствующие во всех записях. *Курсивом* показаны редкие данные, присутствующие менее чем в четверти записей.

Табл. 6. Данные CSDB

| *Группа* | *Данные* | *Индексация и интеграция с другими проектами* | *Примечания* |
|---|---|---|---|
| Структура | **Первичная структура** | выводится в форматах SweetDB, SNFG, в виде структурных формул или 3D-моделей. | Хранится в виде таблицы связности остатков, для ввода-вывода кодируется на языке CSDB Linear. Поддерживаются как точные, так и неоднозначные структуры. |
| | *Ошибки в структуре* | | При наличии ошибок также приводится неправильная опубликованная структура. |
| | **Тип молекулы** | контролируемый словарь | мономер, олигомер, химическое повторяющееся звено, биологическое повторяющееся звено, звено циклического полимера, фрагмент, мотив, гомополимер |
| | *Число повторов* | | для полимеров |
| | *Мол. вес* | | |
| | **Мономерный состав** и *брутто-формула* | неявно задан в структуре | Состав рассчитывается из структуры, брутто-формула – для олигомеров. |
| | *Агликон и позиция его присоединения* | | При возможности включён непосредственно в первичную структуру. Если это невозможно - закодирован в виде SMILES, IUPAC или суперклассов. |
| | *Молекулярная геометрия* **(или ее расчёт)** | | Для опубликованных данных - только факт наличия; расчётная геометрия хранится в базе для всех структур. |
| | Класс, роль | индексированы | например, «О-антиген», «гликосфинголипид» и т.п. |

| | | | |
|---|---|---|---|
| | *Тривиальное название* | индексированы | |
| | *Биосинтетические и генетические данные, названия связанных ферментов* | | Для *E. coli* и *A. thaliana* – полный набор данных группы «гликозилтрансферазы»;<br>для биосинтеза и генетики остальных углеводов - только наличие данных в публикации. |
| | *Тип искусственного синтеза природной структуры* | индексированы | при наличии - химический, ферментативный, *in vivo*, моделирование и т.д. |
| Гликозилтрансферазы | **Гены** | индексированы | название, ссылка на GenBank, ссылка на кластер в GenBank |
| | **Ферменты** | индексированы | название, группа, ссылка на GenBank / Uniprot |
| | **Синтезируемая связь** | индексированы | в контексте полной структуры и синтезирующего её организма |
| | **Донор и субстрат** | индексированы | представлены как виртуальные структуры в CSDB |
| | **Степень достоверности** | индексированы | *in silico*, *in vitro*, три градации *in vivo* |
| | **Методы подтверждения активности** | | |
| Запись (уникальная комбинация структуры и статьи, в которой она описана) | **Признак новизны** | | установлена ли структура впервые в этой статье |
| | **Методы, применённые к структуре** | индексированы | |
| | **Локализация структуры в статье** | | номер рисунка, схемы и т.д. |

| | | | |
|---|---|---|---|
| | Комментарии | | все, что не удалось закодировать в типизированных полях, включая информацию об ошибках в публикациях |
| Библиография | **Авторы** | индексированы | |
| | **Название** | | |
| | **Название журнала, сборника или книги** | индексированы; для журналов – доп. внешний индекс NCBI NLM ID | |
| | *Издательство, редакторы* | редакторы индексированы | для книг |
| | **Выходные данные** | | год, том, страницы |
| | Ссылки на внешние библиографические базы | внешние индексы DOI и/или PMID | DOI, NCBI PubMed ID, веб-адрес |
| | *e-mail автора-корреспондента* | | |
| | аффилиации авторов | индексированы | |
| | ключевые слова | индексированы | |
| | реферат (abstract) | | |
| Таксономия и природный контекст | **Царство, тип** | индексированы | |
| | **Род, вид, штамм / серогруппа** | индексированы (три индекса) | включая неполные сочетания, а также гибриды и мутанты |
| | *Переименования и переклассификации таксонов* | | по отношению к опубликованному названию |
| | *Орган, ткань, стадия развития* | индексированы | |
| | *Болезнь организма-хозяина* | индексированы | ассоциированная с микроорганизмом или со структурой |

| | | | |
|---|---|---|---|
| | *Организм-хозяин* | индексированы | для микроорганизмов |
| | **Ссылка на базу NCBI Taxonomy** | внешний индекс NCBI TaxID | |
| ЯМР | Спектр $^{1}$H | | включая отнесение сигналов |
| | Спектр $^{13}$C | | включая отнесение сигналов |
| | Температура и pH | | |
| | Растворитель | индексированы | в том числе поддерживаются смеси, указание концентраций и стандартов калибровки шкалы химических сдвигов |
| Универсальные данные | **Взаимосвязи между остальными группами данных и идентификаторами** | | |
| | Свойства мономеров | индексированы; контролируемый словарь; ссылка на MSDB | название, возможные конфигурации остатка (аномерная, абсолютная, размер цикла), стереоконфигурации атомов, число протонов в каждой позиции, тип заместителя в каждой позиции, запись WURCS, запись SMILES |
| | Суперклассы мономеров | индексированы | разновидность неопределённости структуры на уровне остатков |
| | Топологии соединения остатков | индексированы | до 12 остатков |
| | Эмпирические эффекты гликозилирования в спектрах ЯМР $^{13}$C | индексированы | 215 эффектов |
| | Данные ЯМР модельных структур | индексированы | 2679 спектров (313 остатков) |
| | Разрешённые типы связей | индексированы | для всех комбинаций типов атомов |

| | | | |
|---|---|---|---|
| Вспомогательные данные | **Идентификаторы записи** | | запись = уникальная комбинация структуры и статьи, в которой она описана |
| | **Идентификаторы соединений** | | соединение = уникальная комбинация первичной структуры, её типа, аглико-на, степени полимеризации и других свойств молекулы |
| | **Идентификаторы статей** | | |
| | **Идентификаторы спектров** | | |
| | **Идентификаторы организмов** | | организм = уникальная комбинация таксона (царство, тип, род, вид), штамма и/или серогруппы |
| | **Идентификаторы активностей гликозилтранс-фераз** | | |
| | **Идентификаторы генов и ферментов** | | |
| | *Ссылки на записи в других базах* | | GlycomeDB, CCSD (CarbBank), Chemical abstracts, USA patent, Uniprot, Genbank и др. |
| | Ссылки на родственные записи в CSDB | | Например, структуры, синтезируемые тем же организмом, структуры, отличающиеся только рамкой полимеризации, или идентичные структуры, опубликованные в другом контексте. |
| | **Данные для отслеживания процесса аннотирования** | | аннотатор, контролёр, дата, ссылка на запись в лабораторной базе, ссылка на файл со статьёй, ошибки в других базах, исправленные при импорте данных. |

### *3.1.2. Покрытие CSDB и источники данных*

Таксономический охват базы данных CSDB включает микроорганизмы (бактерии, археи, одноклеточные грибы, простейшие), растения и грибы. Покрытие по прокариотам близко к полному (т.е. >90% всех опубликованных данных попадает в базу), что обеспечивает научную ценность даже отрицательного ответа на поисковые запросы. В настоящее время CSDB является единственной постоянно обновляемой базой, обеспечивающей практически значимое покрытие по углеводам прокариот. Данные по прокариотам попадают в базу в среднем через год после их опубликования. Покрытие по грибам близко к полному до 2010 года включительно и активно расширяется, покрытие по растениям – до 1997 года. Углеводы человека и других многоклеточных животных не включены в CSDB, так как, в отличие от упомянутых доменов, эта таксономическая группа представлена в других углеводных базах. База данных CSDB пополняется на систематической основе по результатам аннотирования 500-1000 публикаций в год.

Объем аннотированных данных CSDB на 2018-й год представлен в Табл. 7. с разбивкой по доменам. Покрытие в пределах меньших таксономических групп (Рис. 16) отражает относительную изученность таксонов в публикациях.

Табл. 7. Покрытие CSDB (количество объектов)

| *Домен* | *Углевод-содержащие структуры* | *Таксоны (организмы и их группы)* | *Публикации* | *Отнесённые спектры ЯМР ($^{13}C$ и $^{1}H$)* | *Активности гликозилтрансфераз* |
|---|---|---|---|---|---|
| Бактерии | 12695 | 7333 | 5022 | 5460 | 829 |
| Растения | 4892 | 1046 | 1532 | 3073 | 926 |
| Грибы | 1616 | 889 | 663 | 752 | 0 |
| Простейшие | 230 | 41 | 54 | 23 | 0 |
| Археи | 100 | 82 | 38 | 20 | 0 |
| *Всего* * | 19483 | 9150 | 7285 | 9427 | 1755 |

* В строке «Всего» представлено логическое объединение значений всех групп, включая небольшое число записей для углеводов других доменов. При попадании объекта в несколько групп в строке «Всего» он учитывается один раз.

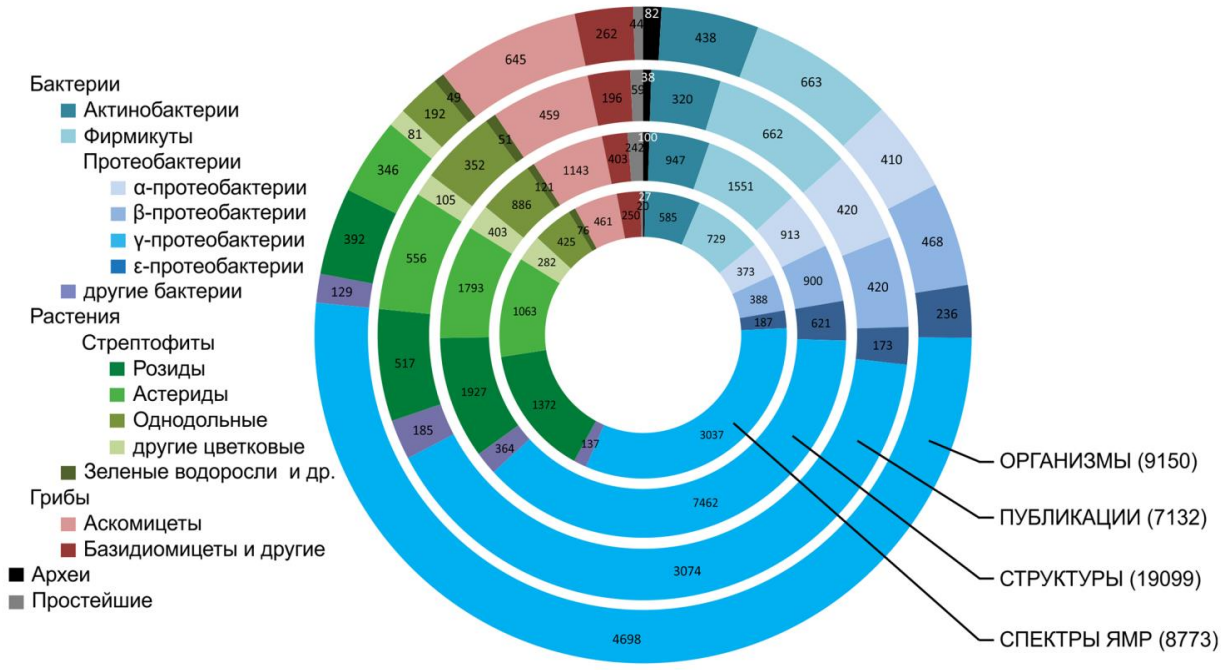



Рис. 16. Количество записей CSDB в основных таксономических группах (на 2017-й год).

CSDB является первичной базой, т.е. содержит данные, полученные непосредственно из научной литературы (включая данные экспериментов). Данные попадают в базу следующими способами:

1. Ретроспективный анализ и аннотирование научной литературы. Отбор публикаций проводится по критериям наличия в статье хотя бы одной явно или неявно заданной структуры, удовлетворяющей следующим критериям: она содержит хотя бы один углеводный остаток, но не является нуклеиновой кислотой; она определена достаточно полно, чтобы иметь возможность судить о мономерном составе и не менее, чем о 50% связей и конфигураций остатков; она соотносится со структурой из однозначно идентифицированного природного источника (организма), принадлежащего к микроорганизмам, растениям или грибам. Под соотнесением с природной структурой понимается одно из следующего: структура выделена из природного источника; структура является частью выделенной природной молекулы большего размера; синтетическая структура, идентичная природной или отличающаяся от неё только агликоном; выделенная структура, полученная модификацией природной структуры (напр., в процессе анализа). Ежегодный процесс импорта данных включает поиск в системах Web Of Sci-

ence [258] и NCBI Pubmed [36] по ключевым словам, первичный отбор на основании рефератов, вторичный отбор на основании анализа текстов статей, аннотирование статей в виде текстового дампа, автоматическое выявление ошибок и несоответствий данных, проверку аннотаций другим аннотатором, исправление найденных ошибок, загрузку дампа в базу.

2. Загрузка пользователями собственных опубликованных данных. При этом проводится автоматическая и ручная проверка качества данных, после чего данные попадают в дамп текущего года.
3. Другие базы данных. Структуры, опубликованные до 1996 года (около 40% всех структур в CSDB), были отобраны по критериям таксономической принадлежности из базы данных CCSD (Carbbank) [62] с полным покрытием до 1996 года и проверены. Приблизительно 50% статей из записей Carbbank были повторно аннотированы с внесением недостающих и исправлением ошибочных данных. Библиографические и таксономические данные при импорте сопоставляются с базами NCBI NLM Catalog (идентификация журналов [259]) и NCBI Taxonomy (идентификация организмов, переименований и переклассификаций таксонов [260]). Данные по молекулярным свойствам моносахаридов сопоставляются с базой MonosaccharideDB [101]и обновляются вместе с ней.
4. Обобщённые и опосредованные данные (напр., молекулярная геометрия, характеристичные химические сдвиги и эффекты гликозилирования в спектрах), полученные на основании собственного статистического анализа и предсказания свойств по собственным моделям. Эти данные попадают в часть базы, изолированную от опубликованных данных, недоступны для поиска и используются для моделирования и предсказания свойств.

### *3.1.3. Контроль ошибок*

База данных CSDB является курируемой, т.е. данные проверяются экспертами, что делает CSDB одним из самых надёжных источников данных об углеводах. В рамках доменов прокариот, растений и грибов CSDB является единственной первичной курируемой базой. Качество данных традиционно является краеугольным камнем для биохимических баз. В CSDB оно обеспечивается автоматическим выявлением более ста видов ошибок в данных и «подозрительных» сочетаний данных. Проверка и верификация аннотаций также включает ручное выявление ошибок, не обнаруживаемых автоматически (напр., химически возможный, но не соответствующий оригинальной публикации способ связывания моносахаридов в структуре). В порядке распространённости к ошибкам относятся: ошибки, пришедшие из других баз; некорректные аннотации; ошибки в оригинальных публикациях; ошибки в программах обработки данных. Основные группы ошибок, верифицируемых при работе с дампом, перечислены в

Табл. 8. Тип «программа» означает, что ошибка автоматически выявляется и исправляется с выдачей предупреждения, «выявление» - что ошибка выявляется автоматически, но для исправления нужно повторное аннотирование, «эксперт» - что ошибку можно выявить только путём ручного сопоставления записи с оригинальной публикацией.

Табл. 8. Типы ошибок в данных

| *группа* | *примеры ошибок* | *тип* |
|---|---|---|
| молекулярная структура | неканоническая сортировка боковых цепей и выявление главной цепи;<br>то же самое химическое повторяющееся звено полимерной структуры присутствует в других записях с другим положением рамки полимеризации;<br>для одной и той же олигомерной молекулы в разных записях указан разный мол. вес или брутто-формула;<br>избыточная информация о стереохимии атомов | программа |
| | некорректный синтаксис записи структуры на CSDB Linear;<br>нераспознанный тип молекулы;<br>указанный тип молекулы противоречит её записи;<br>неканонические / нераспознанные наименования или модификации | выявление |

| | | |
|---|---|---|
| | остатков либо модификации, не совместимые с остатком;<br>в соседних остатках связаны химически несовместимые положения либо позиция замещения не укладывается в размер остатка;<br>положение связи не соответствует правилам нумерации атомов в остатках данного типа (напр., для углеводного остатка указано буквенное обозначение положения)<br>невозможная стехиометрия боковой цепи;<br>указанные конфигурации (аномерная, абсолютная или размер цикла) несовместимы между собой, с типом остатка или с его атомарным дескриптором либо отсутствуют в тех остатках, где должны быть;<br>структура содержит нестандартные компоненты, не объяснённые в комментариях, или на один компонент приходится несколько комментариев;<br>нехарактерные для природных углеводов типы связывания (замещение обоих протонов аминогруппы неалкильными заместителями, C-C-связь между углеводными остатками, C-C-связь в дезоксигенированном положении или в точке замыкания цикла, более двух связей у одной фосфатной группы, более одной C-C-связи в одно положение, связь остатков через их аминогруппы и т.д.)<br>количество связей с заместителями в одном и том же положении остатка не соответствует типу атома;<br>список альтернативных фрагментов содержит единственную альтернативу;<br>количество заместителей в неустановленных позициях превышает число свободных позиций в остатке;<br>в качестве способа синтеза указан неизвестный способ;<br>наличие в записи структуры неподдерживаемых топологических особенностей (альтернативные ветви с тройной вложенностью, альтернативные фрагменты на восстанавливающем конце, нестехиометрическое присутствие корневого остатка, более двух связей между одними и теми же остатками) | |
| | химически возможный, но неправильный мономерный состав;<br>химически возможные, но неправильные топология, последовательность, конфигурации остатков или позиции образования связей;<br>неправильные границы биологического повторяющегося звена полимера;<br>химическое повторяющееся звено указано как биологическое или наоборот; | экс-<br>перт |

| | | |
|---|---|---|
| | тривиальное название или класс структуры не соответствует молекуле;<br>нестандартная нумерация атомов в агликонах | |
| библио-<br>графия | нераспознанный тип публикации;<br>нераспознанный тип внешней ссылки;<br>некорректный формат PubMed ID;<br>неполные или нераспознанные выходные данные;<br>название статьи, журнала или книги не совпадает с названием с этими же выходными данными в других записях;<br>неправильно записанное название журнала;<br>нераспознанный формат контактов для переписки;<br>набор библиографических данных не соответствует типу публикации;<br>для этой же книги в других записях указаны другие редакторы или издательство | выяв-<br>ление |
| | указанный NLM ID непериодического издания относится к журналу;<br>ошибки в написании авторов (в том числе в национальных символах);<br>опечатки в названиях статей, аффилиациях, рефератах;<br>набор ключевых слов не соответствует авторскому;<br>возможные, но неправильные выходные данные;<br>внешняя ссылка недоступна или ведёт не на ту публикацию;<br>неправильная локализация структуры в статье;<br>неправильно записанное название книги | экс-<br>перт |
| таксо-<br>номия | в качестве основного указан более высокий ранг, чем род;<br>для одних и тех же таксонов в разных записях указан разный NBCI Tax ID;<br>не указан таксономический тип; | про-<br>грамма |
| | нераспознанный формат описания таксонов;<br>нераспознанный формат описания синонимов таксонов или противоречивые синонимы таксонов из разных записей;<br>несуществующие или неправильно записанные наименования родов и видов (не имеют NCBI Tax ID);<br>наличие неподдерживаемых таксономических описаний (вложенные переименования, тройные гибриды); | выяв-<br>ление |

| | | |
|---|---|---|
| | указан идентификатор неопределённого вида (sp.), но не указан штамм<br>количество или значение NBCI Tax ID не соответствует указанным таксонам;<br>множественное указание организмов с одинаковыми NCBI Tax ID;<br>указанный тип или царство не существуют или не соответствуют роду;<br>указано несколько типов или царств, но их количество не соответствует числу таксонов;<br>в качестве организма-хозяина указан прокариотический организм<br>указанное наименование таксона соответствует более чем одному таксону в NCBI Taxonomy;<br>неоднозначное соответствие списка таксонов и списка NCBI Tax ID;<br>неоднозначное соответствие списка таксонов и списков типов или царств;<br>NCBI Tax ID гибрида указан для негибридного организма или наоборот | |
| | указан NCBI Tax ID более высокого ранга, чем таксон, хотя для данного таксона существует точный NCBI Tax ID;<br>некорректное наименование штамма или серогруппы;<br>неправильный орган, ткань, стадия развития;<br>возможный, но неправильный род или вид<br>возможный, но неправильный организм хозяин и/или заболевание; | эксперт |
| ЯМР | некорректный порядок хим. сдвигов для протонов при одном атоме углерода | программа |
| | некорректный формат описания спектра;<br>химические сдвиги, не попадающие в характерную область;<br>число сигналов в отнесении остатка не соответствует числу атомов в остатке;<br>количество или наименования остатков в отнесении спектра не соответствуют структуре;<br>локализация остатка в отнесении не соответствует ни одному остатку с таким именем в структуре;<br>в структуре присутствует более одного остатка с указанной локализацией, но ни один не соответствует наименованию;<br>пара «наименование-локализация» не уникальна; | выявление |

| | | |
|---|---|---|
| | не указана или не распознана температура или растворитель экспериментов ЯМР<br>температура образца ЯМР не укладывается в характерный диапазон значений<br>не распознано соотношение растворителей или pH образца | |
| | неправильные хим. сдвиги, тем не менее попадающие в характерную область (для их выявления написана отдельная программа для статистического поиска отклонений с учётом моделирования сигналов каждого атома);<br>для некоторых остатков химические сдвиги отсутствуют, хотя опубликованы;<br>не указан опубликованный pH образца ЯМР | эксперт |
| запись | в записи отсутствует обратная ссылка на цитирующую её запись | программа |
| | указанная запись (комбинация структуры и статьи) присутствует в базе под другим номером в сочетании с другими данными;<br>некорректное количество полей или отсутствие критически важных полей;<br>некорректные или повторяющиеся идентификаторы;<br>запись помечена, как неиспользованная, но причина не указана;<br>внешняя ссылка содержит идентификатор неизвестной базы;<br>внутренняя ссылка ведёт на несуществующую запись;<br>некорректные вспомогательные данные (аннотатор, дата и т.д.) | выявление |
| | запись не соответствует контексту статьи (напр., «лишняя» структура)<br>запись, соответствующая контексту статьи, отсутствует в базе;<br>запись помечена как неиспользованная по причине, не следующей из контекста статьи;<br>неправильные методы установления структуры (некорректный синтаксис выявляется отдельной программой, смысловые ошибки – только экспертом);<br>ссылка на внешнюю базу ведёт не на ту запись | эксперт |
| характеристики | стерео-конфигурации атомов отсутствуют или противоречат данным MSDB | программа |
| мономеров (вспо- | неполный набор дескрипторов;<br>протонирование не соответствует типам атомов;<br>позиция связи по умолчанию не соответствует типу атома; | выявление |

| | | |
|---|---|---|
| мога-тельный дамп) | невозможная комбинация базового остатка и модификаций;<br>длина дескрипторов не соответствует числу атомов;<br>дескрипторы указаны для невозможного размера цикла;<br>дескрипторы разных циклических форм имеют разную длину;<br>длина дескриптора протонирования не соответствует числу атомов | |
| | непротиворечивые, но неправильные дескрипторы;<br>неправильная стереоконфигурация атома или двойной связи;<br>возможный, но неправильный код WURCS;<br>возможный, но неправильный код SMILES;<br>неправильное текстовое описание (комментарий) остатка | экс-перт |
| журна-лы (вспо-мога-тельный дамп) | некорректное число полей;<br>нераспознанный формат NLM ID | выяв-ление |
| | некорректное название журнала;<br>журнал описан как книга;<br>некорректная аббревиатура названия журнала;<br>некорректное издательство | экс-перт |
| взаимо-дей-ствие с базой на уровне СУБД | Каждое действие, затрагивающее данные в базе, отслеживается на предмет корректности запроса SQL и соответствия количества возвращаемых результатов (0, 1, 2, много) смыслу запроса. Если запрос или результат некорректен, это свидетельствует либо о наличии новых типов ошибок в данных, проверка которых ещё не реализована, либо об ошибках в коде программ управления CSDB. Во всех подобных случаях проводится ручная проверка данных и кода и выяснение причин ошибки. | выяв-ление |
| сервер | Технические ошибки среды (ошибки во вспомогательных файлах, ошибки файловой системы, ошибки настройки веб-сервера, недостаточные права доступа, недостаточные аппаратные ресурсы, нет связи с внешними базами данных, аппаратные сбои и сбои операционной системы и т.д.) | выяв-ление |

Практически во всех современных углеводных базах, включая CSDB, информация, опубликованная до 1996 года, импортирована из CCSD (CarbBank) [62]. В этом контексте выявление и исправление ошибок CarbBank представляет чрезвычайно важную задачу для всей области знания. В процессе импорта CarbBank в CSDB были автоматически проверены все записи и вручную – около половины записей. Соотношение типов выявленных и исправленных ошибок в

первичной структуре или таксономии получено на случайной выборке из дампа CCSD и приведено в Табл. 9. Доля ошибок является оценкой снизу, так как на основании отбора кандидатов на проверку по публикациям не все проверяемые записи были признаны подозрительными. Доля ошибок указана по отношению к общему количеству проверенных записей. Более подробно с этим разделом можно ознакомиться в публикации [42].

Табл. 9. Выявленные ошибки CarbBank

| *домен* | *грибы* | *бактерии* | *растения* |
|---|---|---|---|
| Проверено записей | 857 | 301 | 464 |
| Проверено записей по оригинальным публикациям | 498 | 223 | 263 |
| Записей с одной или более ошибками | 35% | 49% | 19% |
| Неправильный штамм | 22% | 26% | 2.1% |
| Структуры нет в статье | 1.6% | 3% | 0.2% |
| Неправильная структура | 2.1% | 10% | 5% |
| Предполагаемая структура указана как определённая | 2.1% | 2.0% | 5% |
| Неправильный организм | 2.2% | 9% | 9% |
| Структура из статьи отсутствует в базе (недостающие структуры / проверено статей) | 60/107 | 33/68 | 191/68 |

### *3.2. Поиск данных*

Поиск данных позволяет перейти от известных данных к другим связанным с ними данным. Общая схема таких переходов в CSDB показана на Рис. 17. Поисковые запросы различных типов можно проводить по всему покрытию базы либо комбинировать при помощи логических операторов И (AND, искать в результатах предыдущего запроса), ИЛИ (OR, объединить с результатами предыдущего запроса), НЕ (NOT, получить все результаты, кроме удовлетворяющих запросу) и И НЕ (AND + NOT, вычесть результаты, удовлетворяющие запросу, из результатов предыдущего запроса). Особенности работы поисковой системы при более чем двукратном комбинировании запросов разного типа можно изучить в справочной системе проекта[a]. Пример комбинирования разнородных критериев в запросе показан на Рис. 18.

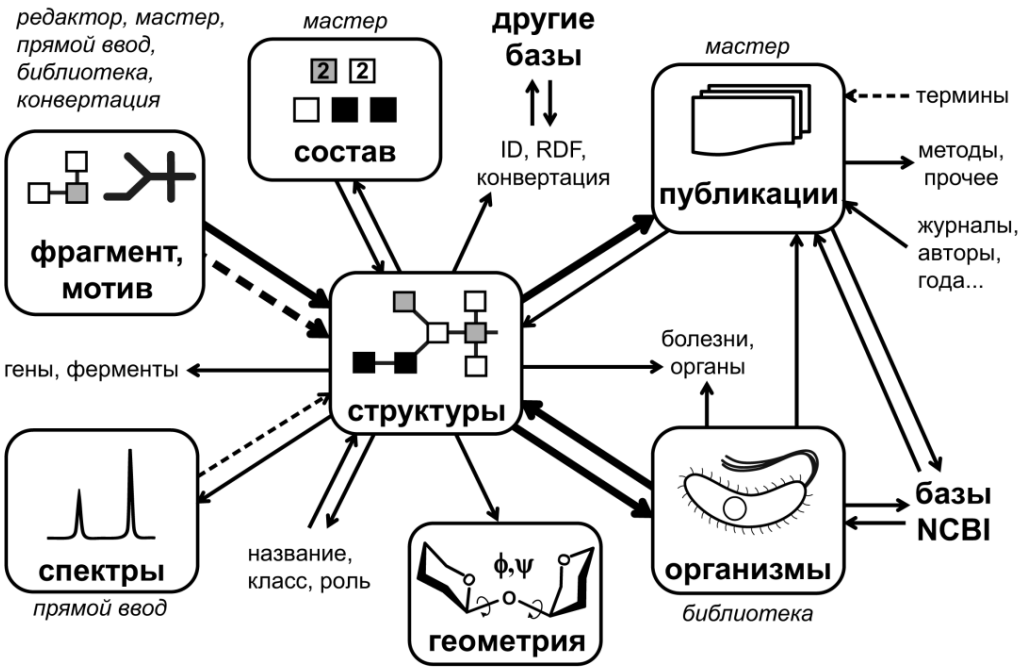



Рис. 17. Типы данных CSDB и возможные переходы между ними (толстые стрелки – наиболее распространённые переходы, тонкие – прочие переходы, пунктир – переходы с нечёткой логикой). Способы ввода приведены курсивом рядом с типами наиболее распространённых данных.

[a] http://csdb.glycoscience.ru/help/usage.html#scope

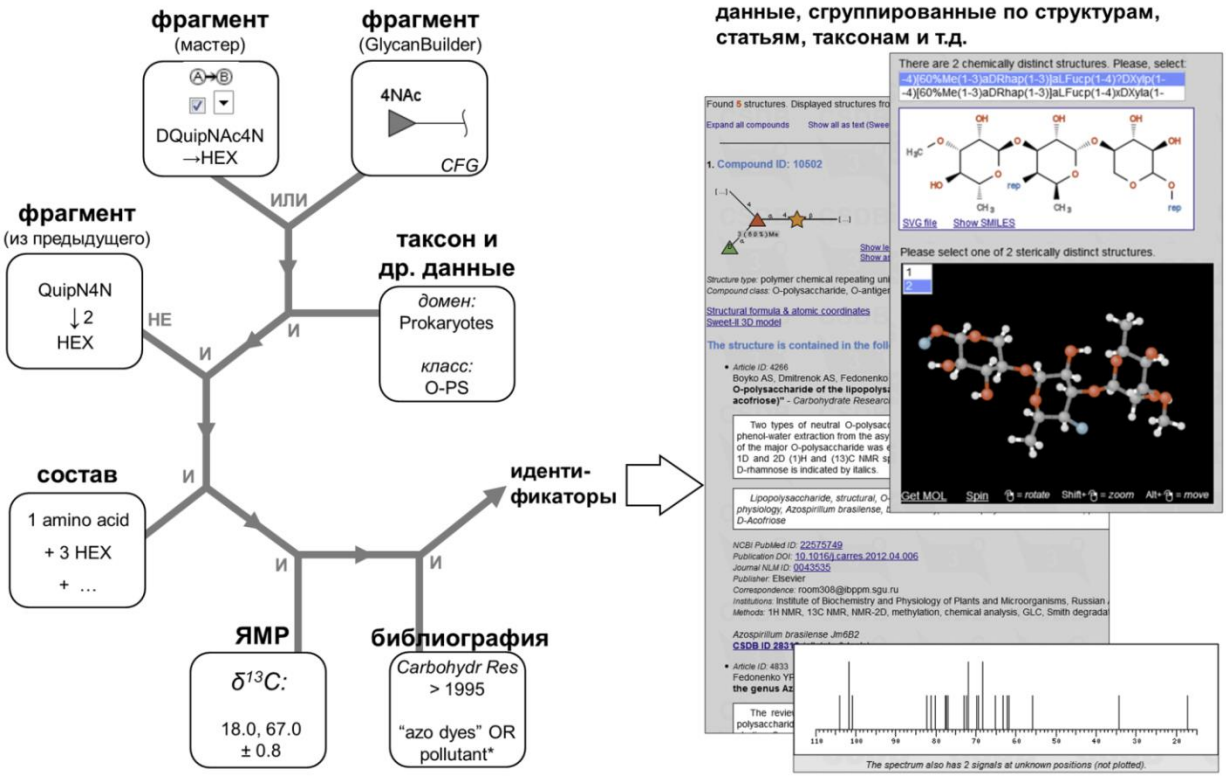



Рис. 18. Пример комбинирования разнородных критериев поиска.

Результаты поиска объектов любого типа сопровождаются: ссылками на записи, из которых можно получить полную информацию, связанную с объектом поиска; инструментами навигации, сортировки и уточнения результатов; инструментами работы со структурой (перевод на другие семантические языки, моделирование молекулярной геометрии и спектров ЯМР и др.) и набором индексов, перечисленных в Табл. 10, для идентификации объектов в других базах данных.

Табл. 10. Внешние стандартные индексы в CSDB

| *Объект* | *Индекс* | *База данных* | *Статус и комментарии* |
|---|---|---|---|
| Первичная структура | GlycomeDB ID; CA registry number; CA access number | GlycomeDB; Chemical Abstracts | Поддерживается, но ссылки присутствуют лишь в отдельных записях. |
| | CCSD ID | CarbBank | Поддерживается и заполнено для записей до 1996 года. |
| | GTC ID | GlyTouCan | Планируется с автозаполнением. |

| | | | |
|---|---|---|---|
| Остаток | MSDB ID | MonosaccharideDB | Поддерживается и заполнено для моносахаридов. |
| Конформация | PDB ID | Protein Data Bank | Анализируется целесообразность для гликанов. |
| Таксон | Tax ID | NCBI Taxonomy | Поддерживается и заполнено для всех объектов. |
| Заболевание | ICD-11 code | ICD | Планируется. |
| Публикация | DOI | Система Digital Object Identifier | Поддерживается и заполнено для большинства объектов. |
| | PMID | NCBI PubMed | |
| | NLM ID | NCBI NLM Catalog | Полная поддержка для книг. Для журналов – сохраняется в базе, но не доступно пользователю. |
| | сетевой URL | *Интернет* | Поддерживается, но ссылки присутствуют лишь в отдельных записях. |
| | USPT | USA Patent | |
| Информация из статей | термин MeSH | NCBI Medical Subject Headings | Планируется для методов исследования и ключевых слов. |
| Спектр ЯМР | CCPN ID | Collaborative Computational Project for NMR | Анализируется целесообразность. |
| Ген и фермент | GenBank ID; Uniprot ID | NCBI Gene; NCBI Enzyme; Uniprot | Поддерживается и заполнено для всех активностей гликозилтрансфераз. Не заполнено для углеводных структур в основной части CSDB. |

Веб-интерфейс базы предоставляет пользователям возможности поиска по следующим критериям:

1. *Структуры, их фрагменты и номенклатура (Substructure).* Обязательной частью запроса является описание структурного фрагмента на языке CSDB Linear. Его можно ввести: в графической форме с помощью визуального инструмента, работающего в формате SNFG; с помощью мастера, позволяющего «собрать» структуру посредством визуальных операций мышью;

путём прямого ввода на языке CSDB Linear либо редактирования структуры, сгенерированной мастером; путём копирования и редактирования предыдущего запроса; путём выбора из библиотеки поименованных распространённых углеводных структур; путём трансляции с языка GlycoCT. База данных возвращает структуры, содержащие заданный фрагмент (с учётом неопределённостей в структурах и сдвига рамки полимеризации) либо совпадающие с заданным фрагментом с указанным уровнем строгости сравнения. Возможна предварительная фильтрация результатов по типу молекулы, по классам соединений (функциям в клетке), по наличию данных ЯМР, по биологической привязке на уровне царств. Также поддерживается поиск указанных фрагментов текста в записях структур на CSDB Linear, в названиях агликонов и в тривиальных названиях соединений. Результатом поиска являются соответствующие запросу структуры, а также сопутствующая структурная информация и список публикаций, в которых они описаны, с привязкой к таксономии.

5. *Мономерный состав* (*Composition*). Этот критерий позволяет найти структуры с указанным полным или частичным мономерным составом. В качестве единиц могут использоваться не только имена остатков, но и их суперклассы (напр., «гексоза» или «аминокислота»). Результат и способы его фильтрации аналогичны характеристикам структурного поиска.
6. *Биологическая привязка* (*Taxonomy*). Этот критерий позволяет перейти от организмов, из которых выделены структуры, к самим структурам, а также получить записи, соответствующие всем микроорганизмам, инфицировавшим указанный организм-хозяин. Полная или частичная таксономия вводится путём последовательного сужения области поиска с помощью четырёх меню: домен, род, вид и штамм. При изменении критерия более высокого ранга возможности выбора более низких рангов пересчитываются. Подвиды и серогруппы относятся к рангу «штамм» и могут быть введены в свободнотекстовом виде, в том числе с поддержкой символов-заменителей. Альтернативным способом идентификации таксона является прямой ввод NCBI Tax ID с указанием использовать только этот таксон либо его и все включённые в него таксоны более низких рангов. Результатом поиска яв-

ляются соответствующие запросу организмы вместе с сопутствующей биологической информацией и списком выделенных из них структур, с привязкой к библиографии.

7. *Библиография* (*Bibliography*) позволяет найти статьи, главы и книги, удовлетворяющие заданным условиям, чтобы затем перейти к опубликованным структурам и организмам. Поисковый запрос учитывает любую комбинацию следующих критериев: фамилии и инициалы авторов, наличие заданного термина или его производных в названии, реферате или ключевых словах, название журнала, ограничения на год публикации (больше, меньше или равен указанному), номер тома и номер страницы. Для ввода авторов используется авторский указатель, фильтруемый по началу фамилии. Текст для поиска в названиях, рефератах и ключевых словах обрабатывается с поддержкой логических операций И, ИЛИ и НЕ, группировкой операций с помощью скобок, неразделимых терминов (в кавычках) и символов-заменителей (* и ?). Более подробно про язык запросов можно прочитать в справочной системе проекта [a]. Возможна предварительная фильтрация результатов по факту установления новой структуры в публикации и по биологической привязке на уровне царств. Результатом поиска являются соответствующие запросу публикации и их метаданные, а также список описанных в них структур, с привязкой к таксономии.
8. *Сигналы ЯМР* (*NMR signals*). Критерием поиска являются химические сдвиги $^{1}$H или $^{13}$C. Этот вид поиска позволяет найти структуры, ЯМР-спектры которых содержат искомые сигналы с указанным уровнем похожести. Возможна фильтрация по признаку нахождения отнесения сигналов в пределах одного остатка. Результат аналогичен структурному поиску и содержит дополнительные ЯМР-спектроскопические данные, включая таблицы отнесения сигналов и числовую метрику соответствия экспериментального спектра ЯМР его указанному фрагменту.
9. *Идентификаторы (CSDB IDs).* Этот тип поиска позволяет найти записи, структуры, публикации, организмы и спектры по их идентификаторам в

[a] http://csdb.glycoscience.ru/help/usage.html#query

случае, если идентификаторы или их диапазоны известны заранее. Результаты могут быть представлены как в виде веб-страниц, аналогичных другим видам поиска, так и в виде формализованной выдачи для автоматической обработки на одном из пяти распространённых языков кодирования данных, применяемых в том числе в биохимии (Thomson Reuters DCI, RDF Turtle, RDF N-triples, RDF JSON, RDF XML).

С пользовательским интерфейсом поисковых запросов и представлением их результатов можно ознакомиться в справочной системе на сайте проекта[a] и в публикациях автора [18, 64, 257].

Поиск активностей гликозилтрансфераз реализован в виде отдельного модуля [74], позволяющего перейти как к описанным гликозилтрансферазам, синтезируемым ими структурам, используемым донорам и субстратам и другим данным, присутствующим в CSDB, так и к записям фермента и кодирующего его гена в протеомных и геномных базах данных. Взаимосвязь поисковых критериев с получаемым результатом схематично представлена на Рис. 19.


Рис. 19. Получение информации о биосинтезе углеводов с помощью модуля CSDB GT.

[a] http://csdb.glycoscience.ru/help/usage.html ; http://csdb.glycoscience.ru/help/examples.html

## *3.3. Описание углеводных структур*

### *3.3.1. Кодирование структур*

В качестве промежуточного редактируемого хранилища аннотаций и резервной копии, из которой импортируется база CSDB, используется текстовый дамп, содержащий данные из статей в формализоавнном виде. В процессе работы над организацией заполнения дампа и операций ввода-вывода были проанализированы существующие языки описания углеводов (нотации) и разработана нотация CSDB Linear (линейный код CSDB), лишённая недостатков конкурентов. Сравнение CSDB Linear и существующих углеводных и общехимических языков приведено в Табл. 3 в разделе 2.3.

Линейный код CSDB представляет собой однозначный, полный, машино- и человекочитаемый язык описания биогликанов, разработанный в рамках проекта для записи в строковой форме углеводов и их производных, включая экзотические случаи. С точки зрения пользователя знание этого языка требуется для создания сложных запросов, добавления данных в CSDB и взаимодействия с другими базами. В рамках интерфейса CSDB возможен перевод с CSDB Linear на языки IUPAC Extended (для визуализации), SNFG (для визуализации; совместим с нотацией CFG), WURCS (для взаимодействия с другими базами), SMILES (для перехода к атомарному описанию и молекулярной геометрии), GLYCAM (для моделирования конформаций), в формат MOL (для описания связности атомов и молекулярной геометрии), а также обратный перевод с языка GlycoCT, используемого многими другими проектами [46], на CSDB Linear. Уровень поддержки различных структурных особенностей, характерных для биогликанов, подытожен в колонке «Поддержка в CSDB Linear» Табл. 13.

Нотация CSDB Linear использует распространённый в гликоинформатике подход, кодирующий природные молекулы в виде направленных графов, в которых остатки соответствуют вершинам, а связи между ними – рёбрам. Эти графы записываются в виде одной строки текста и используются для аннотирования структур в дампе, операций импорта и экспорта структур, контроля ошибок и прямого ввода сложных структур пользователем. При этом собственно в базе структуры хранятся в нечеловекочитамом внутреннем формате, более удобном для машинного поиска.

Структура биогликанов и родственных соединений в большинстве случаев подразумевает мономерные остатки, связанные друг с другом с отщеплением воды. На уровне мономеров CSDB Linear использует контролируемый словарь из 486 базовых наименований остатков (`Glc`, `Gal` и т.д.), встречающихся в биогликанах. Эти наименования комплектуются жёстко типизированными префиксами (аномерная и абсолютная конфигурация) и суффиксами (размер цикла – пираноза, фураноза, линейная форма – или признак полиола; модификации, связанные с окислением или функционализацией).

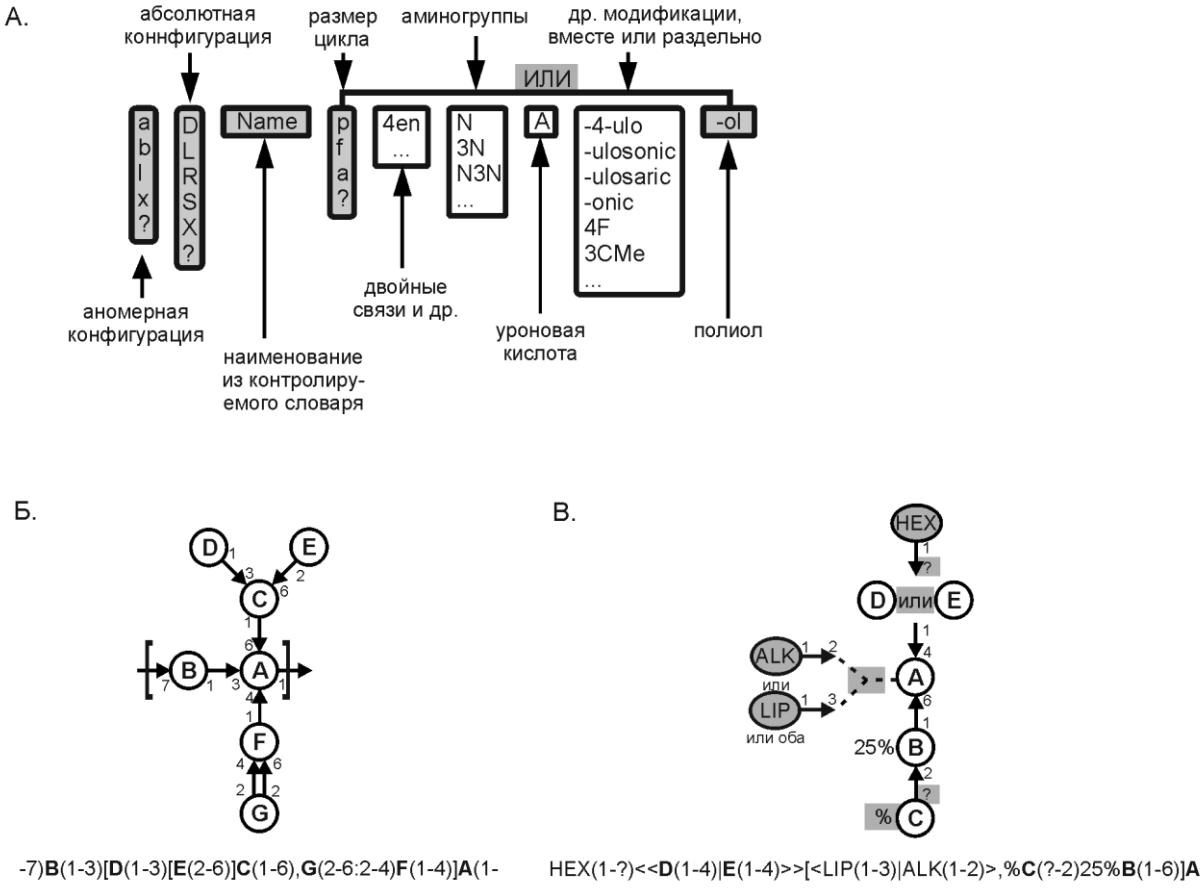



Рис. 20. Возможности нотации CSDB Linear. **А.** Компоненты кодировки мономерных остатков (обязательные показаны серым). **Б.** Пример кодировки топологии и связности (A,B – основная цепь полимера; A,C – точки разветвления; E,D,G – терминальные остатки; G и F связаны двумя связями). **В.** Пример кодировки неоднозначностей вне остатков: суперкласс (HEX = любая гексоза), неизвестные позиции замещения (?), альтернативные фрагменты (D и E); альтернативные фрагменты и связи (ALK и LIP); нестехиометрические компоненты (B,C).

Признак дезоксигенирования кодируется в базовом имени остатка, так как приводит к либо изменению стереодескрипторов (напр., глюкоза, дезоксигенированная по C2 теряет один хиральный центр и становится арабиногексозой) либо к остаткам, имеющим устоявшиеся тривиальные названия (Rha, Fuc, Qui и др.). Допустимые значения префиксов и суффиксов показаны на Рис. 20A. В случае, если остаток имеет единственный асимметрический центр, в качестве абсолютной конфигурации используется R/S, в остальных случаях - D/L. Если модификация приводит к изменению количества асимметрических центров, базовое имя остатка заменяется на соответствующее новой стереохимии. Конфигурации и размер цикла могут быть неопределёнными (знак «?»). В случае ахиральных одновалентных остатков, напр., `Ac` (уксусная кислота) конфигурации не являются обязательными. Комбинирование перечисленных признаков приводит к интуитивно понятному написанию имён остатков, близкому к историческому, а жёсткая типизация и контролируемый словарь мономеров позволяет сохранить машиночитаемость и однозначность интерпретации, а также облегчить контроль ошибок ввода. *Примеры*: `aDTalfA` = α-D-талофуранозуроновая кислота; `?XKdop` = кетодезоксиманноктоновая кислота в пиранозной форме с неизвестной аномерной конфигурацией; `xLGro` = L-глицерин; xDManN-ol = 2-амино-D-маннит; `Ac` = уксусная кислота; `xRPyr` = пируват, приобрётший при связывании новый стереоцентр в конфигурации R; `bDFuc?N3N` = 2,3-диамино-β-D-фукоза в неизвестной (а)циклической форме. Словарь мономеров, их классификация, атомарные свойства и систематические названия по IUPAC доступны в виде специального сервиса на сайте проекта[a]. Возможные значения префиксов и суффиксов-модификаторов приведены в Табл. 11.

[a] http://csdb.glycoscience.ru/database/core/residues.php

Табл. 11. Префиксы и суффиксы в наименовании мономерных остатков

| *Поле* | *Значение* | *Объяснение* |
|---|---|---|
| Аномерная конфигурация | a | циклизованный α-аномер |
| | b | циклизованный β-аномер |
| | ? | моносахарид с неизвестной аномерной конфигурацией |
| | l | алифатический остаток с карбоксильной группой в положении 1 (в основном, жирные кислоты) |
| | x | прочие неуглеводные остатки или моносахарид в открытой форме |
| Абсолютная конфигурация | D | D-форма мультихирального остатка |
| | L | L-форма мультихирального остатка |
| | R | остаток, имеющий единственный ассиметрический R-центр |
| | S | остаток, имеющий единственный ассиметрический S-центр |
| | X | ахиральный остаток либо остаток, хиральность которого закодирована в базовом имени (напр., Kdo или DLmanHep) |
| | ? | хиральный остаток с неизвестной абсолютной конфигурацией |
| Размер цикла | p | пираноза |
| | f | фураноза |
| | a | моносахарид в линейной форме |
| | ? | неизвестный способ циклизации либо неуглеводный остаток, базовое имя которого заканчивается символом «p», «f» или «a». |
| Основные модификаторы (в любых комбинациях) | A | уроновая кислота |
| | N | аминогруппа в положении 2 |
| | #N | аминогруппа в положении # |
| | #en | двойная связь в положении # |
| | #-ulo | кето-группа в положении #, отличном от 2 |
| | #-ulosonic | #-улозоновая кислота |
| | #-ulosaric | #-улозаровая кислота |
| | #S | тиогруппа в положении # |
| | #CMe | C-связанная метильная группа в положении # |
| | #CHm | C-связанная гидроксиметильная группа в положении # |
| | #F | фтор в положении # |
| Признак полиола | -ol | полиол (моносахарид восстановленный по аномерному атому) |

На топологическом уровне (Рис. 20Б) одна цепь остатков всегда является основной (в полимерах её выбор очевиден, для олигомеров разработаны правила сортировки цепей по старшинству). Боковые цепи, присоединённые к остаткам основной цепи, перечисляются через запятую в квадратных скобках слева от остатка. Боковые цепи могут иметь собственные боковые цепи. На уровне связности (Рис. 20Б) каждый остаток, кроме восстанавливающего конца олигомера (корня графа), снабжён дескриптором исходящей из него связи (в круглых скобках), состоящим из пары связанных позиций в соответствии с нумерацией атомов в остатках. Правила определения донора и акцептора близки к устоявшимся в гликохимии. Донор всегда записывается слева от акцептора; любой остаток может иметь один или ни одного акцептора и произвольное число доноров (включая ноль для терминальных остатков). На Рис. 20 (Б и В) остатки для краткости показаны латинскими буквами, упрощённая кодировка CSDB Linear приведена под схемами. Полимерные структуры кодируются в виде повторяющегося звена с двумя «висящими» связями. *Примеры*: линейный трисахарид A(1-3)B(1-4)C, разветвлённый трисахарид A(1-3)[B(1-4)]C, трисахаридное повторяющееся звено -6)A(1-3)B(1-4)C(1-. На уровне связности (см. Табл. 12) предусмотрен специальный синтаксис для остатков фосфорной и серной кислот (напр., `xXEtn(1-P-6)??Glc?` для фосфоэтаноламина, замещающего C6 глюкозы), для моновалентных агликонов на восстанавливающем конце (напр., `bDGlcp(1-Me` ), для C-C-связей между остатками (напр., в C-гликозидах), для связей в нестандартные положения, не кодируемые нумерацией углеродного скелета (напр., 3’), для второй связи с тем же остатком (напр., если донор - пируват или бифосфат). Если углеродный скелет остатка имеет части, не соединённые C-C-связями, это позволяет предположить, что остаток образован двумя остатками с отщеплением воды. Несмотря на несоответствие исторически сложившемуся подходу, применённое в CSDB Linear описание такой системы как структурного фрагмента из двух остатков (напр., в случае N-ацетилглюкозамина: 1-2-связанные остатки уксусной кислоты и глюкозамина `Ac(1-2)bDGlcpN`, но не единый остаток `bDGlcpNAc`) позволяет лучше формализовать первичную структуру молекул и на два порядка снизить объем требуемого словаря мономеров.

Табл. 12. Кодировка связей между остатками

| *Синтаксис (X, Y – остатки; # - числа)* | *Объяснение* | *Примеры* |
|---|---|---|
| X(1-#)Y<br>X(2-#)Y | Связь между остатком-донором X и положением # остатка Y | aDGlcp(1-4)bDManp; Ac(1-2)bDGlcpN<br>(альдозные и неуглеводные доноры);<br>aXKdo(2-6)bDGlcp; xLLys(2-6)bDGlcpA<br>(доноры-кетозы и N-связанные аминокислоты) |
| X(#-#')Y<br>X(#-#")Y | Связь в нестандартное положение акцептора | bDGlcp(1-5')Subst // Subst = oxymarmesin<br>(связь в положение 5') |
| X(?-#)Y<br>X(#-?)Y<br>X(?-?)Y | Связь с неустановленным положением в доноре и/или акцепторе | Ac(1-?)[xLLys(?-6)]bDGlcpA<br>(неизвестна локализация ацетата, а также какая из двух аминогрупп лизина образует связь) |
| X(#-P-#)Y<br>P-#)Y<br>X(#-P<br>P-P-#<br>#-P-P-#<br>и т.д. | Фосфо- и сульфодиэфирные связи | xXEtn(1-P-P-6)aXKdop;<br>xDRib-ol(1-P-4)bDGlcp;<br>aXKdop(2-P; S-3)[S-4)]bDGlcp<br>(этаноламиндифосфат, полиолфосфат, фосфат на восстанавливающем конце, терминальные сульфаты) |
| X(1-Y<br>X(2-Y | Связь с одновалентным донором-агликоном | bDGlcp(1-Me<br>(алкильный агликон на восстанавливающем конце) |
| X(#-#:#-#)Y | Две связи между одной и той же парой остатков | xRPyr(2-4:2-6)bDGalp;<br>P(0-3:0-4)aDGlcp<br>(циклические пируват и бифосфат) |
| X(1C-#)Y<br>X(2C-#)Y | Связь без отщепления воды, в которой участвует углеродный атом донора | bD1dGlc(1C-6)Subst // Subst = luteolin<br>C-гликозид: C-C-связь с удалением гидроксильной группы из положения 1 |

| | | |
|---|---|---|
| %X(#-#)Y<br>X(#-%P-#)Y<br>#%X(#-#)Y<br>и др. комбинации | Связь, присутствующая не во всех структурах из выборки | %Ac(1-2)aDGlcp, xXEtN(1-%P-4)aXKdop (нестехиометрическое замещение);<br>-4)[50%Me(1-4)33%bDXyl(1-6)]aDGlcp(1- (ксилоза присутствует в одной трети повторяющихся звеньев, половина остатков ксилозы 4-O-метилирована) |

Значительная доля установленных природных структур содержит неоднозначности (Рис. 20B). В простых случаях они кодируются знаками «?» вместо конфигураций мономеров или положений связей. На уровне мономеров предусмотрены 15 суперклассов[a] (напр., `PEP` = любая аминокислота) и возможность создавать собственные обозначения для неподдерживаемых или неявно заданных остатков. В последнем случае смысл обозначений должен быть расшифрован, напр., `Subst1(1-3)bDGlcp // Subst1 = unidentified 3-deoxy-manno-nonose`. Для обозначения альтернативных фрагментов структуры используются одинарные (для логики «ИЛИ») или двойные (для логики «исключающее ИЛИ») угловые скобки и вертикальная черта, см. пример в кодировке на Рис. 20B. Такие альтернативные наборы остатков, в свою очередь, могут быть замещёнными, в том числе другими альтернативными наборами. Стоит отметить, что для сохранения человекочитаемости и избегания потенциальных неоднозначностей топологии следует минимизировать длину цепей в угловых скобках, т.е. использовать синтаксис вида AB<<C|D>>E, но не A<<BC|BD>>E, где буквы A..E – отдельные остатки. Предусмотрен специальный синтаксис для нестехиометрических боковых цепей и модификаций с установленной или неизвестной стехиометрией (число и знак процента перед названием остатка).

[a] По числу углеродных атомов: TET, PEN, HEX, HEP, OCT, NON; по типу остатка: ANY (любой), MVA (моновалентный), SUG (моносахарид), ALK (спирт), LIP (карбоновая кислота), PEP (аминокислота), SPH (сфингоид), CER (N-ацилированный сфингоид), INO (инозитол).

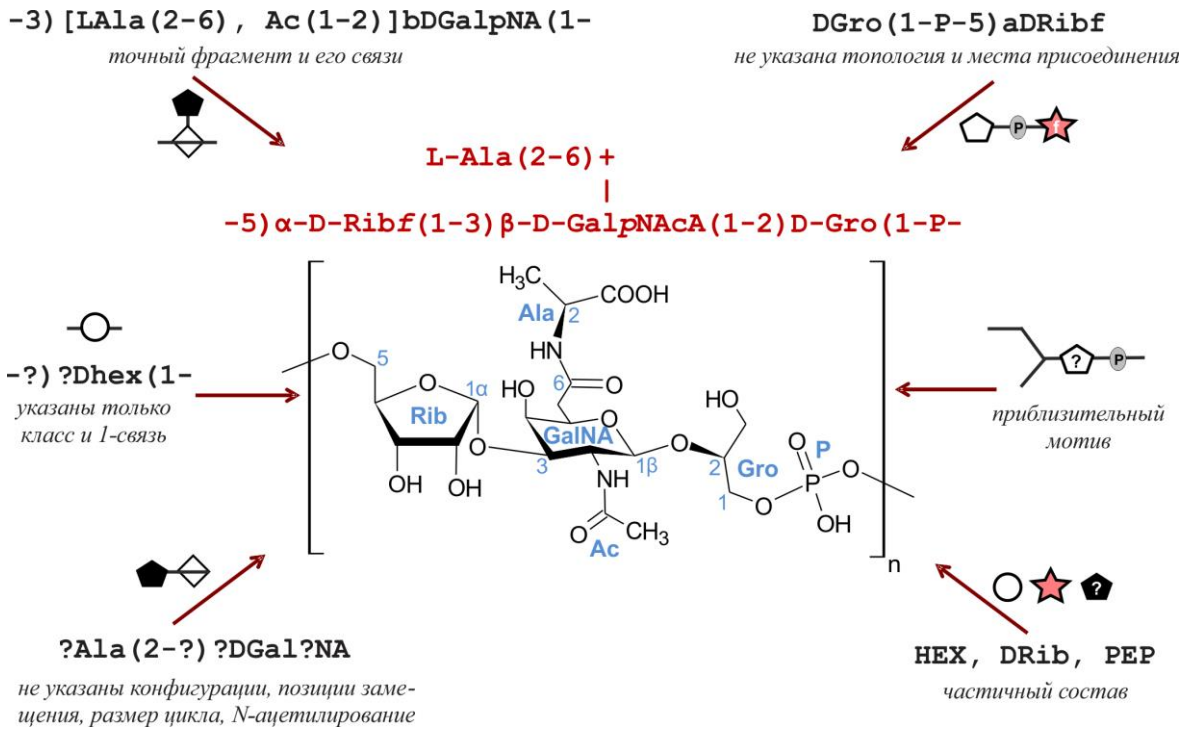
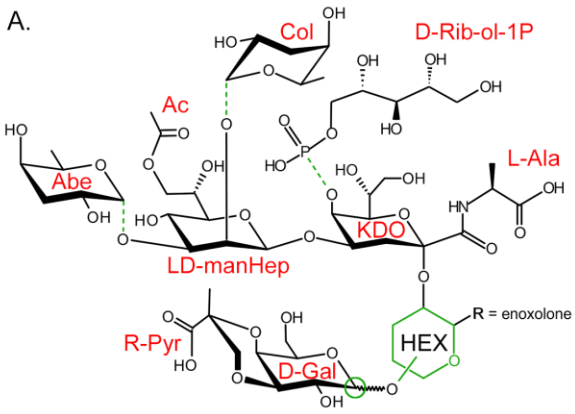



Рис. 21. Уровни абстракции поисковых запросов. Для пяти примеров приведены запросы на языке CSDB Linear и их визуализация в SNFG. Пример «приблизительный мотив» в настоящее время не поддерживается нотацией.

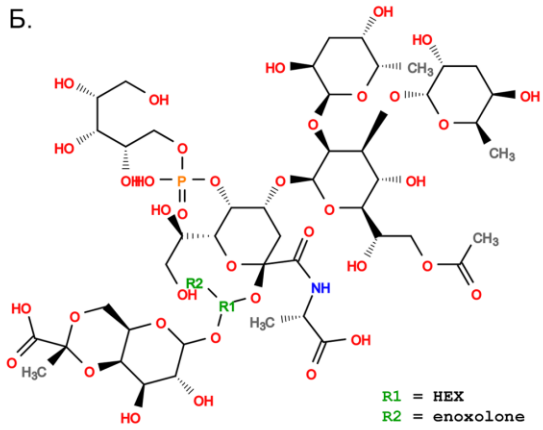
Использование CSDB Linear в поисковых запросах подразумевает произвольное задание уровня абстракции: от точного указания структурного фрагмента, замещённого указанным образом, до указания класса мономерных остатков. Примеры, показанные на Рис. 21, приводят к нахождению модельной структуры. Отдельные структурные особенности не поддерживаются языком CSDB Linear: чередование повторяющихся и уникальных звеньев в пределах молекулы; присоединение цепи к неизвестному узлу топологии; статистическое распределение боковых цепей (гетерогенные полимеры); указание степени полимеризации; циклические полимеры; компоненты структуры, не связанные ковалентно. Эти особенности кодируются в дополнительных полях базы данных; их планируется включить в CSDB Linear по мере его дальнейшего развития. При поисковых запросах дифференциация уровней абстракции, связанная с этими особенностями (например, способ обработки «висячих» связей на границах структурной единицы) реализована с помощью отдельных функций базы CSDB.

Этот раздел содержит тезисное описание языка CSDB Linear. С его функциональной полнотой (списком поддерживаемых структурных особенностей)

А.




Б.




В.

```
<aXAbep(1-3)|aXColp(1-2)>[Ac(1-7)]bXLDmanHepp(1-4)[xDRib-ol(1-%P-5),xLAla?(2-1)]aXKdop(2-
2)[xRPyr(2-4:2-6)?DGalp(1-?)]HEX(1C-24')Subst // Subst=enoxolone
```

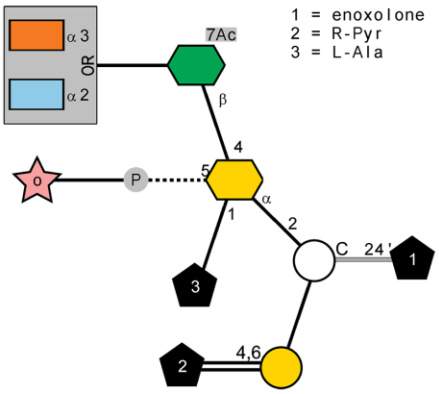
Г.

```
                                R-Pyr-(2-4:2-6)-D-Galp-(1-?)-+
                                                             |
                          D-Rib-ol-(1--P--5)--+              |
                                              |              |
/Variants 0/-L-gro-b-D-manHepp7Ac-(1-4)-a-Kdop-(2-2)-HEX-(1C-24')-Subst
                                              |
                                   L-Ala-(2-1)-+
```

```
/Variants 0/ =
  a-Abep-(1-3)
  OR (inclusively)
  a-Colp-(1-2)

Subst =
  enoxolone
```

Д.




Е.

```
[2*][1*](OC1O[C@@H]2CO[C@@](C)(C(=O)O)O[C@
@H]2[C@H](O)[C@H]1O)O[C@]1(C(=O)N[C@@H](C)
C(=O)O)C[C@@H](O[C@@H]2O[C@H]([C@@H](O)COC
(C)=O)[C@@H](O)[C@H](O[C@H]3O[C@H](C)[C@H]
(O)C[C@H]3O)[C@@H]2O[C@@H]2O[C@@H](C)[C@@H
](O)C[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
```

Ж. 2-O-α-3,6-дидеокси-L-ксилогексопиранозил-(3-O-α-3,6-дидеокси-D-ксилогексопиранозил-)-4-O-7-деокси-7-ацетамидо-L-глицеро-β-D-манно-пиранозил-(5-O-D-рибитол-1-фосфо-)-(1-N-L-аланил)-2-O-α-кето-3-деокси-D-маннокт-2-улозонил-(4,6-O-R-пирувоил-O-галактопиранозил-)-24'-гексозил-еноксолон

Рис. 22. Возможности кодировки и вывода структур в углеводных форматах. А. Гипотетическая структура, содержащая различные типы остатков (показаны красным) и неопределённостей (показаны зелёным), показанная в виде «человекочитаемой» структурной формулы. Пунктиром обозначены связи, которые могут присутствовать или отсутствовать. Б. Автоматически сгенерированная структурная формула. В. Кодировка структуры на языке CSDB Linear. Г. Визуализация структуры в формате SweetDB. Д. Визуализация структуры в формате SNFG. Е. Кодировка структуры в формате SMILES. Ж. Название соединения по IUPAC.

можно ознакомиться в Табл. 13. Более подробно возможности и синтаксис CSDB Linear описаны в справочной системе на сайте проекта[a] и в публикациях автора [64, 70].

Для демонстрации возможностей кодировки на Рис. 22 приведена структура, содержащая множество характерных для биогликанов особенностей, записанная традиционным способом (А), структурная формула, полученная автоматически на веб-сайте CSDB (Б) и её кодировка в нотации CSDB Linear (В). Неопределённые элементы структуры показаны зелёным. Для наглядного сравнения по критерию человекочитаемости эта же структура отображена с применением распространённой человекочитаемой нотации SweetDB[142] (Г), наиболее используемой машиночитаемой нотации SMILES [125] (Е, для варианта с 100% присутствием всех нестехиометрических компонентов) и в виде условно человекочитаемого названия по IUPAC с применением устоявшихся тривиальных названий моносахаридов и других остатков [122] (Ж). По данным опросов пользователей об удобстве отображения первичной углеводной структуры на экране и в статьях учёные-химики отдают предпочтение варианту SweetDB (Г), а учёные-биологи – графическим нотациям (напр., SNFG, Д). Относительно кодировки для ввода и передачи информации о структуре консенсус в научном сообществе не достигнут.

[a] http://csdb.glycoscience.ru/help/rules.html

### *3.3.2. Визуализация структур*

Не менее важной является возможность визуализации структур в виде, привычном гликохимикам и гликобиологам. CSDB представляет три возможности визуализации: семантическое, атомарное (структурная формула) и геометрическое (трёхмерная молекула). В научной литературе, посвящённой природным углеводам, традиционно использовались семантические способы записи: многострочный текстовый формат IUPAC extended [123] или графическая нотация CFG [133]. Для текстового отображения CSDB использует модифицированный формат SweetDB (пример на Рис. 22Г), зарекомендовавший себя в CarbBank [62]. Он представляет собой формализованный псевдографический вариант IUPAC extended, в котором в рамках CSDB был повышен уровень строгости описаний мономеров, а также добавлены другие возможности.

Графическая нотация CFG версии 2, отображающая остатки в виде пиктограмм, была разработана в 1970-х годах Консорциумом по функциональной гликомике, но быстро набрала популярность, особенно в гликобиологических и медицинских статьях. В настоящее время около 30% публикаций по углеводным структурам содержит записи в этом формате. Однако с открытием множества новых моносахаридов и распространением гликомики в область углеводов прокариот возможностей нотации CFG стало не хватать. В рамках работы автор в составе консультативной группы по гликоинформатике при NCBI разработал третью версию нотации Консорциума, получившую название Symbol Nomenclature for Glycans (SNFG, пример для той же структуры см. на Рис. 22Д). В настоящее время её поддержали основные проекты гликоинформатики[a] и рекомендовали к использованию углеводные журналы: Glycobiology, Glycoconjugate Journal, Journal of Biological Chemistry, Journal of Cell Biology, Molecular & Cellular Proteomics, Carbohydrate Research. В 2018 году углеводные структуры в крупнейшей общехимической базе данных PubChem [261] были дополнены визуализацией в формате SNFG в разделе «Biologic depiction»[b]. Кроме того, появи-

[a] https://www.ncbi.nlm.nih.gov/glycans/snfgorg.html

[b] Пример: https://pubchem.ncbi.nlm.nih.gov/compound/44456859

лись компьютерные сервисы для перевода на SNFG с других углеводных языков (в том числе в проекте CSDB).

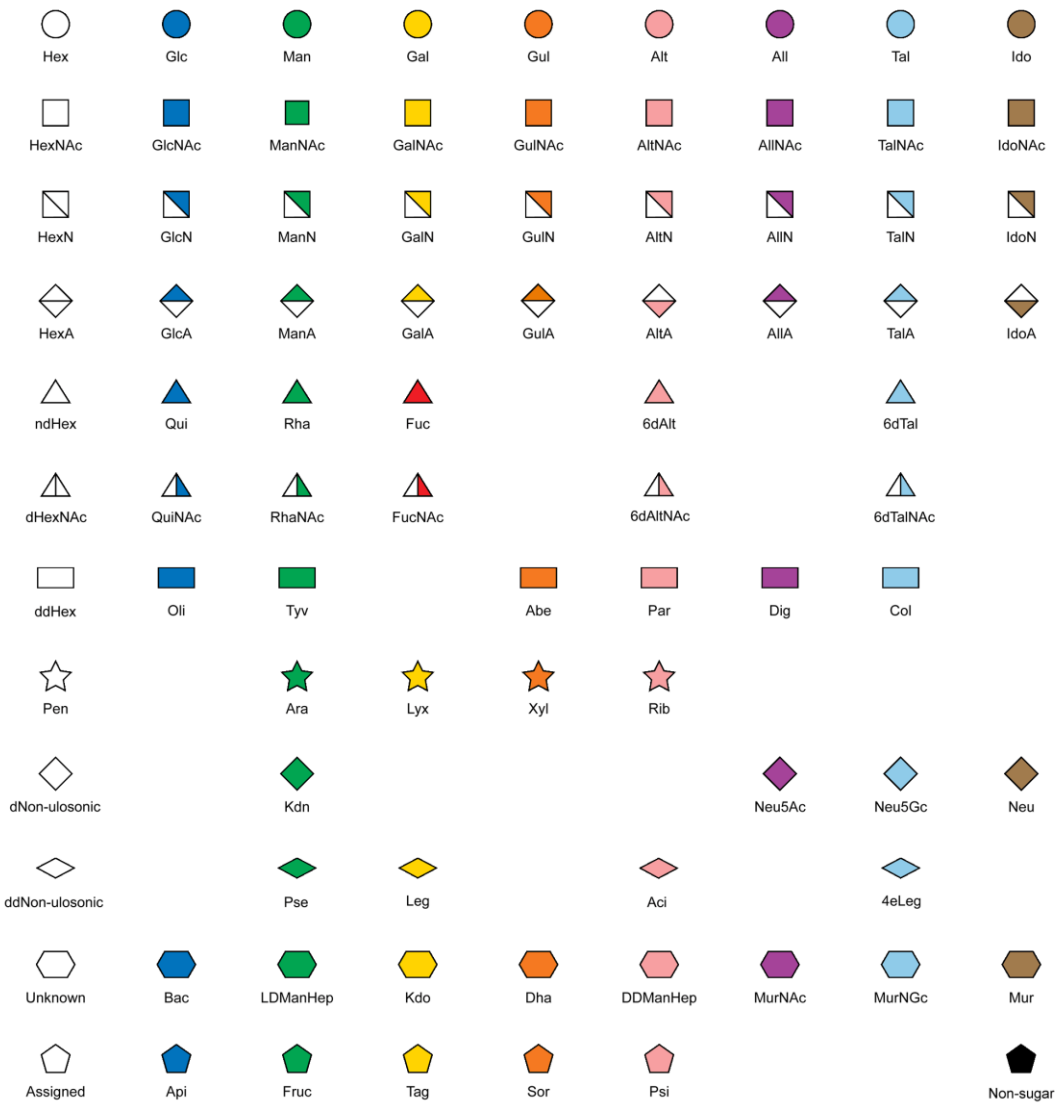



Рис. 23. Пиктограммы, используемые в номенклатуре SNFG [145].

SNFG стандартизирует пиктограммы для 75 моносахаридов и 12 суперклассов (Рис. 23), подобранные так, чтобы обеспечивать обратную совместимость с CFG (т.е. структура, записанная по правилам CFG, является верной и в нотации SNFG). Наиболее распространённые N-ацетилпроизводные представлены как отдельные остатки, остальные модификации моносахаридов указываются текстом рядом с пиктограммами. Для каждого моносахарида существует наиболее распространённая конфигурация по умолчанию, а в случае если абсолютная конфигурация, состояние восстановления или размер цикла отличаются, это указывается символами внутри пиктограмм (напр., «о» означает полиол рибозы (рибит) на Рис. 22Д). Связи между остатками обозначаются линиями, соединя-

ющими пиктограммы, с указанием позиций замещения и аномерных конфигураций. В отличие от CSDB Linear и SweetDB, схема SNFG не является функционально полной, но охватывает наиболее распространённые случаи в химии углеводов и может быть расширена пользователями. Версия SNFG, использованная в CSDB, расширена по отношению к канонической версии для обеспечения возможности визуализировать любые структуры из CSDB, в том числе содержащие неопределённости. С синтаксисом и особенностями SNFG можно ознакомиться на странице NCBI, посвящённой графической номенклатуре углеводов[a], и в публикации автора [134].

[a] https://www.ncbi.nlm.nih.gov/glycans/snfg.html

### *3.3.3. Атомарное описание*

Возможность использования общехимического программного обеспечения для природных углеводов долгое время ограничивал медленный ручной перевод распространенного и доступного в статьях семантического описания в атомарное, требуемое этим программам. На платформе CSDB реализован алгоритм перевода семантических описаний в наиболее распространённый формат хемоинформатики (SMILES [125, 262]) с учётом неопределённостей в структурах. Основные шаги этого процесса представлены на Рис. 24. Все остатки, поддерживаемые в CSDB Linear, сводятся к 943 прототипам (немодифицированным остаткам в D-форме, со свободными аминогруппами и в определённой конфигурации, при её наличии). Атомарные описания этих прототипов, включая каноническую нумерацию атомов, получены заранее и сохранены во вспомогательной базе данных.

Для перевода произвольной структуры в атомарное описание код CSDB Linear подвергается синтаксическому анализу и интерпретируется во внутреннюю структуру в памяти, где каждый остаток представлен отдельным объектом, имеющим номер и содержащим всю сопутствующую информацию о конфигурациях и связях с другими остатками. К каждому такому объекту добавляется код SMILES, полученный модификацией прототипа с учётом конфигураций стереоцентров, отличающихся от прототипа. При этом атомы перенумеруются (Рис. 24, блок 2) так, чтобы в их номерах фигурировали как номера остатков (разряд сотен), так и номера атомов (разряд единиц). Коды SMILES отдельных остатков комбинируются в структуру с помощью виртуальных реакций SMARTS [263] с учётом позиций, в которых каждый из остатков образует связь. Если при этом возникает новый стереоцентр, его конфигурация берётся из исходного семантического описания структуры.

В случае если структура содержит неопределённости, приводящие не более чем к одному рацемическому атому в каждом остатке, конфигурации этих атомов не указываются в результирующем коде SMILES. Во всех остальных случаях структурной неопределённости (когда не определены конфигурации нескольких взаимозависимых стереоцентров, например, неизвестна абсолютная конфигурация моносахарида, либо когда неопределённость подразумевает изо-

мерию на уровне связности атомов) код SMILES не способен описать все возможные структуры, оперируя только конфигурациями стереоцентров. В этих случаях неопределённая структура предварительно превращается в набор определённых структур, для каждой из которых генерируется отдельный код SMILES. Структурные характеристики, встречающиеся в природных углеводах и их производных, как и их поддержка транслятором из CSDB Linear в SMILES перечислены в Табл. 13. Они систематизированы с использованием четырёх уровней детализации: остатков, связей, топологии и неопределённостей. Разработанная схема позволяет адекватно перевести в атомарное описание те из

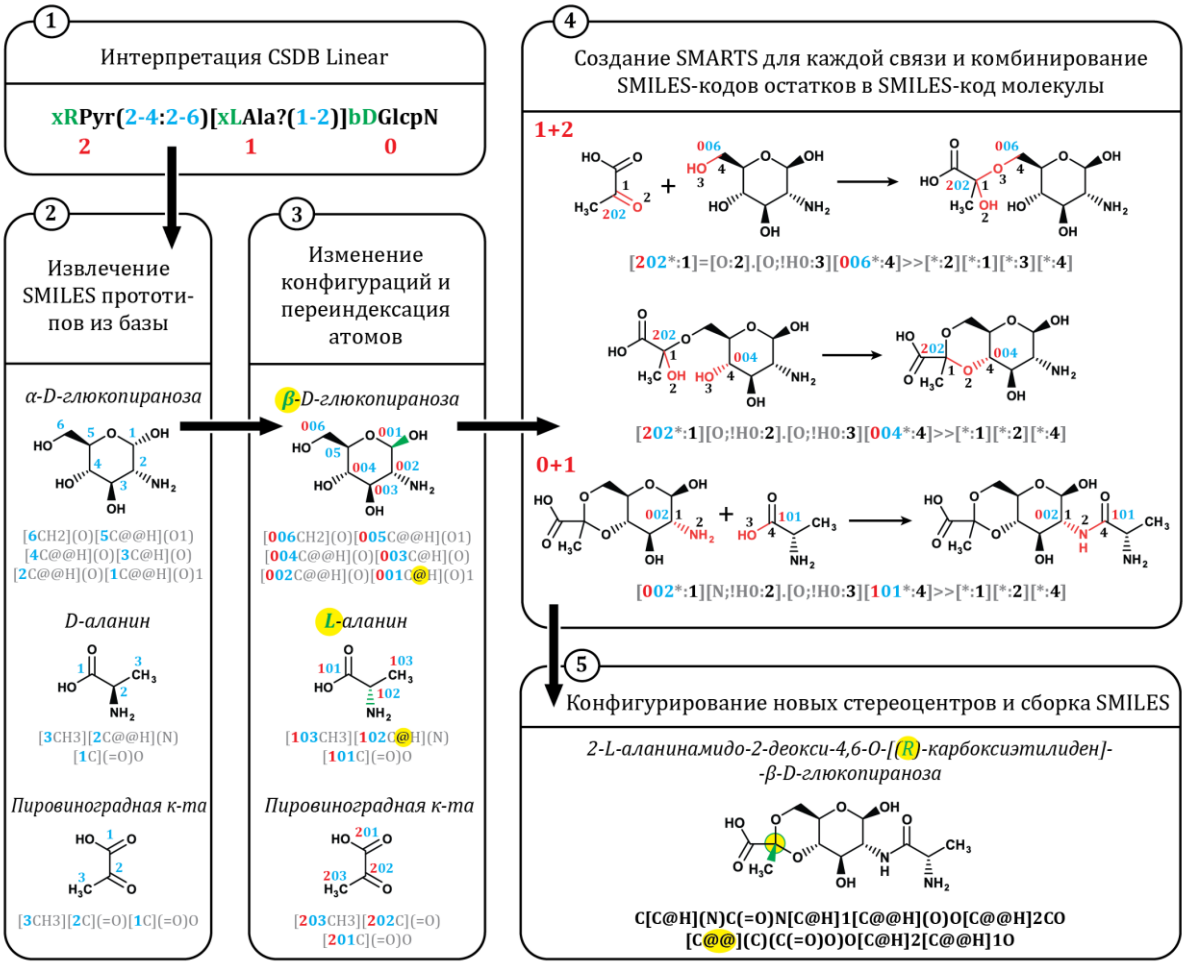



Рис. 24. Получение поатомного описания в SMILES на основе семантического описания CSDB Linear на примере R-пируватизированного β-D-глюкопиранозамина, амидирующего L-аланин. Красными числами пронумерованы остатки, голубыми – атомы в остатках, черными – положения в реакциях SMARTS. Жёлтым обозначены изменяемые стереоконфигурации.

них, что поддерживаются в CSDB Linear. Более подробно с ней можно ознакомиться в публикации автора [146].

Табл. 13. Возможности кодировки CSDB Linear и её перевода в атомарное описание.

| *Особенность* | *Поддержка в CSDB* | *Перевод в SMILES* | *Примеры в нотации CSDB Linear* | *SMILES для первого примера* * | *Комментарии* |
|---|---|---|---|---|---|
| *Уровень остатков* | | | | | |
| Моносахариды | + | + | aDGlcpN,<br>aXKdop,<br>aXLDmanHepp | N[C@H]1[C@@H](O)O[C@H](CO)[C@@H](O)[C@@H]1O | 234 прототипа остатков ** |
| Пиранозы, фуранозы и ациклические формы | + | + | aDGlcp,<br>aDGlcf,<br>aDGlca | OC[C@H]1O[C@H](O)[C@H](O)[C@@H](O)[C@@H]1O | |
| Полиолы, инозитолы и альдоновые кислоты | + | + | xDGro,<br>xDGlcN-ol,<br>xXmyoIno<br><br>myoIno = *мио*-инозитол | OC[C@@H](O)CO | 63 прототипа остатков |
| Фосфаты и сульфаты | + | + | xXEtN(1-P-P-5)[P-4)]aXKdop,<br>aDGlcp(1-P-5)xDRib-ol<br><br>EtN = этаноламин | NCCOP(=O)(O)OP(=O)(O)O[C@@H]1[C@H](OP(=O)(O)O)C[C@](O)(C(=O)O)O[C@@H]1[C@H](O)CO | как терминальные, так и в цепи |
| Жирные кислоты | + | + | lXPam,<br>lX3HOLau<br><br>Pam = пальмитиновая кислота, 3HOLau = 3-гидроксилауриновая кислота | CCCCCCCCCCCCCCCC(=O)O | 167 прототипов остатков (включая 25 сфингоидов) |

| | | | | | |
|---|---|---|---|---|---|
| Аминокислоты | + | + | xLLys,<br>xXPmN2,<br>xXSRCetLys<br><br>SRCetLys = S,R-карбокисэтиллизин, PmN2 = диаминопимелиновая кислота | NCCCC[C@H](N)C(=O)O | 42 прототипа остатков |
| Прочие неуглеводные остатки | + | + | xSPyr,<br>xXCho,<br>xXSuc<br><br>Pyr = пировиноградная кислоты, Cho = холин, Suc = янтарная кислота | C[C@](=O)C(=O)O | 70 прототипов остатков (включая 9 нуклеотидов) |
| Нетипичные остатки (агликоны и подстановки свободного текста) | + | +/– | aDGlcp(1-3)Subst // Subst = enoxolone | ***[1*]***O[C@H]1O[C@H](CO)[C@@H](O)[C@H](O)[C@H]1O | представлены в SMILES как изотопно-меченные псевдоатомы |
| Суперклассы остатков | + | +/– | LIP(1-3)xDGro(1-P-2)HEX | ***[1*]***OP(=O)(O)OC[C@H](O)CO***[2*]*** | 28 суперклассов; представлены в SMILES как изотопно-меченные псевдоатомы |
| *Уровень связей* | | | | | |
| Модификации остатков | + | + | Ac(1-2)[Me(1-3)]aDGlcp | CO[C@@H]1[C@@H](OC(C)=O)[C@@H](O)O[C@H](CO)[C@H]1O | алкилирование, ацетилирование и т.д. |

| | | | | | |
|---|---|---|---|---|---|
| Две связи между остатками | + | + | xSPyr(2-4:2-6)aDGalp | C[C@@]1(C(=O)O)OC[C@H]2O[C@H](O)[C@H](O)[C@@H](O)[C@H]2O1 | бифосфаты, O-пируваты, 1,1-связанные ацетали моносахаридов |
| Межостаточные связи C-C и C-N | + | +/– | Me(1C-3)aDGlcp | C[C@@]1(O)[C@@H](O)[C@@H](O)O[C@H](CO)[C@H]1O | C-гликозиды, N-гликаны; C-C-связи с неопределённой позицией связывания не поддерживаются |
| Нестехиометрические связи | + | +/– | -4)[30%Ac(1-3),xXEtN(1-%P-6)]aDGlcp(1- | [*]O[C@H]1O[C@H](COP(=O)(O)OCCN)[C@@H]([*])[C@H](OC(C)=O)[C@H]1O | 30% глюкозы ацетилировано, неизвестная доля глюкозы фосфорилирована; SMILES генерируется для структур со 100% присутствием связей |
| Сложноэфирные и амидные связи | + | + | Ac(1-2)xLLys(1-2)aDGalpN | CC(=O)N[C@@H](CCCCN)C(=O)N[C@H]1[C@@H](O)O[C@H](CO)[C@H](O)[C@@H]1O | |
| *Уровень топологии* | | | | | |
| Олигомерные структуры | + | +/– | bDGlcp(1-2)aDFruf | OC[C@H]1O[C@@H](O[C@@]2(CO)O[C@H](CO)[C@@H](O)[C@@H]2O)[C@H](O)[C@@H](O)[C@@H]1O | |

| | | | | | |
|---|---|---|---|---|---|
| Регулярные полимеры | + | +/– | -9)[Ac(1-5)]aXNeup(2- | ***[*]***C[C@@H](O)[C@@H](O)[C@@H]1O[C@@](O***[*]***)(C(=O)O)C[C@H](O)[C@H]1NC(C)=O | границы повторяющегося звена представлены в SMILES псевдоатомами |
| Нерегулярные полимеры | – | – | | | |
| Циклические полимеры | + | – | CYCLO -4)bDGlcp(1- | | поддержка в отдельном поле CSDB «тип молекулы» |
| Вложенные повторяющиеся звенья | – | – | | | |
| Повторяющиеся фрагменты в олигомерах | – | – | | | |
| Биологические повторяющиеся звенья | + | – | BIOL -4)aLRha(1-3)[Ac(1-2)]aDGlcpN(1- | | поддержка в отдельном поле CSDB «тип молекулы» |
| *Неопределённости в структуре* | | | | | |
| Неизвестные аномерные конфигурации | + | + | ?DGlcp | OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O | |

| | | | | | |
|---|---|---|---|---|---|
| Неизвестные абсолютные конфигурации | + | + | a?Rhap | C[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O<br>C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O)[C@H]1O | для остатков с единственным хиральным атомом этот атом попадает в SMILES с не-определённой конфигураци-ей; для каждого мультихираль-ного остатка генерируется два энантиоме-ра (разные структуры) |
| Неизвестный тип циклиза-ции | + | + | bDGal? | OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@H]1O<br>OC[C@@H](O)[C@@H]1O[C@@H](O)[C@H](O)[C@H]1O | генерируются возможные изомеры (раз-ные структуры) |
| Неизвестное положение свя-зи | + | + | aDGlcp(1-?)aLRhap | C[C@@H]1O[C@@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H](O)[C@H](O)[C@H]1O<br>C[C@@H]1O[C@@H](O)[C@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H](O)[C@H]1O<br>C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O[C@H]2O[C@H](CO)[C@@H](O)[C@H](O)[C@H]2O)[C@H]1O<br>C[C@@H]1O[C@@H](O)[C@H](O)[C@H](O)[C@H]1O[C@H]1O[C@H](CO)[C@@H](O)[C@H](O)[C@H]1O | генерируются все химически разрешённые структуры |

| | | | | | |
|---|---|---|---|---|---|
| Альтернативные фрагменты | + | + | <Ac(1-2)\|Me(1-3)>aDGlcp,<br><br>-3)<<aDGlcp(1-4)\|aDGalp(1-4)>>aLRhap(1- | CC(=O)O[C@H]1[C@@H](O)O[C@H](CO)[C@@H](O)[C@@H]1O<br>CO[C@@H]1[C@@H](O)[C@@H](O)O[C@H](CO)[C@H]1O<br>CO[C@@H]1[C@@H](OC(C)=O)[C@@H](O)O[C@H](CO)[C@H]1O | поддерживается логика «ИЛИ» (OR) и «ЛИБО» (XOR), число альтернатив не ограничено; в SMILES генерируется несколько структур |
| Неизвестный узел присоединения боковой цепи | – | – | | | |
| Известен только мономерный состав | – | – | | | |

[*] Фрагменты кодировки SMILES, обсуждаемые в комментариях, показаны ***курсивом***.

[**] Прототипы остатков – это остатки, не несущие ацетильных групп на атомах азота. При образовании структур каждая из аминогрупп может быть или не быть ацетилированной (напр., прототип GlcN даёт два возможных фрагмента структуры: мономер глюкозамина GlcN и димер N-ацетилглюкозамина Ac(1-2)GlcN).

Так как каждая структура с неопределённостями может приводить к появлению десятков, а иногда и тысяч определённых кодов SMILES и соответствующих им структурных формул, визуализация структур в виде двумерных формул на веб-сайте проекта может занимать длительное время. Для преодоления этого ограничения, присутствующие в базе данных структуры были преобразованы в списки возможных кодов SMILES, а для каждого кода SMILES – в набор графических структурных формул полностью определённых вариантов структуры. Эти изображения (56610 файлов) были кэшированы и используются для ускорения интерфейса пользователя.

### 3.3.4. *Молекулярная геометрия*

Коды SMILES обрабатываются модулем моделирования молекулярной геометрии, что позволяет проводить простые конформационные расчёты в ручном и потоковом режимах для произвольных углеводных структур. Процесс перехода от атомарной связности к атомным координатам показан на Рис. 25, а интерфейс соответствующего веб-инструмента - на Рис. 26.

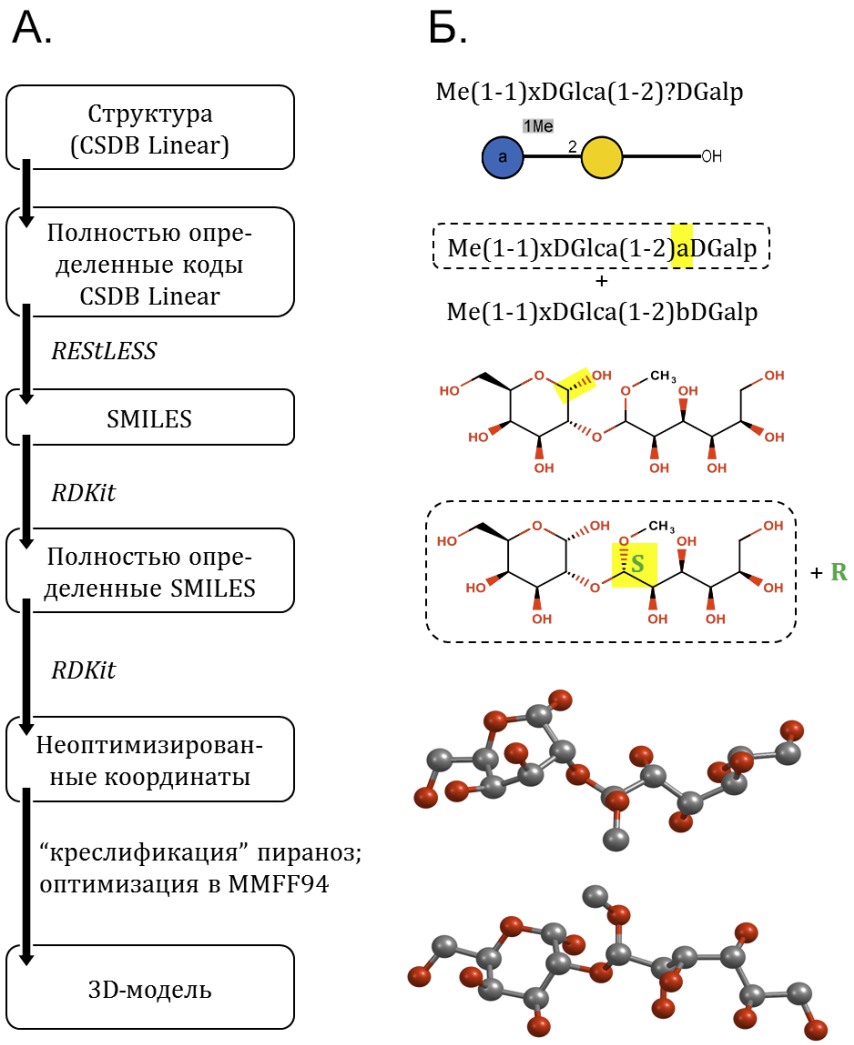



Рис. 25. А. Алгоритм получения пространственной модели углеводов на основе семантического описания. Б. Объекты, используемые на каждом шаге алгоритма на примере модели одной из четырёх возможных структур 1-O-метил-D-глюкозил-2-D-галактопиранозы. Пунктирными рамками выделены объекты, используемые на следующем шаге. Жёлтым обозначен выбор одной из двух стереоконфигураций.

Структуры, записанные в виде SMILES, трансформируются в наборы молекул с полностью определёнными конфигурациями стереоцентров, и для каждой молекулы предсказывается и записывается в формате MOL [124] простран-

ственное расположение атомов. Было обнаружено, что существующие программы для потокового моделирования молекулярной геометрии ошибаются в выборе начальной конформации пираноз в составе более сложных структур, прогнозируя «ванну», твист-форму или инвертированное «кресло». Для устранения

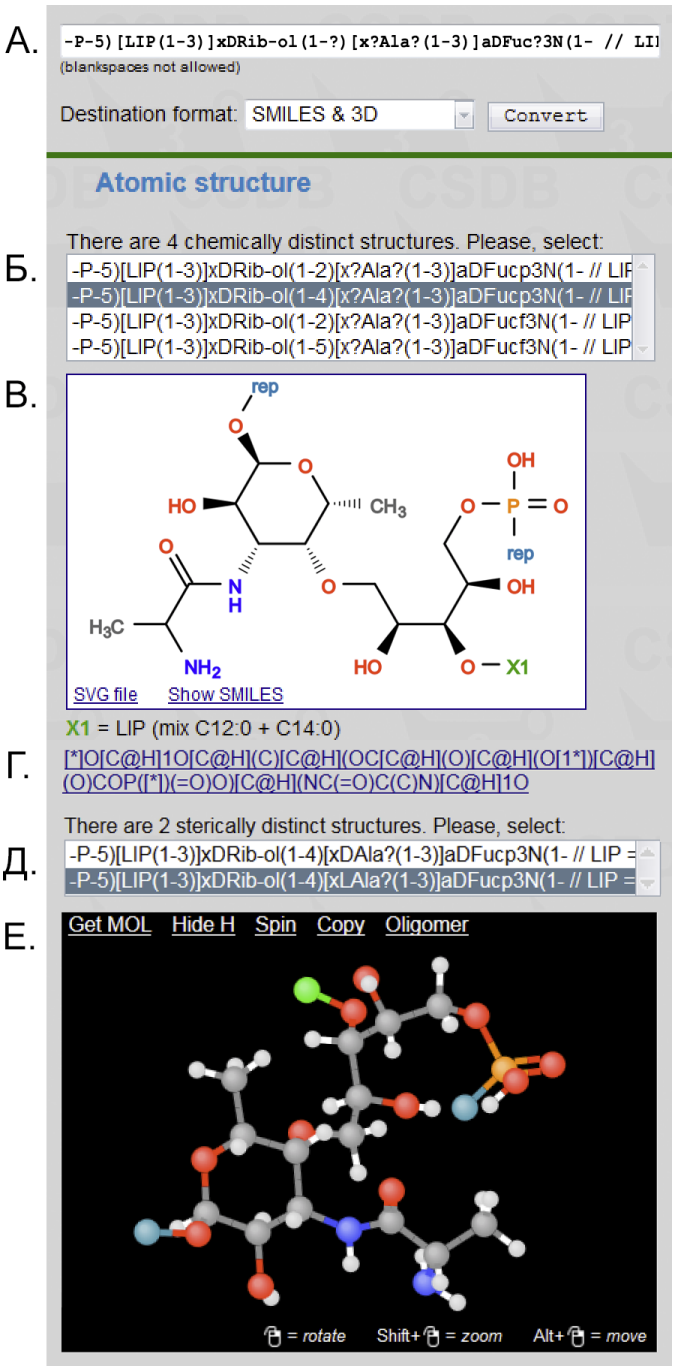



Рис. 26. Интерфейс модуля моделирования геометрии. А. структура, записанная в CSDB Linear. Б. Выбор из химически различных вариантов структуры. В. Сгенерированная формула выбранного варианта. Г. SMILES выбранного варианта. Д. Выбор из возможных стереомеров. Е. Оптимизированная трехмерная атомная модель выбранного стереомера и инструменты работы с ней.

проблем, привносимых неправильной начальной конформацией, 381 моносахарид в пиранозной форме был обсчитан с помощью высокотемпературной молекулярной динамики в силовом поле MM3-2001. Преимущественная конформация пираноз ($^1C_4$ или $^4C_1$) была выбрана на основании анализа количества шагов, в течение которых мономер находился в той или иной конформации в молекулярно-динамической модели.

После внедрения в MOL правильных конформаций мономеров они объединяются друг с другом в полную структуру, которая затем оптимизируется путём релаксации в молекулярно-механическом силовом поле MMFF94 [264]. Этот инструмент предназначен для получения начальных геометрий для последующих ресурсоёмких расчётов молекулярно-механическими или квантово-механическими методами. Более подробно с подготовкой начальной геометрии можно ознакомиться в публикации [146].

На данном этапе начальные конформации гликозидных мостиков выбираются произвольно, и результирующая структура подразумевает конформации, соответствующие первому минимуму, найденному в процессе оптимизации. В настоящее время завершается работа по созданию вспомогательной базы данных конформационных карт подвижных мостиков в ди- и тримерных фрагментах, содержащих один, два или три торсионных угла. После её развёртывания и валидации расчётов на основании экспериментального ЯЭО значения торсионных углов, соответствующие минимумам этих конформационных карт, будут использоваться для моделирования взаиморасположения остатков в каждом фрагменте, который может быть получен из полной структуры. Наличие нескольких минимумов в каждой конформационной карте подразумевает, что параллельно необходимо проводить расчёты на основании многих начальных геометрий.

Для природных структур, уже содержащихся в CSDB, геометрическая модель востребована в дальнейших пользовательских расчётах, однако её получение в рамках поисковых запросов требует значительного времени, особенно при наличии неопределённостей, приводящих к комбинаторному росту количества возможных структур. Для решения проблемы своевременного предоставления информации о пространственных моделях структур, присутствующих в CSDB, их вариации с полностью определёнными конфигурациями (42894 молекулы)

были обсчитаны заранее, помещены в кэш в виде MOL-файлов и предоставлены пользователям. Из-за неопределённостей в ряде структур количество кэшированных молекул приблизительно вдвое превышает количество соединений в CSDB. Для произвольных структур, введённых пользователем, расчёт проводится только первый раз, после чего результаты также помещаются в кэш, из которого извлекаются при последующих запросах. Пределом производительности является расчёт новых структур, содержащих 200-250 неводородных атомов, в течение пользовательской сессии.

### *3.4. Обработка данных и прогнозирование*

На платформе CSDB были разработаны инструменты анализа химической и биологической информации. Эти инструменты позволяют выявлять и обобщать данные, присутствующие в базе неявно.

#### *3.4.1. Моделирование спектров ЯМР*

Для создания инструментов помощи экспертам в интерпретации спектров природных углеводов и для последующей разработки модуля предсказания структуры по спектрам были улучшены имеющиеся и разработаны новые подходы к теоретическому расчёту ЯМР-спектральных параметров углеводов и их производных. Из эмпирических, статистических, квантово-механических, регрессионных и нейросетевых подходов к предсказанию наблюдаемых данных ЯМР были выбраны первые два как имеющие наибольший потенциал для повышения точности расчётов в химии углеводов в сочетании с разумными требованиями к вычислительным ресурсам. Практическое сравнение методов ЯМР-моделирования углеводов опубликовано автором в обзоре [166].

Эмпирическая схема расчёта спектров ЯМР $^{13}C$ углеводов, известная более 25 лет и доведённая до универсального практического использования в рамках кандидатской диссертации автора (2001 г. [116], программный продукт BIOPSEL[a]), была расширена на все классы природных углеводов и родственных соединений, кроме нуклеиновых кислот, дополнена данными по теоретическим эффектам замещения и химическим сдвигам в ди- и тримерных фрагментах, дополнена модулем оценки достоверности моделирования и снабжена веб-интерфейсом, использующим структурные модули CSDB. В современной реализации она представляет собой инкрементную схему, основанную на 9-13 дескрипторах уровня остатков и учитывающую отклонения от аддитивности химических сдвигов, привнесённые стерическим влиянием близкорасположенных заместителей. Для расчётов химических сдвигов используются спектры моно-, ди- и трисахаридов, а также эмпирические эффекты замещения, полученные усреднением наблюдаемых данных при различных комбинациях дескрипторов.

[a] http://toukach.ru/ps.htm

Эти данные хранятся во вспомогательной базе, заполненной на основании анализа литературы, и включают 300 эффектов замещения, спектральные и структурные характеристики 440 мономеров и 3300 димеров и тримеров.

Появление обширной и регулярно пополняемой базы данных CSDB, содержащей более 9200 спектров[a] природных углеводов, открыло возможности для статистического моделирования их спектров. Была разработана модель влияния структуры на химические сдвиги $^{1}$H и $^{13}$C, основанная на идее иерархии сферического окружения атома (HOSE), но учитывающая наличие в сферах HOSE не атомов, как в оригинальном подходе [191], а структурных дескрипторов, характеристичных для углеводов. Эта модель применима к предсказанию любых атомарных параметров углеводов, информация о которых поддаётся формальному описанию и хранению в базах данных. В общем виде разработанная статистическая схема предсказания химического сдвига конкретного атома подразумевает следующие шаги:

1. Выделение из структуры фрагмента, содержащего остаток, включающий предсказываемый атом, и все соседние остатки. Биохимический смысл термина «остаток» (часть структуры, соединяющаяся с другими аналогичными частями в результате реакций с отщеплением воды) в большинстве случаев совпадает с ЯМР-спектроскопическим смыслом (остаток как изолированная протонная спиновая система и связанные с ней атомы углерода).
2. Многошаговое последовательное обобщение структурных характеристик (дескрипторов) фрагмента, начиная с наиболее удалённых от предсказываемого атома, происходящее по мере увеличения изменений в структуре фрагмента и приближения точки их применения к предсказываемому атому. Этот процесс продолжается, пока в базе CSDB не будет найдено статистически значимое количество структур, содержащих обобщённый фрагмент.
3. Усреднение химического сдвига предсказываемого атома в найденных фрагментах с учётом выбросов и оценка достоверности предсказаний на

[a] Данные на 2018-й год.

основании количества и веса проведённых обобщений и дисперсии значений из базы.

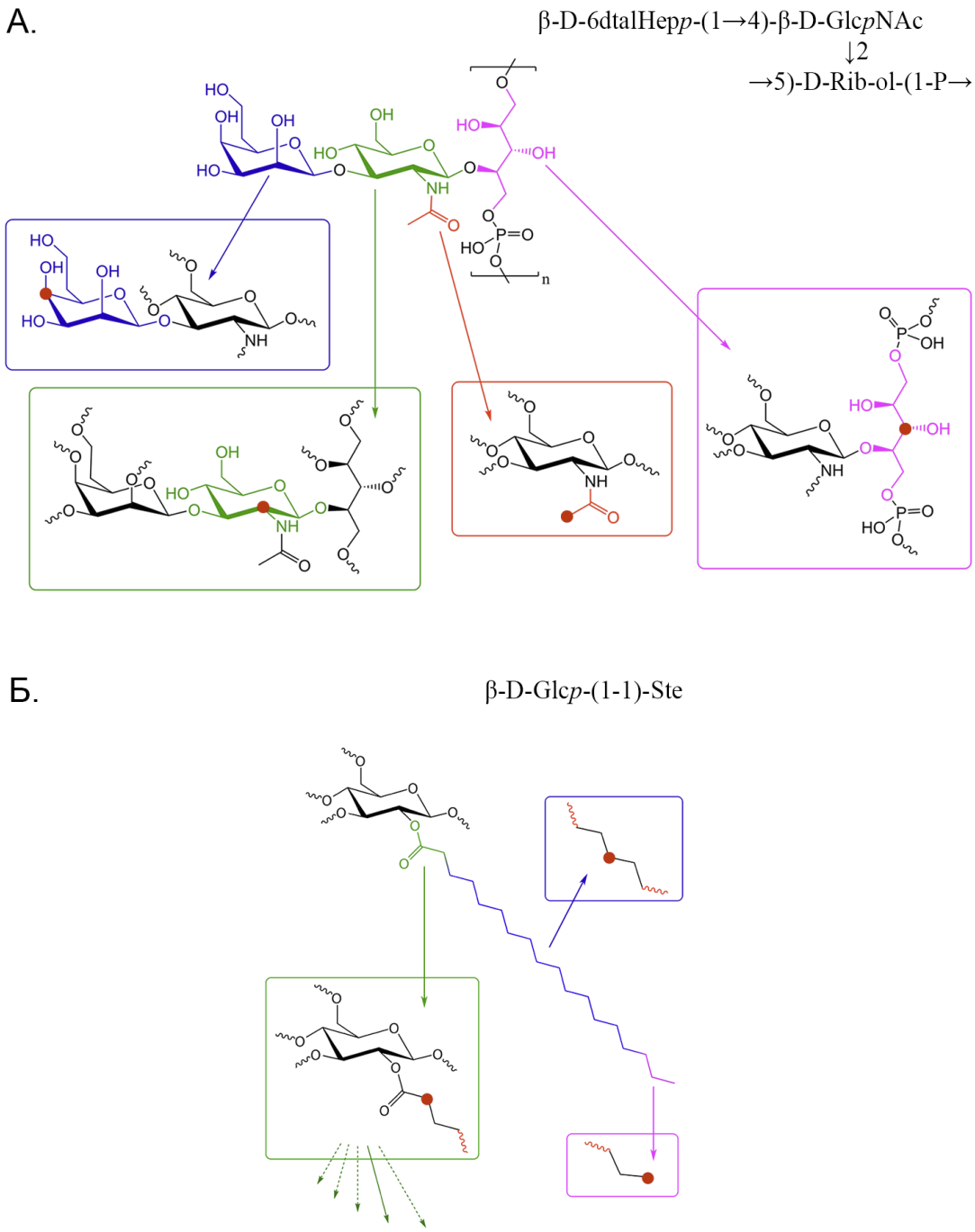



Рис. 27. Фрагментация структуры. А. Разбиение структуры на олигомерные фрагменты. Центральные остатки каждого фрагмента выделены цветом. В каждом фрагменте для примера показан один из предсказываемых атомов (красная точка). Б. Разбиение длинных алифатических цепей на структурно различные области («голова», «середина», «хвост»).

Табл. 14. Классификация углеродных атомов, встречающихся в природных гликанах.

| *Параметр* | *Значение* | *Объяснение* |
|---|---|---|
| Тип | o | >CH-OH (гидрокси) или -$CH(OH)_2$ (дигидрокси) |
| | n | >CH-NH- (амино) |
| | a | -COOH (карбоксил) или -$XO_nOH$, (например, остаток фосфорной или серной кислоты) |
| | A | -CHO (альдегид) или -$C(OH)_2$- (углерод, способный образовывать полуацеталь) |
| | d | -$CH_2$-, >CH- или >C< (дезокси) |
| | D | -CH= или >C= (дезокси $sp^2$- или sp-углерод) |
| | O | -C(OH)= (гидрокси $sp^2$) |
| | N | -C(NHR)= (амино $sp^2$) или -$CONH_2$ (амид) |
| | x | прочее |
| | ? | любой возможный тип |
| | % | любая последовательность возможных типов |
| Стерео-конфигурация | 1 | *l*-конфигурация (по Фишеру) |
| | 2 | *d*-конфигурация (по Фишеру) |
| | 0 | ахиральный углерод |
| | ? | любая конфигурация |
| | % | любые конфигурации нескольких атомов |

Для получения полного спектра структуры все её атомы моделируются независимо. При разбиении структуры на фрагменты образуются подструктуры, содержащие центральный остаток (включающий предсказываемый атом) и соседние остатки: моносахариды, полиолы, аминокислоты, жирные кислоты и т.д. Пример такой фрагментации показан на Рис. 27А. Взаимное влияние фрагментов исключается благодаря тому, что они имеют достаточный размер для изоляции центрального остатка от остатков за пределами фрагмента. В полимерных структурах фрагмент может содержать не только повторяющееся звено, включающее центральный остаток, но и остатки из соседних звеньев, если они связаны с центральным остатком.

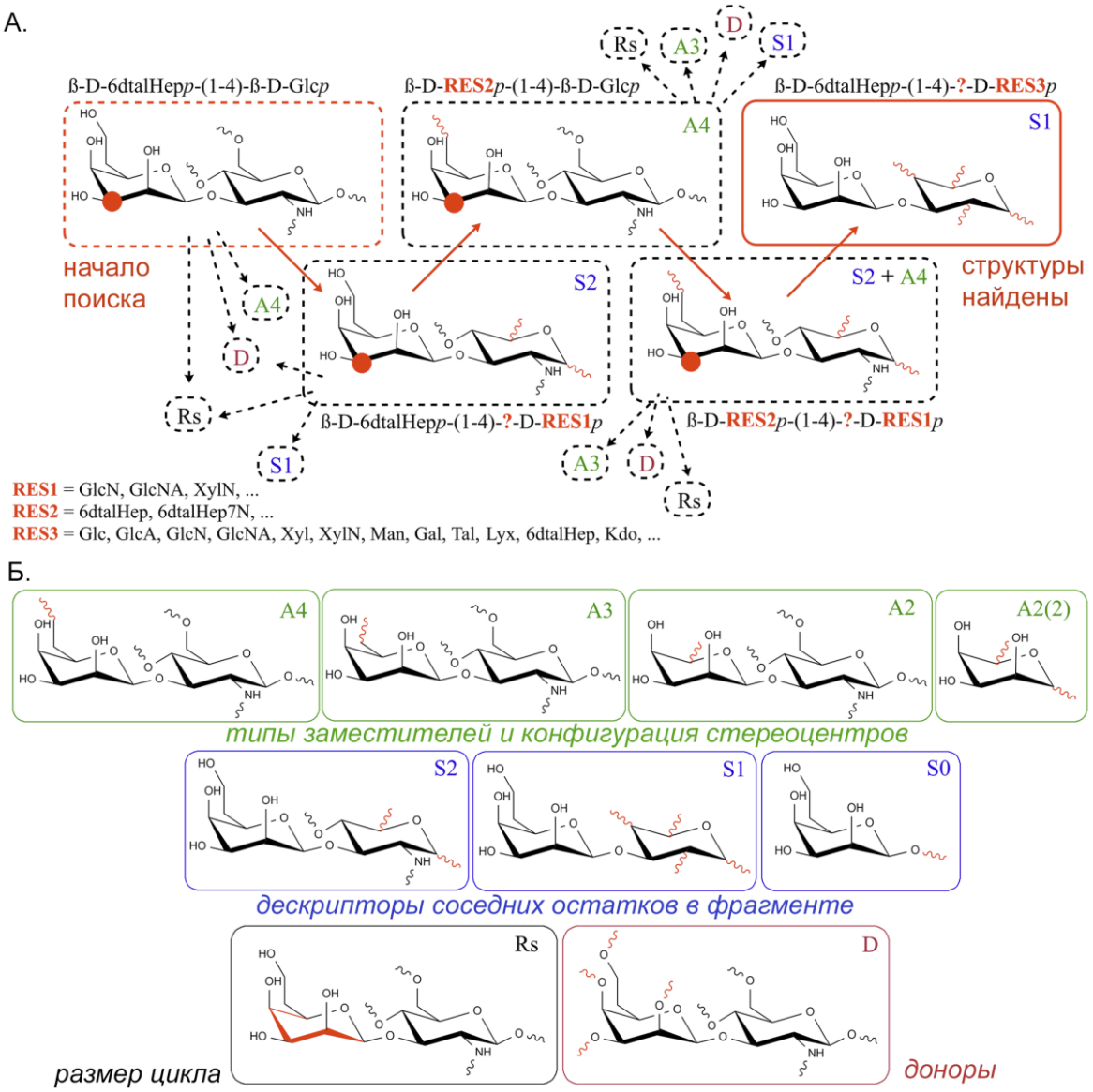



Рис. 28. А. Найденный путь генерализации (красные стрелки) на примере моделирования сигнала С3 6-дезокситалогептозы в структурном фрагменте β-D-6dTalHep*p*-(1→4)-β-D-Glc*p*NAc. Пунктирными стрелками показаны другие возможные пути. Б. Элементарные обобщения. Обобщаемые дескрипторы показаны красным. Волнистая связь подразумевает произвольную стереоконфигурацию и тип заместителя. Генерализации обозначены: A, S, R, D – тип дескриптора (пояснено на рисунке), число после символа – удалённость дескриптора от предсказываемого атома либо от точки образования связи с остатком, содержащим дескриптор.

Полученные фрагменты обобщаются для поиска содержащих их структур. Обобщением называется акт изменения структурных дескрипторов, после которого набору дескрипторов соответствует большее число структур. Например, «превращение» β-D-Fuc*p* в D-Fuc*p* является обобщением дескриптора «аномер-

ная конфигурация», так как первому случаю соответствует одна структура, а второму - две структуры, α-D-Fuc*p* и β-D-Fuc*p*.

Обобщаемые дескрипторы центрального и прилежащих остатков включают:

- пространственные конфигурации и типы атомов углерода (см. Табл. 14);
- размеры циклов (для углеводных остатков – пиранозная, фуранозная или открытая форма);
- абсолютные конфигурации (для оптически активных остатков);
- нахождение центрального остатка на восстанавливающем конце (для корневых остатков – разрешение искать фрагменты, где этот остаток образует исходящую связь);
- нахождение центрального остатка на невосстанавливающем конце (для терминальных остатков - разрешение искать фрагменты, где этот остаток замещён);
- замещение в положениях, не замещённых в исходной структуре (для остатков в цепи - разрешение искать фрагменты, где этот остаток замещён также и в другие положения).

Пример многошагового обобщения фрагмента с использованием разных дескрипторов показан на Рис. 28. Критерием выбора последовательности обобщений является минимизация их суммарного веса (см. ниже).

Тип и стереоконфигурация каждого атома центрального остатка обобщаются одновременно. При этом обобщение атомов, находящихся ближе к рассматриваемому атому, подразумевает обобщение типов и стереоконфигураций атомов, более удалённых от рассматриваемого. Например, если опустить обобщение других дескрипторов остатка, схема структурных обобщений для аномерного атома α-D-Glc*p* C1 будет включать следующие стадии:

1. типы атомов: ooooDo, стереоконфигурации: 221220 (α-D-Glc*p*);
2. типы атомов: ooood%, стереоконфигурации: 22122% (α-D-Glc*p*, α-D-Glc*p*A, L-глицеро-α-D-глюкогептопираноза и др.);
3. типы атомов: ooo?d%, стереоконфигурации: 221?2% (все остатки из п.2, α-D-Glc*p*4N, α-D-Gal*p*, α-D-Gal*p*4N и др.);

4. типы атомов: ooo%, стереоконфигурации: 221% (все остатки из п.3, α-D-Xyl*p* и др.);
5. типы атомов: oo%, стереоконфигурации: 22% (все остатки из п.4, α-D-Fuc*p*3N и др.);
6. типы атомов: o%, стереоконфигурации: 2% (все остатки из п.5, α-D-Man*p*, α-D-Glc*p*N и др.);
7. любой остаток.

Когда центральный остаток образует связь с другими остатками, их дескрипторы также обобщаются. Для ускорения предсказаний обобщение параметров этих остатков происходит ступенчато (Рис. 29). На каждом шаге обобщается определённая часть дескрипторов остатка, оказывающая наименьшее влияние на значение химического сдвига (дескрипторы имеют наименьший вес, см. ниже). Чем более удалён дескриптор от связи между остатками, тем меньше ступеней применяется.

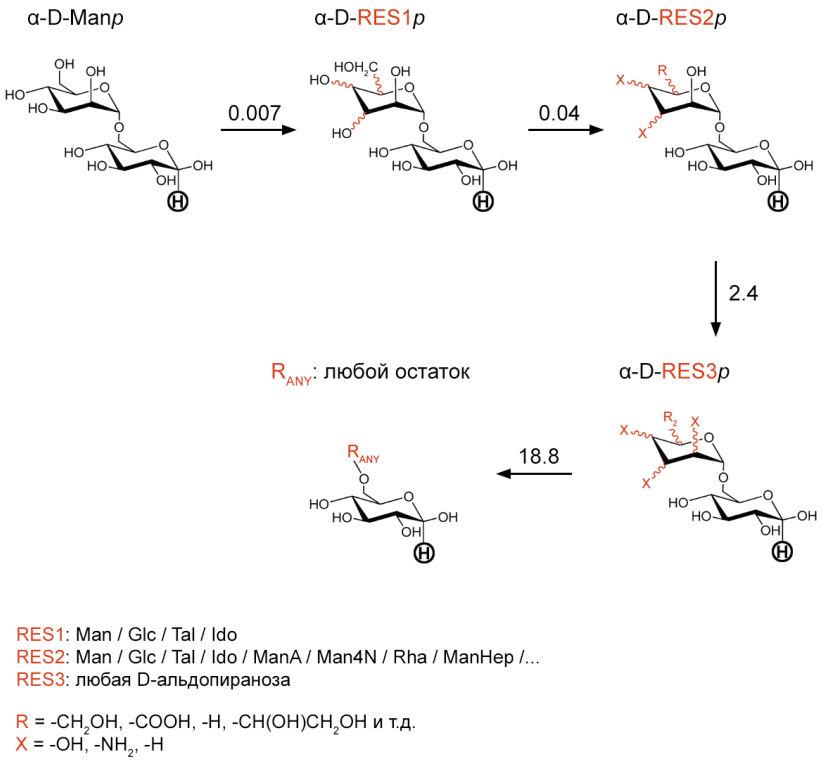



Рис. 29. Пример обобщения остатка-донора в фрагменте α-D-Man*p*-(1→6)-β-D-Glc*p*. Обобщаемые дескрипторы показаны красным. Веса приведены над стрелками для случая предсказания протона Glc H1, показанного в кружке. На первом шаге обобщаются стереоконфигурации удалённых атомов донора, на втором – типы заместителей при этих атомах, на третьем – конфигурация и тип ближайшего к связи атома, на четвёртом – весь остаток-донор.

Каждому структурному параметру фрагмента, который может быть подвергнут обобщению, ставится в соответствие эмпирический весовой фактор («вес»), отображающий влияние этого параметра на химический сдвиг предсказываемого атома. Вес обобщения зависит от числа связей между положением дескриптора и предсказываемым атомом (удалённости параметра), природы параметра и природы центрального остатка. При выработке пути обобщений разработанный алгоритм начинает с обобщений дескрипторов с минимальным весом, которые относятся к наиболее удалённым группам атомов. Целью является нахождение обобщения, вносящего наименьшее искажение в величину предсказываемого химического сдвига. Для иллюстрации этого критерия примем, что в базе данных содержатся отнесённые спектры ЯМР дисахаридов α-D-Man*p*-(1→4)-α-D-Glc*p* и α-D-Tal*p*-(1→4)-β-D-Glc*p* и требуется предсказать химический сдвиг атома C1 глюкозы в дисахариде α-D-Tal*p*-(1→4)-α-D-Glc*p*. Использование для предсказания данных второго дисахарида приведёт к заметной ошибке из-за несовпадения аномерных конфигураций глюкозы, в то время как использование маннозосодержащего дисахарида даст хороший результат, т.к. его единственным отличием от целевой структуры является стереоконфигурация удалённого от Glc C1 четвёртого атома остатка-заместителя. Веса обобщений позволяют формализовать подобные рассуждения о влиянии типа и положения дескрипторов на химические сдвиги в общем виде. Для итеративного нахождения оптимальных значений весов всех возможных комбинаций дескрипторов и предсказываемых атомов в олигомерных фрагментах был использован «генетический» алгоритм ABC [265] («искусственная пчелиная колония»), широко применяющийся для задач оптимизации. Подробности алгоритма оптимизации и результирующие значения весов опубликованы в статьях автора [117, 118].

При обобщении фрагментов требуется соблюдать баланс между максимизацией числа подходящих структурных фрагментов в базе данных и минимизацией влияния обобщений на результирующий химический сдвиг. Так как спектры ЯМР энантиомеров в ахиральном окружении совпадают, инверсия абсолютных конфигураций каждого остатка в фрагменте может увеличить число подходящих структур без уменьшения точности предсказания. Один из двух альтернативных наборов абсолютных конфигураций фрагмента (исходный и инвертиро-

ванный) выбирается на основании данных о встречаемости составляющих его остатков (Рис. 30А).

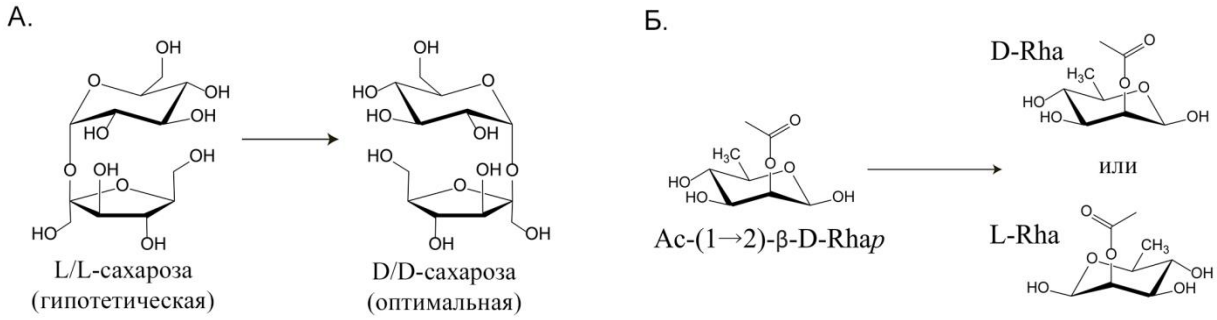



Рис. 30. Примеры предварительных обобщений с нулевым весом: А. Инверсия абсолютных конфигураций всех остатков в фрагменте. Б. Обобщение абсолютных конфигураций остатков, не имеющих оптически активных остатков-заместителей.

Если все соседние остатки оптически неактивны (напр., остаток уксусной кислоты в фрагменте на Рис. 30Б), предварительное обобщение абсолютной конфигурации центрального остатка не повлияет на спектры фрагмента, и следовательно, будет иметь нулевой вес, повышая при этом вероятность нахождения подходящих структур в базе.

Алгоритм обобщения имеет параметр, регулирующий выбор в пользу точности либо скорости моделирования. Он лимитирует число шагов и максимальный вес обобщений в соответствии со значениями, приведёнными в Табл. 15. При достижении критических значений химический сдвиг обозначается, как непредсказанный.

Табл. 15. Режимы работы в контексте приоритета точности или скорости.

| *режим* | *fast (быстрый)* | *accurate (оптимум)* | *extreme (точный)* |
|---|---|---|---|
| максимальное число обобщений на один атом | 5 | не ограничено | не ограничено |
| максимальный вес обобщений | 100 | 100 | не ограничен |
| число шагов обобщения остатков, соседних с содержащим предсказываемый атом | 1 | 1-2 * | 4 |

* в зависимости от удалённости предсказываемого атома от точки присоединения соседнего остатка (по приведённому итеративно-оптимизированному числу связей).

Режим *accurate* рассматривается как оптимальный по соотношению точности и скорости предсказаний. Режим *fast* может использоваться для более грубой и быстрой оценки химических сдвигов. В некоторых случаях использование режима *extreme* может приводить к увеличению точности, однако для экзотических остатков или остатков, образующих связь с несколькими заместителями, предсказание может отнимать более десяти минут. Более подробно особенности указанных режимов описаны в публикации [117].

Природные гликоконъюгаты часто содержат остатки с большим числом атомов углерода, что приводит к лавинообразному нарастанию числа возможных комбинаций обобщений. В этом случае скорость моделирования становится критичной для обработки множества структур, например, при автоматизированном установлении структуры по спектру. В то же время остатки, состоящие из одинаковых многократно повторяющихся групп атомов, не требуют тонко настроенной для углеводов схемы обобщений. Для наиболее распространённого из таких случаев – длинных алифатических цепей (жирные кислоты, сфинголипиды, алифатические спирты и т.д.) – был разработан специальный алгоритм обобщений. Такие остатки разбиваются на три сектора (Рис. 27Б): «голову» (ближайшие два углеродных атома), «хвост» (концевые группы из трёх углеродных атомов) и «середину» (оставшиеся атомы). Для атомов «головы» первое обобщение затрагивает алифатические атомы, удалённые от предсказываемого атома более чем на две связи, и далее обобщения происходят по углеводной схеме, включая соседние остатки (глюкоза на Рис. 27Б). Обобщение остатка для атомов из «середины» и «хвоста» не оставляет информации об остатке, связанном с «головой», и алгоритм ищет в базе фрагменты, включающие атомы в пределах двух связей от предсказываемого.

После обобщения структурного окружения и поиска в базе структур, содержащих обобщённые фрагменты, экспериментальные химические сдвиги предсказываемого атома усредняются. Во избежание потери точности из-за ошибочных данных или данных, полученных в нестандартных экспериментальных условиях, перед усреднением выборка проверяется на наличие статистических отклонений («выбросов») с помощью критерия Шовене [266]. Для удаления

скрытых выбросов (которые появляются только после удаления явных) проверка повторяется, пока в выборке не останется выбросов.

Растворитель, используемый для приготовления образца для ЯМР-анализа, может влиять на химические сдвиги сигналов вплоть до нескольких м.д. в случае ЯМР $^{13}$C и до 0.5 м.д. в случае ЯМР $^{1}$H [267]. В случае природных углеводов большинство исследований методом ЯМР проводится в водных растворах, поэтому по умолчанию моделирование предполагает воду в качестве растворителя. Однако для гликоконъюгатов растительного происхождения спектры ЯМР часто снимают в пиридине-*d5* и других растворителях. Для моделирования спектров ЯМР таких структур введён параметр, указывающий, данные для какого растворителя разрешено использовать при поиске и выборке данных, что делает разработанный метод единственным неквантово-механическим способом предсказания спектров углеводов с учётом растворителя. Точность статистического моделирования зависит от заполненности базы данных спектрами, снятыми в конкретном растворителе. Наиболее распространёнными растворителями в CSDB являются вода и пиридин; более подробную статистику можно посмотреть по ссылке «Coverage» на сайте GODDESS[a].

Кислотность среды влияет на химические сдвиги атомов, близких к заряженным группам и к протонам, склонным к химическому обмену [268]. Это влияние учитывается путём задания допустимого интервала pH. Публикуемые спектры ЯМР углеводов чаще всего записаны без дополнительного подкисления или подщелачивания, поэтому, учитывая, что свободные аминогруппы в этих соединениях встречаются редко, а высокая точность предсказания сигналов карбоксильных групп не востребована, отсутствие ограничений на pH даёт хорошую точность предсказаний.

Влияние температуры [269] учитывается аналогичным образом. Концентрация растворённого вещества в типичном для природных соединений диапазоне, как правило, слабо влияет на химические сдвиги сигналов. Химические сдвиги в CSDB уже приведены к единому стандарту, поэтому проблемы калибровки шкалы при моделировании не возникает. Полученные химические сдвиги

[a] http://csdb.glycoscience.ru/database/core/nmrsim.html

нормированы относительно тетраметилсилана ($^{1}$H и $^{13}$C - 0 м.д.), моделируемого исходя из положения сигнала добавки ацетона в водный раствор ($^{1}$H 2.25 м.д., $^{13}$C 31.45 м.д.). Для находящихся в CSDB спектров ЯМР, стандартом в которых выступает DSS, химические сдвиги $^{1}$H и $^{13}$C откорректированы сдвигом в сильное поле на 0.017 м.д. [270] и 2.67 м.д.[271], соответственно. Рабочая частота спектрометра и его настройка не влияет на химические сдвиги, выраженные в м.д.

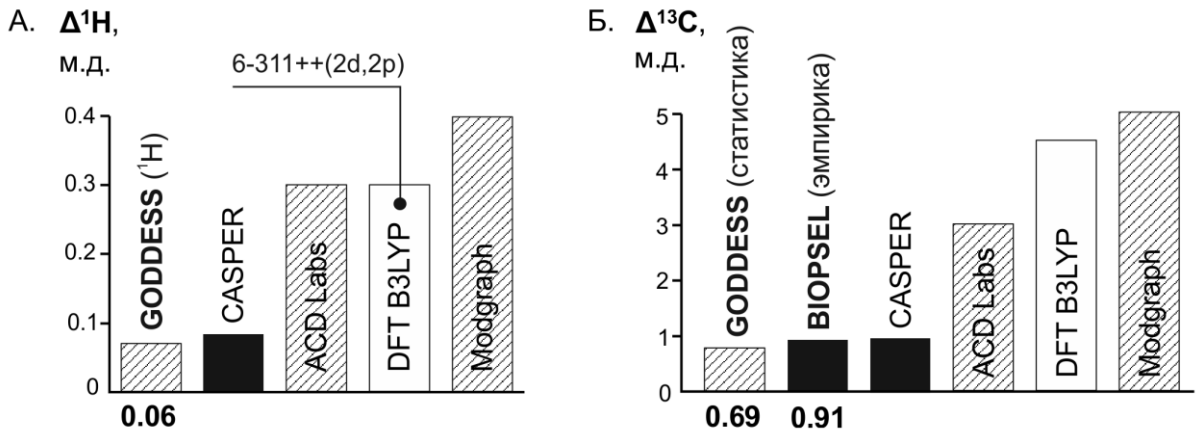



Рис. 31. Сравнение предсказательной силы алгоритмов GODDESS (статистика) и BIOPSEL (эмпирика) с другими алгоритмами моделирования химических сдвигов. Эмпирические подходы показаны черным, статистические – штриховкой. Средняя точность разработанных алгоритмов на природных углеводах (м.д.) показана в нижней строке. А. Моделирование протонных спектров; Б. Моделирование углеродных спектров.

Разработанные подходы к моделированию спектров ЯМР углеводов были реализованы в инструменте GODDESS (Glycan-Optimized Database-Driven Empirical Spectrum Simulation; «моделирование спектров углеводов, основанное на базе данных»). Получаемые модели были валидированы двумя способами: на десяти отобранных тестовых структурах, отражающих разнообразие природных углеводов, и на статистической выборке природных структур, для которых опубликованы спектры ЯМР. Первый способ использовался, в основном, для измерения производительности и сравнения списка поддерживаемых структурных особенностей с возможностями других существующих методов. По результатам этого сравнения можно сделать вывод о существенном преимуществе методов, оптимизированных для углеводов (GODDESS [117, 118], BIOPSEL [116],

CASPER [203]), перед методами, основанными на использовании «общеорганического» HOSE, нейронных сетей и квантово-механических расчётов на высоких уровнях теории в больших базисных наборах, которые считаются достаточными для моделирования спектров ЯМР органических соединений (Рис. 31). Из оптимизированных методов ближайший конкурент, CASPER, демонстрирует сравнимую точность, но поддерживает лишь ограниченный набор структурных особенностей, не покрывающий разнообразия углеводов прокариот.

Поддержка большинства структурных особенностей природных биогликанов является отличительной характеристикой GODDESS. Так, с помощью инструмента GlyNest [58]удалось предсказать спектры ЯМР только для глюкозы и модельных структур, представленных на сайте GLYCOSCIENCES.de: предсказание спектров полимеров не поддерживалось, а спектры олигомеров не содержали сигналов остатков, находящихся на восстанавливающем конце. CASPER имеет ограниченный набор поддерживаемых компонентов структуры: например, с помощью этого метода невозможно предсказать спектры ЯМР простого галактана, содержащего фуранозный остаток. Инкрементный подход, основанный на BIOPSEL, поддерживает большее число структурных особенностей, однако с его помощью возможно предсказывать только спектры ЯМР $^{13}$C. Как эмпирические методы общего назначения, так и методы *ab initio* не поддерживают предсказания спектров ЯМР полимерных молекул; кроме того, последние работают приблизительно на пять порядков медленнее. В большинстве случаев предсказание спектра ЯМР типичной природной структуры программой GODDESS занимает около одной минуты.

Характерные примеры суперпозиции предсказанных и экспериментальных сигналов приведены на Рис. 32. В качестве модельных структур здесь выбраны бактериальные гликополимеры, структура которых была установлена автором ранее [272, 273]. Они использовались для валидации как положений, так и предсказанной ширины сигналов (см. ниже). В случае химических сдвигов $^{1}$H совпадение сигналов находится в пределах экспериментальной погрешности. По углеродной оси в $^{1}$H,$^{13}$C HSQC визуально заметные несовпадения наблюдаются для атомов, положение сигналов которых зависит от pH раствора (C2/H2 лизина, C5/H5 галактуроновой кислоты) и конформационно подвижного C6/H6 4,6-

дизамещенной глюкопиранозы, которая может быть подвержена необычным стерическим эффектам.

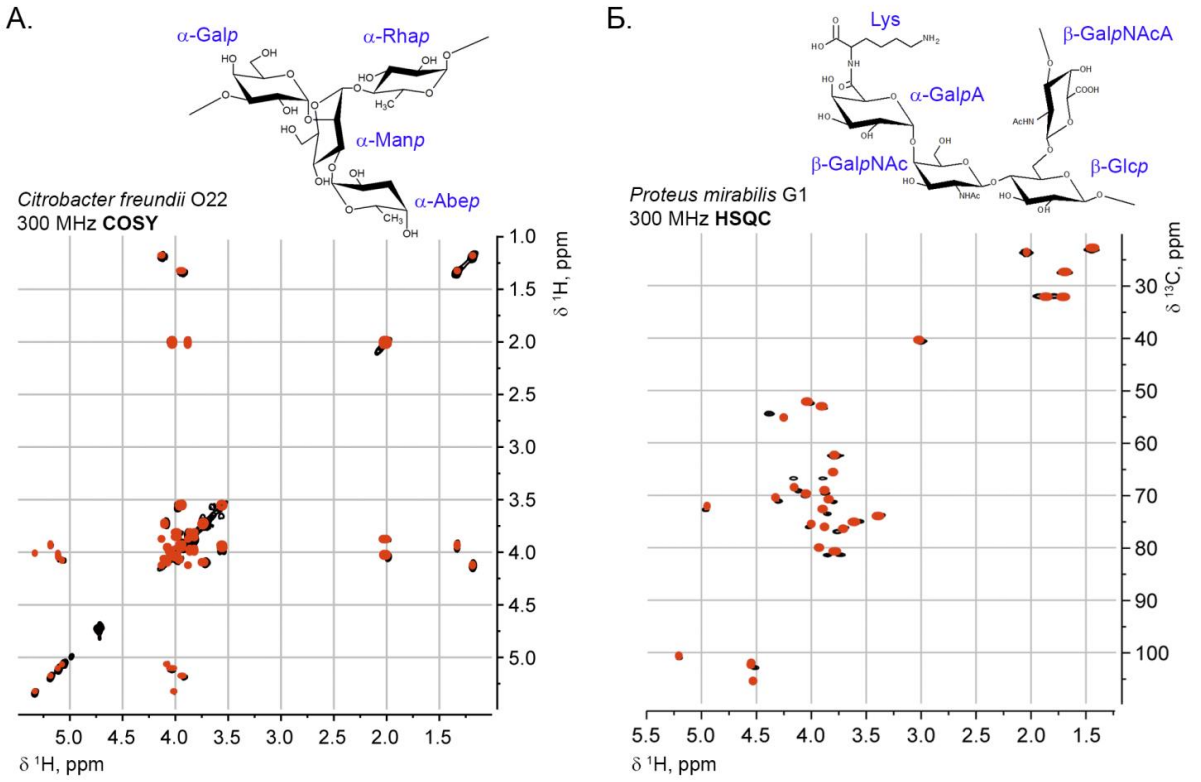



Рис. 32. Суперпозиция предсказанного(красный) и экспериментального (чёрный) спектров на примерах А. COSY полисахарида *Citrobacter frendii* O22 ($D_2O$, 30°C, 600 МГц); Б. $^{1}H,^{13}C$ HSQC полисахарида *Proteus mirabilis* G1 ($D_2O$, 45 °C, 500 МГц). Предсказанные спектры получены для условий: $D_2O$, 25 °C, 300 МГц.

Для статистического анализа точности моделирования из базы CSDB было отобрано 36385 химических сдвигов $^{13}C$ ЯМР, симулирование которых поддерживается обоими методами, и 40441 химических сдвигов $^{1}H$ ЯМР. Критерием отбора было отсутствие неопределённостей в опубликованной углеводной структуре и доступность экспериментальных одномерных спектров ЯМР в водных растворах. Выборки характеризовали разнообразие природных углеводов, а именно содержали остатки пираноз, фураноз, высших сахаров, полиолов, аминокислот, жирных кислот, другие остатки неуглеводной природы. Структуры и экспериментальные спектры ЯМР, с которыми сравнивалась полученные модели, присутствовали в CSDB в явном виде, создавая заведомо идеальные условия для моделирования. Поэтому для предотвращения искажения результатов в

лучшую сторону соответствующая запись (структура и спектр) виртуально удалялась из базы данных на время моделирования.

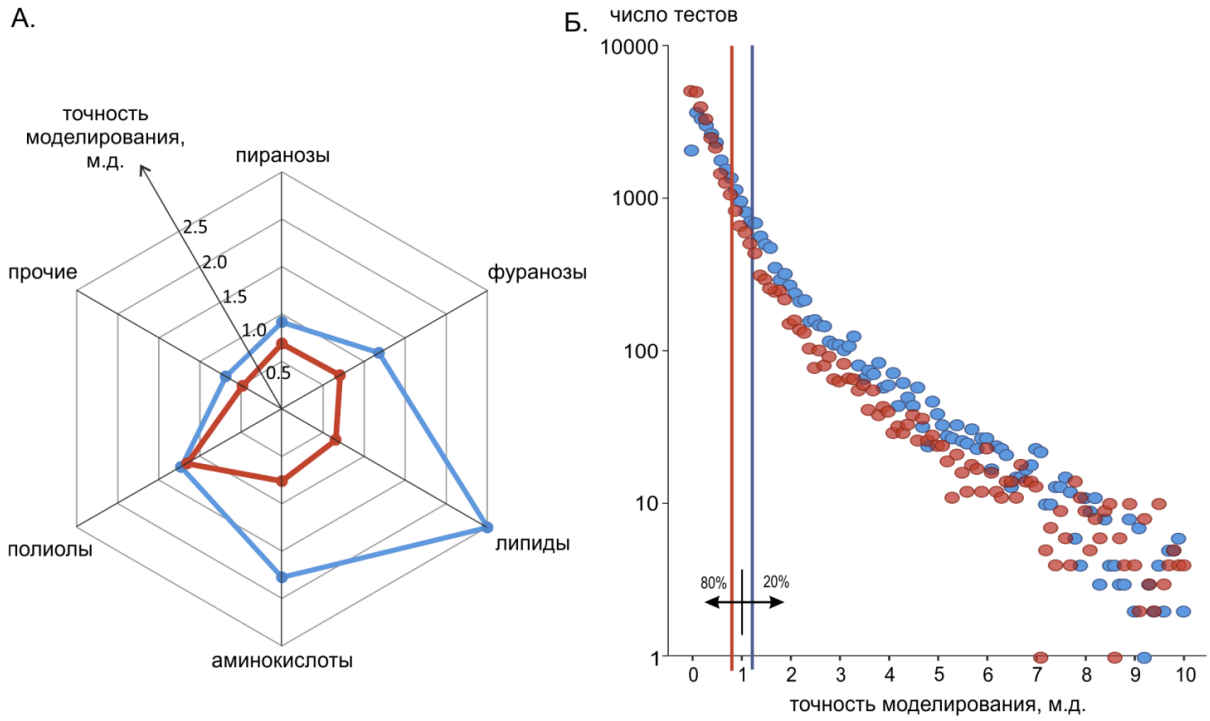



Рис. 33. А. Точность моделирования углеродных химических сдвигов для остатков разных классов. Синие точки –эмпирическое моделирование, красные – статистическое. Б. Распределение точности моделирования ~36000 атомов в природных структурах. Вертикальные линии показывают уровень точности, разделяющий 80% лучших и 20% худших предсказаний.

На Рис. 33 приведены распределения абсолютных отклонений предсказанных значений от экспериментальных, а также средняя точность моделей для разных типов остатков. При использовании статистического подхода средняя точность составила 0.69 м.д. и 0.06 м.д. для $^{13}$C и $^{1}$H спектров ЯМР, соответственно; для 95% предсказаний ошибка лежала в пределах 2.6 м.д. и 0.23 м.д., соответственно. При использовании эмпирического подхода средняя точность предсказания $^{13}$C спектров ЯМР составила 0.91 м.д., для 95% предсказаний ошибка лежала в пределах 3.0 м.д. 80% химических сдвигов $^{13}$C было предсказано с точностью, превышающей 0.8 м.д. и 1.2 м.д. для статистического и эмпирического методов, соответственно. Преимущество статистического подхода особенно заметно проявилось для остатков, которые в силу конформационной лабильности или нехватки данных по эффектам замещения плохо поддаются тео-

ретическому моделированию: фураноз, аминокислот и жирных кислот (Рис. 33А). Для полиолов разница в точности оказалась минимальной. Это связано с проблемой неоднозначности выбора абсолютной конфигурации остатков симметричных полиолов в сочетании с позициями замещения (напр., -4)D-Rib-ol(1- или -2)L-Rib-ol(5- ), пока не решённой в GODDESS в силу ограничений нотации CSDB Linear.

Достоверность предсказания оценивается для каждого атома и нормируется с получением числа от 0 до 100. При использовании эмпирического расчёта она зависит от того, использовались ли значения химических сдвигов из базы данных и применялись ли сохранённые для данного структурного окружения теоретические эффекты замещения, а также от числа пермутаций структурного фрагмента, которое потребовалось для нахождения в базе химического сдвига или эффекта. При использовании статистического метода достоверность зависит от суммарного веса применённых обобщений $W$ (чем больше вес, тем меньше достоверность), размера ($N$) и стандартного отклонения ($\sigma$) выборки химических сдвигов (чем больше выборка и чем меньше отклонение, тем больше достоверность). Эта зависимость формализована в виде суммы полиномов второго порядка с коэффициентами, итеративно подобранными по критерию максимизации коэффициента линейной корреляции между реально наблюдаемыми отклонениями экспериментальных значений от предсказанных и величинами достоверности для выборки структур:

$$T\ (\text{достоверность}) = 100 - P_W(W) - P_N\left(\frac{1}{N}\right) - P_\sigma(\sigma)$$

где $P_X(X) = x_1X + x_2X^2$, где коэффициенты $x_1$ и $x_2$ зависят от природы параметра $X$ ($W$, $N$ или $\sigma$) и типа остатка, которому принадлежит рассматриваемый атом.

Для перевода значения достоверности в ожидаемую ошибку моделирования с помощью регрессии получены линейные зависимости (Рис. 34). Положение кругов на графике отражают результаты сопоставления достоверностей предсказания химических сдвигов с разницей между моделью и экспериментальными данными. Более подробно с анализом достоверности и точности созданных моделей можно ознакомиться в публикации [118].

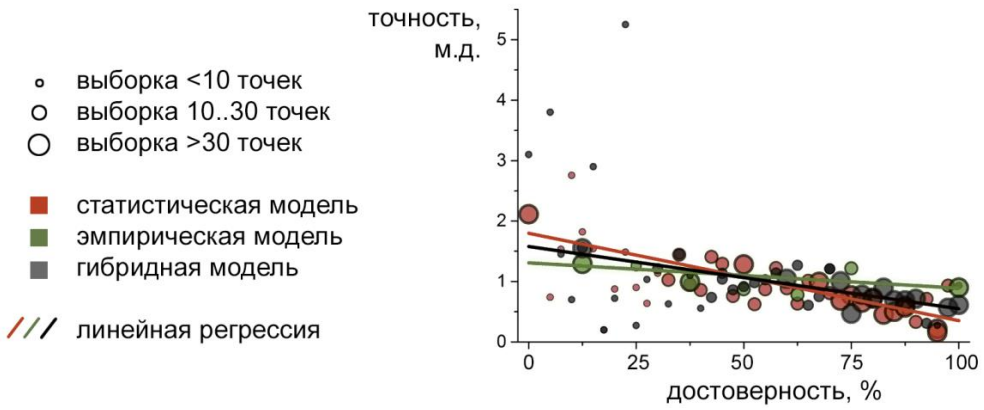



Рис. 34. Регрессионный анализ корреляции между достоверностью и точностью предсказания.

По причине неравномерной полноты баз данных атомы в одной и той же структуре предсказываются каждым из подходов с разной достоверностью в зависимости от химического окружения. Это может приводить к значительным ошибкам, когда в рамках используемого подхода для моделирования одного конкретного атома или группы атомов не хватает данных. В то же время, данные для этой группы, используемые другим подходом, могут обеспечивать более высокую достоверность. Для избежания ошибок такого рода была разработана гибридная модель углеродных химических сдвигов. Она подразумевает смешивание эмпирических и статистических предсказаний на основании значений химических сдвигов и их достоверностей, сгенерированных каждым из методов. Для получения гибридного химического сдвига рассчитывается линейная комбинация эмпирического и статистического значений, коэффициенты в которой зависят от полученных достоверностей каждого из подходов и различия между моделями. Гибридная достоверность учитывает, в какой степени одна модель подтверждает или опровергает другую. Формулы для получения гибридных данных приведены и обоснованы в публикации [117]. «Гибридизация» проводится независимо для каждого атома в моделируемой структуре и в большинстве случаев дает более точные предсказания, чем эмпирический и статистический подходы по отдельности, однако в случае аномально большого различия в точности ЯМР-моделирования молекулы в целом точность гибридной модели может быть меньше, чем точность одной из её составляющих. Во избежание потери точности пользователю

предоставляются все три варианта углеродных химических сдвигов и суммарные достоверности моделирования структуры (см., напр., Рис. 38Б).

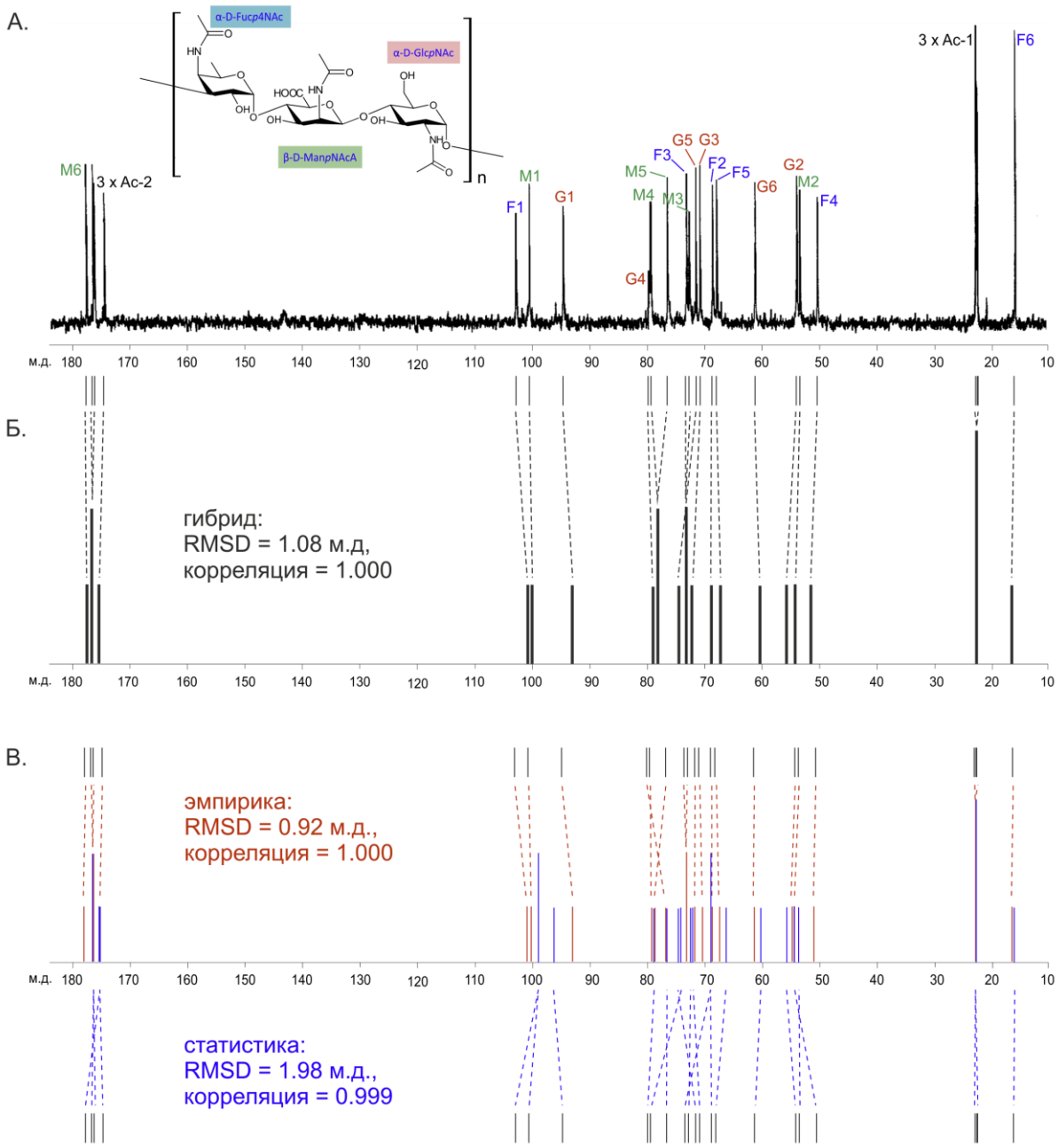



Рис. 35. Моделирование спектра ЯМР $^{13}C$ общего антигена энтеробактерий. Перед получением моделей записи CSDB, содержащие спектры общего антигена, были временно удалены из базы во избежание завышения точности из-за «правильного» подбора примера. А. Структура, экспериментальный спектр ЯМР $^{13}C$ в $D_2O$ (получен автором для полисахарида *Proteus penneri* 17) и отнесение сигналов. Б. Модель, полученная гибридным методом. Соответствие сигналов экспериментальным данным показано пунктирными линиями. В. Эмпирическая (красный спектр) и статистическая (синий спектр) модели. Короткими чёрными линиями продублированы экспериментальные значения.

В качестве характерного примера использования всех трех методов можно привести моделирование углеродного спектра общего углеводного антигена энтеробактерий [274]. Несмотря на распространённость этого трисахаридного повторяющегося звена в гликоме прокариот, инфицирующих млекопитающих, в других структурах такие комбинации остатков встречаются редко, поэтому это соединение является сложным случаем для статистического предсказания. Гибридное моделирование позволило получить отличную корреляцию с экспериментальным спектром ЯМР $^{13}C$ (Рис. 35Б). Кроме того, этот сахарид невозможно промоделировать ни одним другим из существующих специализированных методов из-за нестандартных модификаций остатков, квантово-механические методы неприменимы из-за полимерной структуры, а общехимические методы показывают неудовлетворительную точность из-за несовершенства универсальной стереохимической модели.

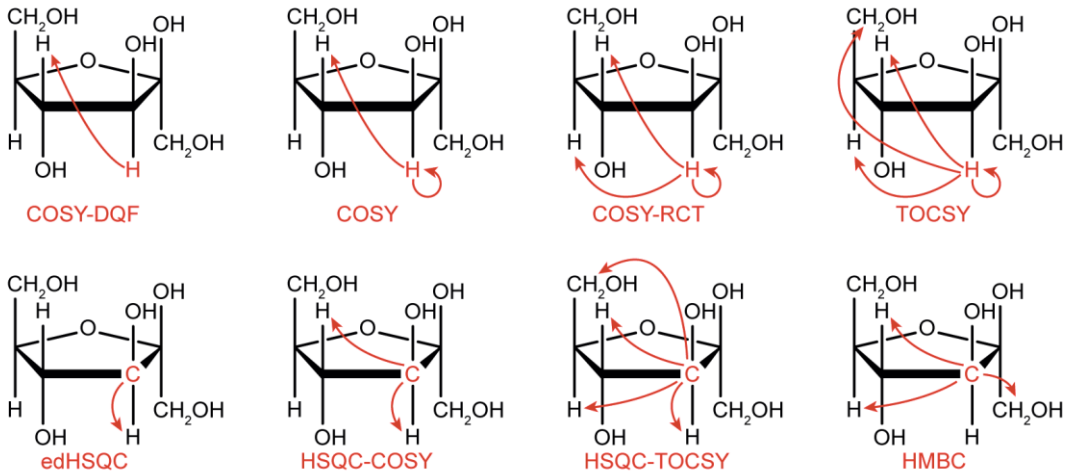



**Рис. 36.** Поддерживаемые модулем визуализации эксперименты ЯМР (красные подписи) и спиновые корреляции (красные стрелки), выявляемые этими экспериментами на примере 3-го положения β-D-фруктофуранозы.

Предсказанные химические сдвиги визуализируются в виде таблиц отнесения, одномерных углеродных спектров и двумерных спектров. До настоящего времени симуляторы двумерных спектров углеводов позволяли получить только спектры HSQC (в проекте CASPER). Разработанный модуль визуализации поддерживает все основные гомо- и гетероядерные спиновые корреляции, востребованные в структурных исследованиях углеводов (Рис. 36),

кроме корреляций, основанных на ядерном эффекте Оверхаузера:

- COSY (корреляция протонов с геминальными и вицинальными соседями-протонами);
- COSY-DQF (COSY без диагональной линии);
- COSY-RCT (COSY с одноступенчатым переносом когерентности по цепочке взаимодействующих вицинальных протонов);
- TOCSY (тотальная корреляция всех протонов в пределах спиновой системы);
- edHSQC (прямая гетероядерная корреляция). Атомам углерода, образующим связь с нечётным числом атомов водорода, ставится в соответствие положительный кросс-пик, а атомам углерода $CH_2$-групп – отрицательный;
- HSQC−COSY (корреляция атомов углерода с собственными протонами, а также с их вицинальными соседями, напр., C1−H2 в глюкозе);
- HSQC−TOCSY (корреляция атомов углерода со всеми протонами спиновой системы собственного протона);
- HMBC (гетероядерная корреляция через две или три связи, включая связи через гетероатомы). Включает трансгликозидные корреляции.

Этот набор спектров доступен для работы в браузере, включая экспорт и наложение предсказанных и экспериментальных спектров для визуального сравнения. Он позволяет химикам проверять структурные гипотезы и делать отнесение сигналов в спектрах сложных объектов с минимальными усилиями.

Отнесение сигналов, представленное в виде таблицы, может быть экспортировано в формате TSV, а сами спектры – в формате Jcamp-DX [275]. Оба формата импортируются распространёнными программами обработки данных ЯМР, напр., сравнение моделей α- и β-лактозы с экспериментальным спектром смеси аномеров (Рис. 37) получено в программе MestreNova 9 [276][a] в автоматическом режиме на основании сгенерированных данных.

[a] http://mestrelab.com/software/mnova/

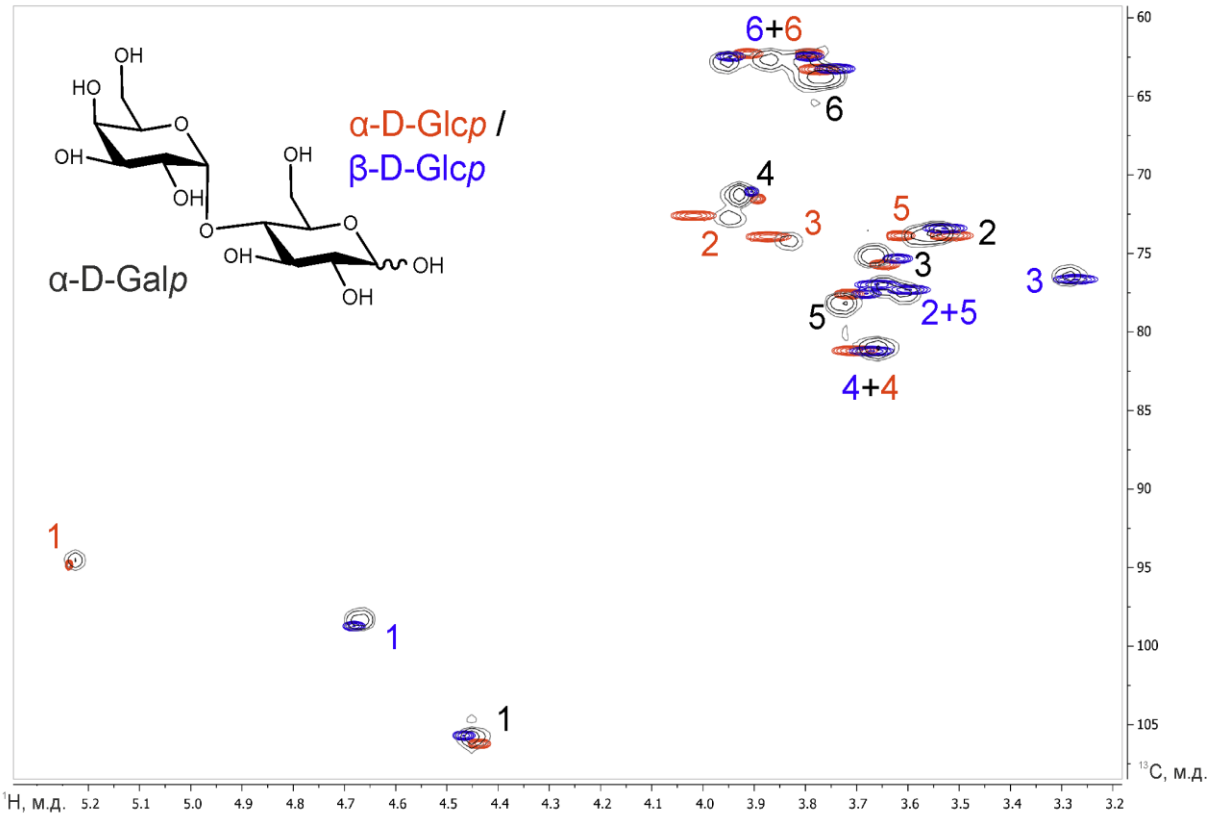



Рис. 37. Симулированные HSQC спектры α-лактозы (красный) и β-лактозы (синий), наложенные на экспериментальный спектр смеси аномеров лактозы. Экспериментальный спектр (в DMSO) был получен с поправкой на DSS/TMS из записи **bmse000938** в Biological Magnetic Resonance Data Bank [277].

В то время как существующие инструменты визуализируют сигналы HSQC в виде крестиков на пересечении протонного и углеродного химических сдвигов, разработанный инструмент предсказывает также и приблизительные размеры и форму кросс-пиков. Для этого используется эмпирический алгоритм оценки ширины сигнала по протонной оси, основанный на грубом предсказании констант спин-спинового взаимодействия (КССВ), исходя из валентных и торсионных углов между соседствующими протонами (Табл. 16). Отдельные слагаемые КССВ усреднены из литературы [278, 279]; их сумма определяет ширину кросс-пика с учетом частоты спектрометра. Ширина кросс-пиков по углеродной оси принята за 125 Гц, исходя из удобства визуализации.

Табл. 16. Приблизительная оценка ширины кросс-пиков по оси 1H.

| *взаимодействие* | *расположение протонов в структуре* | *слагаемое в КССВ, Гц*[*] |
|---|---|---|
| геминальное | при $sp^3$-C | 9 |
| | при $sp^2$-C | 3 |
| вицинальное | при наличии вращения вокруг связи C-C | 6 |
| | при двойной связи | 9 |
| | диаксиальное в пиранозах | 9 |
| | аксиально-экваториальное или диэкваториальное в пиранозах | 3 |
| | анти-перипланарное в фуранозах | 6 |
| | син-перипланарное в фуранозах | 9 |
| | циклический и первый экзоциклический (напр., H5 и группа $CH_2OH$ в пиранозах) | 6 и 3 |
| дальнее | любое | 0 |

[*] с шагом 3Гц

В генерируемых таблицах отнесения сигналов приводятся результирующие значения химических сдвигов, достоверности предсказания, использованные теоретические эффекты, ссылки на пути и веса обощений и ссылки на использованные для усреднения записи в CSDB, от которых можно перейти к оригинальным публикациям. При визуализации каждому остатку (спиновой системе) в таблице отнесения приписывается цветовой код. Сигналы в двумерных спектрах раскрашиваются в соответствии с этим кодом и подписываются номерами взаимодействующих атомов в остатке (Рис. 38). В эксперименте HMBC, содержащем трансгликозидные корреляции, сигналы и подписи могут быть окрашены в два цвета. Кросс-пики имеют форму эллипса, за исключением кросс-пиков с противоположной фазой (визуализированы в форме прямоугольников); в edHSQC фаза зависит от числа протонов при углеродном атоме.

Для сравнительной оценки достоверности положений сигналов спектр может быть раскрашен и подписан в соответствии со значениями достоверности.

Более подробно с получением двумерных ЯМР-моделей и с инструментами работы с ними можно ознакомиться в публикации [119].

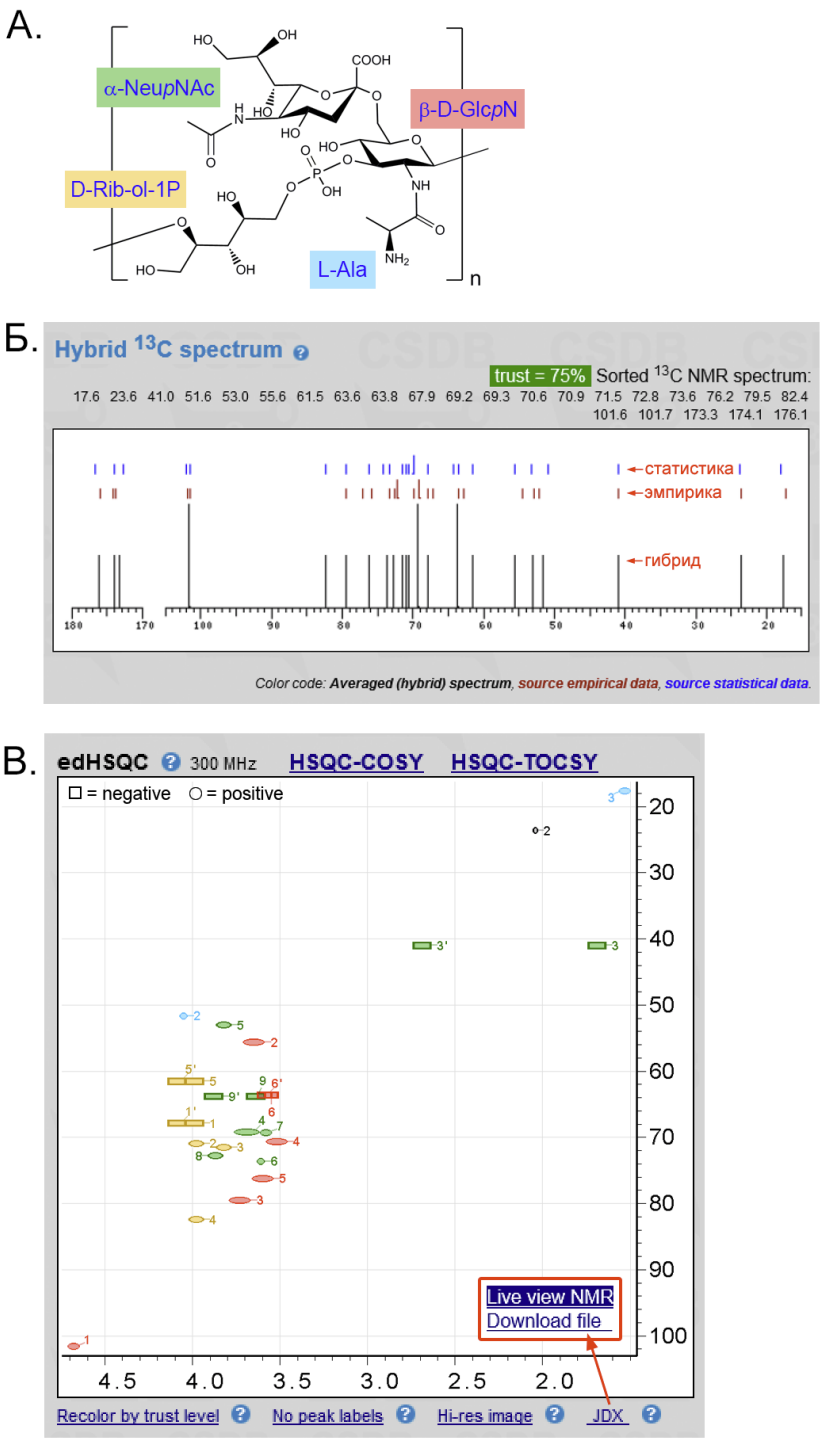



Рис. 38. Вывод предсказанных спектров для модельной структуры, содержащей остатки типичных для биогликанов классов. А. Структура и цветовой код. Цветовой код выводится в таблицах отнесения сигналов (не показаны). Б. Одномерный спектр ЯМР $^{13}$C. В. edHSQC как один из предсказанных двумерных спектров, отображённый в режиме отнесения. Цвет кодирует остаток, число – номер атома в остатке.

### *3.4.2. Прогнозирование строения природных гликанов*

Завершение работ над методами точного моделирования спектров ЯМР углеводов позволило разработать алгоритм и программное обеспечение GRASS (Generation, Ranking and Assignment of Saccharide Structures, «Генерирование, ранжирование и отнесение сахаридных структур») для автоматического предсказания структур биогликанов на основании данных ЯМР и других экспериментов.

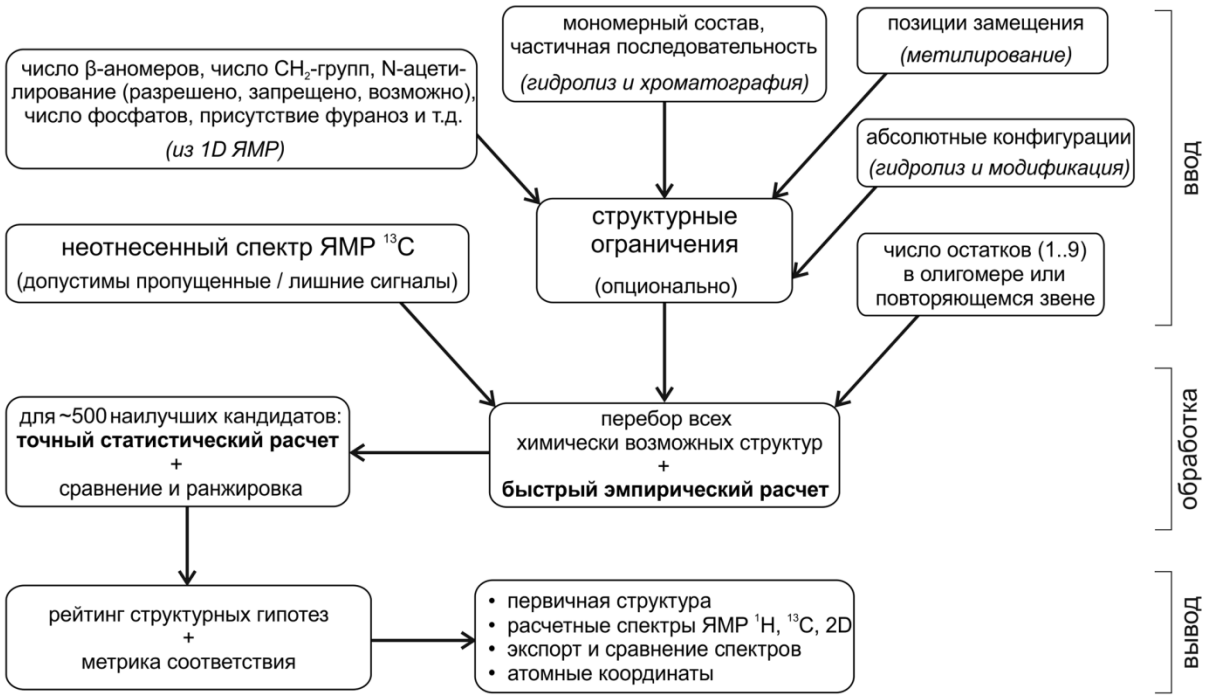



Рис. 39. Входные данные (секция «ввод»), выходные данные (секция «вывод») и последовательность шагов (секция «обработка») при генерировании структурных гипотез с помощью алгоритма GRASS.

Идея такого предсказания (см блок-схему на Рис. 39) включает следующие шаги:

1. Формирование списка структурных ограничений на основании данных ЯМР, ГЖХ, метилирования и других экспериментов. Чем больше данных о структуре получено, тем более точно предсказываются недостающие данные.
2. Перебор всех химически возможных структур, удовлетворяющих заданным ограничениям (см. ниже), и их ранжирование по степени соответствия между неотнесённым экспериментальным спектром ЯМР $^{13}C$ и моделью, полученной быстрым эмпирическим методом (на основе BIOPSEL [116], описанного в кандидатской диссертации автора).

3. Моделирование спектров для 500 лучших гипотез относительно медленным, но более точным статистическим методом (GODDESS [118]), оценка достоверности моделирования и соответствия модели эксперименту, и окончательное выявление наиболее вероятных структурных гипотез.

Обязательными входными данными являются число остатков в олигомере или повторяющемся звене полимера и неотнесённый экспериментальный спектр ЯМР $^{13}C$ водного раствора образца. Для увеличения точности предсказания и получения заметной разницы между соседними гипотезами в результирующем рейтинге (за счёт уменьшения общего числа гипотез) рекомендуется использование как можно большего числа структурных ограничений, как минимум на мономерный состав. Типичным вариантом использования GRASS является установление особенностей строения (топологии структуры, последовательности остатков, аномерных конфигураций), с трудом поддающихся анализу без полного отнесения всех спектров ЯМР, на основании информации о структуре, относительно легко получаемой из экспериментов ГЖХ, метилирования и одномерных спектров ЯМР (хотя бы частичный мономерный состав и позиции замещения остатков). Ниже перечислены поддерживаемые структурные ограничения и в скобках - эксперименты, из которых можно получить эти ограничения. Каждое из них может быть полным, т.е. охватывать все остатки в структуре, или частичным:

- Мономерный состав, в том числе классы остатков (напр., «любая гексоза») (ГХ, ГЖХ, ВЭЖХ, МС).
- Структурная единица: олигомер или повторяющееся звено полимера (хроматография в процессе выделения образца).
- Общее число β-моносахаридов в структурной единице (подсчёт дублетов в аномерной области спектра ЯМР $^{1}H$ или анализ прямых КССВ: в $^{4}C_{1}$-конформерах пираноз $^{1}J_{CH}$ ~ 170 Гц для α-аномеров, и ~160 Гц для β-аномеров, в $^{1}C_{4}$-конформерах – наоборот [280]). Аномерная конфигурация любого из остатков также может быть задана в явном виде.

- Наличие или отсутствие фураноз (наличие сигналов в области 81-88 м.д. в спектре ЯМР $^{13}C$ [281]). Способ циклизации любого остатка (пираноза, фураноза или линейная форма) также может быть задан в явном виде.
- Ограничения на ацетилирование аминогрупп для каждого остатка: «обязательно», «возможно» или «запрещено» (подсчёт сигналов в области 23-24 м.д. спектра ЯМР $^{13}C$ де-О-ацетилированного образца или подсчёт сигналов в области 50-56 м.д. спектра HSQC, дрейфующих при изменении pH [282]).
- Позиции замещения каждого остатка и/или общее число его заместителей (эксперимент по метилированию [283]), а также подтверждённое положение остатка на восстанавливающем или невосстанавливающем конце.
- Абсолютные конфигурации остатков (ГХ продуктов гидролиза, модифицированных хиральными агликонами [284]).
- Общее число $CH_2$-групп в структурной единице – имеет смысл при неполном мономерном составе (ЯМР APT, DEPT-135 [285]).
- Известные частичные последовательности остатков (анализ продуктов гидролиза).
- Количество остатков фосфорной кислоты (подсчёт сигналов в спектре ЯМР $^{31}P$).
- Глубина поиска - обзорная или детальная (на основании здравого смысла).

Использование обзорной глубины поиска (режим “widespread”, рекомендуемый в большинстве случаев) исключает из перебора экзотические структуры, накладывая дополнительные ограничения:

- При интерпретации классов остатков в мономерном составе используются только те остатки, которые встречаются в базе CSDB более, чем в 20 структурах. Это оставляет около 200 остатков из общего числа ~2300.
- Для оптически активных остатков с неизвестной абсолютной конфигурацией используются только распространённые энантиомеры[a] (D для глюко-

[a] На основании частоты встречаемости в CSDB.

зы; L для большинства аминокислот; D и L для рамнозы, и т.д.), если явно не указано обратное.

- Для моносахаридов с неизвестным способом циклизации используются только распространённые размеры цикла (пираноза для глюкозы, фураноза для фруктозы и т.д.) , если явно не указано обратное.
- Исключаются топологии, содержащие сильноразветвлённые узлы (более двух входящих связей, кроме связей с моновалентными остатками типа уксусной кислоты).
- Исключаются полимеры, в которых общее число немоновалентных остатков в боковых цепях превышает таковое в основной цепи.
- Исключаются связи между двумя аномерными центрами (напр., фрагмент Glc(1-1)Glc ).

Для получения разнообразия структурных гипотез используется пятиступенчатый итеративный поиск (Рис. 40). Ограничения, применяемые для фильтрации результатов на каждом шаге, перечислены на рисунке под метками «фильтры».

На первом шаге в соответствии с общим числом остатков создаются несвязанные узлы (наборы остатков). Каждый набор содержит все возможные комбинации типа мономера, аномерной и абсолютной конфигурации и способа циклизации, в соответствии с заданными ограничениями. В примере, показанном на рисунке, ограничения (зелёный текст в блоке 1) приводят к появлению трёх наборов A, B и C , содержащих 1, 4 и 2 возможных остатка, соответственно.

Второй шаг включает получение всех возможных топологий для известного числа остатков в структурной единице. Топология – это направленный граф («дерево»), отражающий связность остатков без учёта их природы. Топологии кодируются строками, где позиция символа (слева направо, начиная с единицы) означает порядковый номер узла, а сам символ – номер узла, в который рассматриваемый узел образует исходящую связь. Для узлов на восстанавливающем конце используется символ 0. Браузер топологий доступен на сайте проекта[a].

[a] http://csdb.glycoscience.ru/biopsel/topology.php?db=database

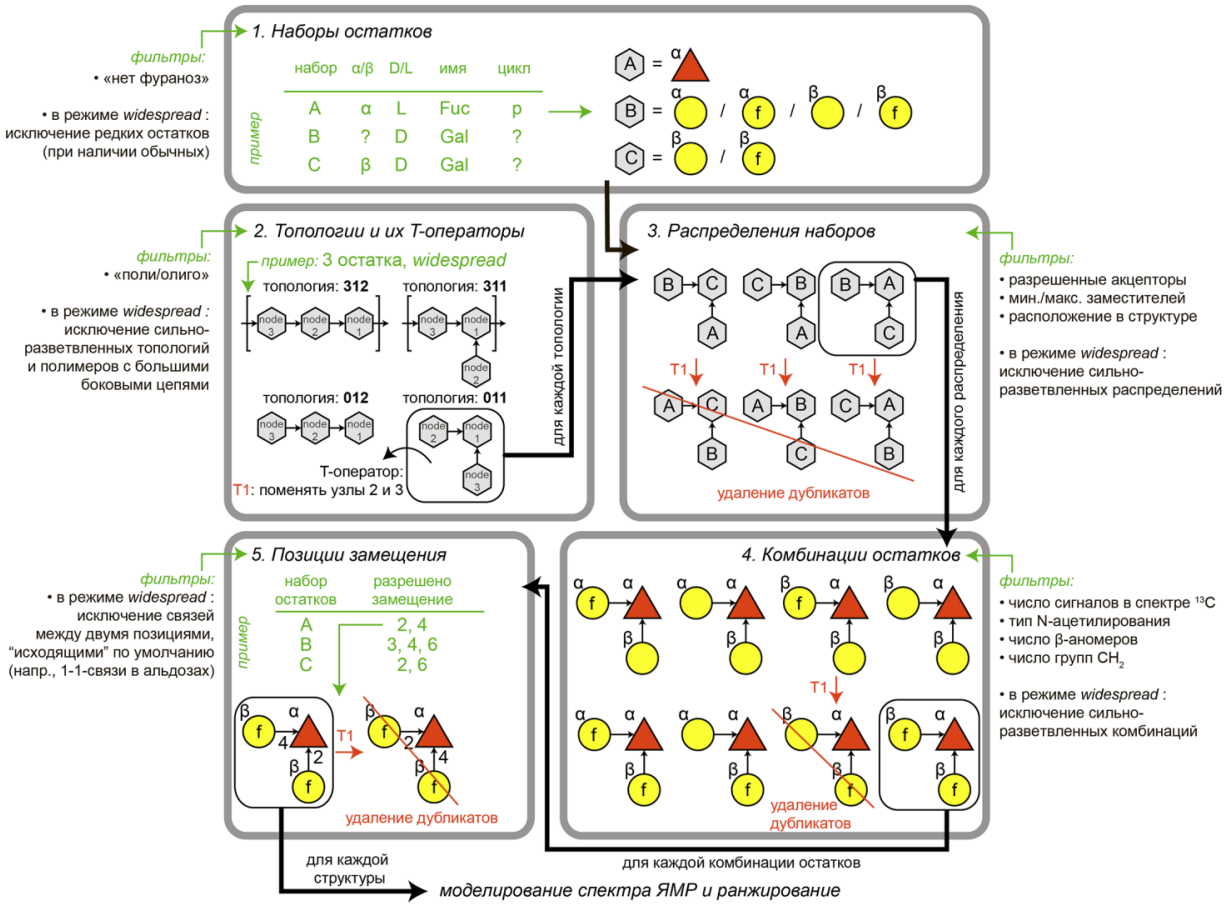
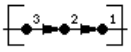


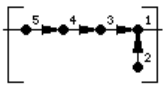

**Рис. 40.** Алгоритм генерирования разнообразия структур на примере трисахарида, определяемого ограничениями, приведёнными зелёным текстом. Структуры и остатки показаны в нотации SNFG. Наборы остатков обозначены шестиугольниками. На каждом шаге показана только одна из возможных ветвей алгоритма. Она соответствует объекту, обведённому чёрной рамкой на предыдущем шаге.

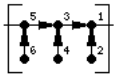
Примеры простейших топологий:

 линейный трисахарид: **012**;

 разветвлённый трисахарид: **011**;

 линейное трисахаридное повторяющееся звено полимера: **312**;

 пентасахаридное повторяющееся звено с боковой цепью из одного остатка: **51134**

 гексасахаридное повторяющееся звено с боковой цепью из одного остатка на каждом из трёх остатков основной цепи: **511335**;

Для каждой топологии, удовлетворяющей структурным ограничениям, вырабатываются операторы транспозиции (Т-операторы). Например, структура α-D-Gal*p*-(1→4)-[β-D-Man*p*-(1→3)-]β-D-Glc*p*, имеющая топологию **011**, не изменится если поменять местами узлы 2 (галактоза) и 3 (манноза) вместе с позициями замещения, поэтому для топологии **011** существует Т-оператор, меняющий местами узлы 2 и 3. В блоке 2 на Рис. 40 он обозначен как Т1. На последующих шагах Т-операторы используются для как можно более раннего отсечения ветвей генератора гипотез, приводящих к одинаковым структурам.

Третий шаг использует результаты первого и второго шагов, чтобы получить все возможные распределения наборов остатков по узлам топологий, убирая эквивалентные распределения с помощью Т-операторов. На рисунке этот шаг показан для топологии **011** (разветвлённый трисахарид).

На четвёртом шаге наборы остатков в каждом распределении наборов по узлам топологий превращаются в конкретные остатки и их конфигурации в соответствии с данными, полученными на первом шаге. Дубликаты, которые могут возникнуть при подстановке остатков в наборы, удаляются Т-операторами (один из примеров показан в 4-м блоке Рис. 40). Результирующие объекты называются сочетаниями остатков.

На пятом шаге каждое сочетание остатков приобретает все возможные химически разрешённые комбинации позиций, в которых образованы связи. На выходе мы имеем минимально возможный набор полностью определённых структур, исчерпывающий заданные структурные ограничения. Каждая из них отправляется на вход модуля моделирования спектров для последующего ранжирования.

Модели, уточнённые для 500 лучших гипотез, анализируются по степени соответствия экспериментальному спектру с помощью трёх численных параметров: коэффициента линейной корреляции, среднеквадратичного отклонения и степени достоверности, усреднённой по всем сигналам, предсказанным для данной структуры. Так как экспериментальные спектры могут содержать трудно обнаруживаемые сигналы четвертичных углеродных атомов, а также иметь нарушения аддитивности интегральных интенсивностей при совпадении двух и более сигналов, разработанная метрика учитывает степень совпадения размеров

спектров (числа сигналов), внося «штраф» за каждую единицу разницы. Для сравнения неравных спектров перебираются все возможные подспектры большего спектра, равные по размеру меньшему спектру и в качестве результирующей метрики выбирается наилучшее значение. Это делает механизм ранжирования толерантным к экспериментальным спектрам с пропущенными или лишними сигналами. Более подробно с анализом влияния неточной оцифровки экспериментального спектра на предсказательную силу алгоритма GRASS можно ознакомиться в секции 3.2 публикации [120] на примере глюкуроноксиломаннана *Cryptococcus neoformans* серотипа A, спектр которого содержит характерные признаки, затрудняющие измерение химических сдвигов - сигналы карбоксильных групп и полностью совпавшие сигналы.

Для тестирования точности и производительности созданного подхода использовалась случайная выборка из 556 полностью определённых структур биогликанов[a] от моно- до нонасахаридов с опубликованными спектрами ЯМР $^{13}C$ водных растворов. Источником структур являлась база CSDB и статьи в журнале Carbohydrate Research за 2013-2018 гг. В случае присутствия тестируемой структуры в базе CSDB точность предсказания оказалась бы завышенной по сравнению с точностью предсказания произвольной структуры по причине нахождения точно совпадающих молекулярных фрагментов при моделировании спектров. Во избежание этого искажения каждая тестируемая структура виртуально удалялась из базы перед тестом. Точность предсказания была протестирована для указанной выборки структур в восьми режимах, отличающихся соотношением известных и неизвестных (предсказываемых) параметров (Табл. 17).

Во всех режимах известными считались мономерный состав, типы циклизации, абсолютные конфигурации, тип структурной единицы (полимер или олигомер) и, для олигомеров – какой остаток находится на восстанавливающем конце.

[a] Полный список структур, их спектры и литературные ссылки приведены в Табл. S3 в дополнительных материалах к статье 120) Kapaev R. R., Toukach P. V. GRASS: semi-automated NMR-based structure elucidation of saccharides // Bioinformatics. – 2018. – Т. 34, № 6. – С. 957-963.: Supplementary Materials

Табл. 17. Режимы тестирования GRASS*.

| *режим* | *позиции замещения* | *число заместителей* | *аномерные конфигурации* | *общее число β-сахаров* | *глубина поиска* |
|---|---|---|---|---|---|
| 1 | + | + | – | + | обзорная |
| 2 | + | – | – | + | обзорная |
| 3 | + | + | – | + | детальная |
| 4 | + | – | – | + | детальная |
| 5 | + | + | – | – | обзорная |
| 6 | + | – | – | – | обзорная |
| 7 | – | – | + | + | обзорная |
| 8 | – | – | – | + | обзорная |

* ‘+’ означает известные параметры, ‘–’ – неизвестные.

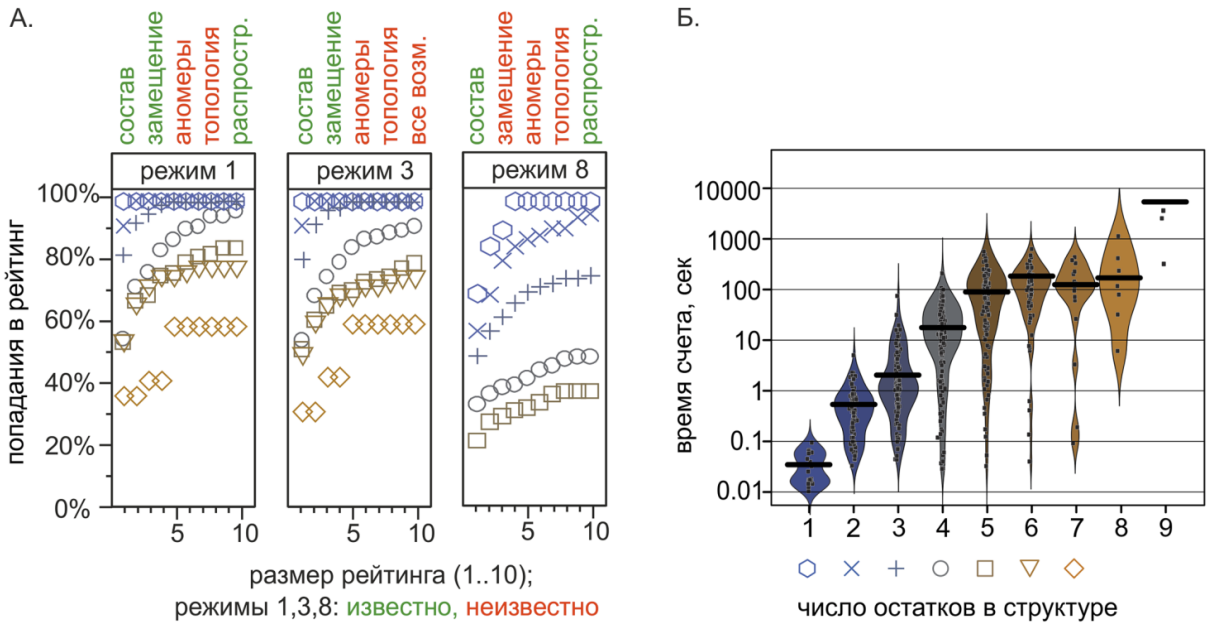



Рис. 41. Результаты валидации алгоритма GRASS на выборке из 556 полностью определённых структур, имеющих опубликованные спектры ЯМР. А. Зависимость числа попаданий правильных структур в рейтинг от размера рейтинга. Три режима отличаются количеством неизвестных параметров на входе: зелёным перечислены известные параметры, красным – предсказываемые. Здесь топология подразумевает также и последовательность остатков. Пиктограммы и цвет точек соответствуют числу остатков в олигомере или в повторяющемся звене полимера (см. ось абсцисс на рис. Б., от синего к оранжевому = от простого к сложному). Б. Статистический анализ производительности расчёта. Средние значения показаны чёрной линией, отдельные измерения – черными точками. Ширина фигуры отражает распределение выборки по времени счета.

Рис. 41А отражает процент попадания правильной структуры в рейтинг в зависимости от размера рейтинга в трёх характеристичных режимах. Наиболее распространённый режим 1, подразумевающий использование GRASS для установления последовательности остатков и их аномерных конфигураций, продемонстрировал 84%, 76% и 74% попаданий правильной структуры в рейтинг пяти наиболее подходящих гипотез для тетра-, пента- и гексасахаридов, соответственно. Правильные структуры меньшего размера, как правило предсказывались как наиболее вероятные. Для наиболее сложных случаев (гепта- и октасахариды), связанных с перебором десятков миллионов возможных структур, предсказательная сила находилась в пределах 60% и могла быть улучшена путём введения дополнительных структурных ограничений. Например, правильная структура антигена группы крови А человека (нонамер, включающий 7 моносахаридов и 2 моновалентных остатка, Рис. 42А) в режиме 1 заняла пятое место в рейтинге и переместилась на первое место, когда аномерная конфигурация фукозы и галактозамина была в явном виде указана как α.

Неуказание общего числа β-аномеров и общего числа заместителей остатка не оказало существенного влияния на точность и производительность предсказаний. Неуказание известных позиций замещения остатков вызвало среднее уменьшение точности и десятикратное уменьшение производительности, в частности по результатам перебора $18\times10^6$ структурных гипотез для антигена группы крови А человека в режиме 7, правильная структура не попала в первую десятку; расчёт занял 26 часов. Установление строения больших молекул в слабоограниченных режимах является фундаментальной проблемой, связанной с появлением огромного числа структурных гипотез с похожими спектрами ЯМР. Поэтому для однозначного предсказания строения тетрасахаридов и бóльших структур требуется указание позиций замещения, полученных из эксперимента по метилированию. Использование слабоограниченных режимов оправдано для структурных единиц, содержащих один или два остатка. Например, при отсутствии ограничений, правильная структура D-галактана I (Рис. 42Б) была предсказана как наиболее вероятная гипотеза из 300000 проверенных; расчёт занял 28 минут.

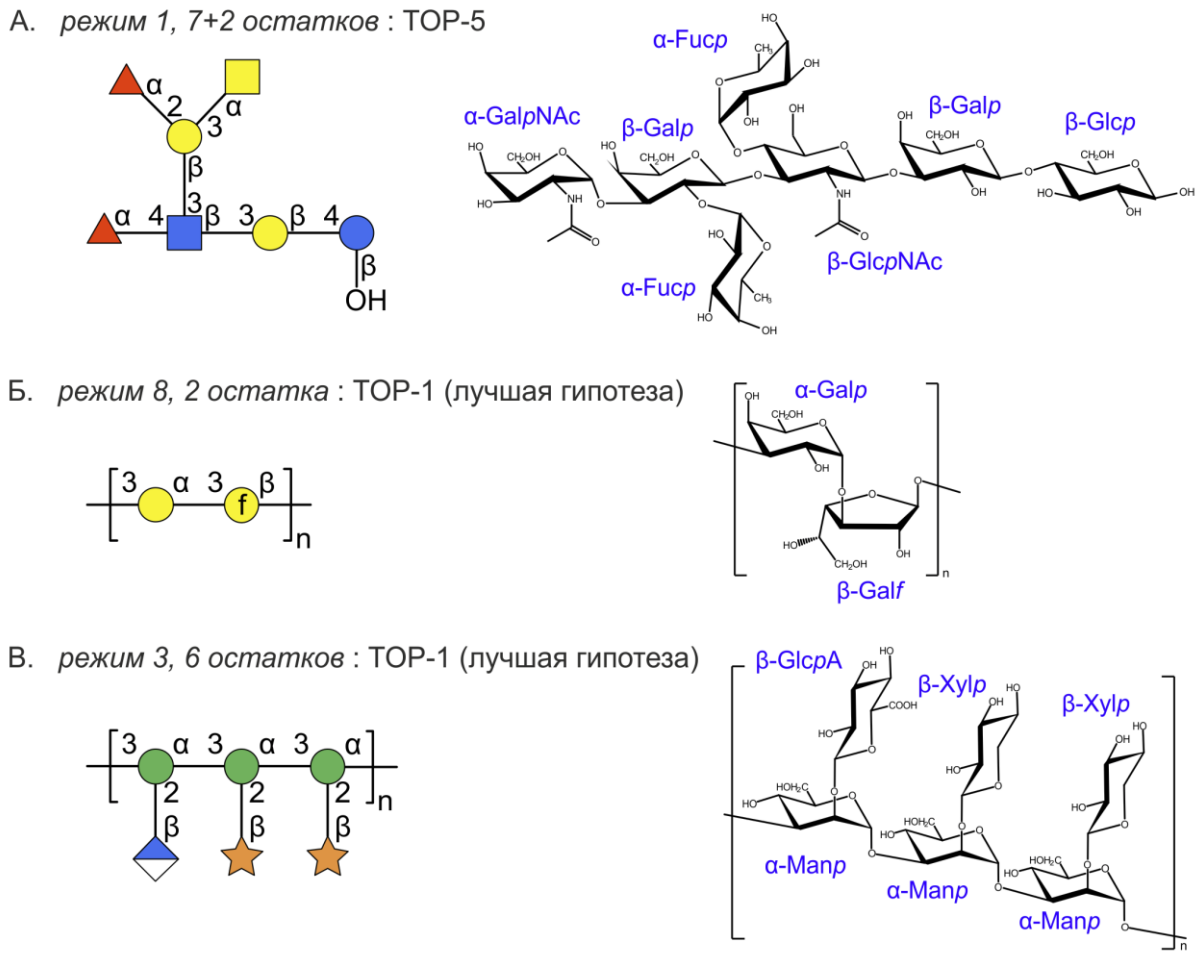



**Рис. 42.** Результаты предсказания модельных структур по экспериментальным спектрам и данным ГЖХ: А. Антиген группы крови А человека (известен состав и позиции замещения; предсказана последовательность и аномерные конфигурации). Б. D-галактан I (известен только состав и данные ЯМР, предсказано все остальное, включая размеры циклов). В. Ксиломаннан *Cryptococcus neoformans* серотипа А (известен состав и позиции замещения; предсказана последовательность и аномерные конфигурации).

Неполнота входных данных (зашумлённые спектры, неточные интегралы сигналов, неполный мономерный состав, отсутствие некоторых позиций замещения и т.д. ) и наличие структурных особенностей, плохо поддающихся эмпирическому расчёту спектров ЯМР, не являются препятствиями для работы алгоритма, лишь количественно влияя на достоверность предсказания. Данный тезис проиллюстрирован примером глюкуроноксиломаннана *Cryptococcus neoformans* серотипа А (Рис. 42В), чьё гексасахаридное повторяющееся звено предсказано как наилучшая гипотеза даже при детальной глубине поиска (режим 3). Подобные структуры, содержащие остатки в близком химическом окружении, тради-

ционно являются сложной задачей для ЯМР-моделирования, особенно при наличии дизамещения в соседних положениях, стерически влияющего на конформации гликозидных мостиков, а следовательно, и на химические сдвиги сигналов (см. остатки маннозы на Рис. 42В).

Веб-интерфейс GRASS позволяет пользователю ввести структурные ограничения и запустить поиск. Производительность расчёта является краеугольным камнем предсказаний, связанных с перебором вариантов. Тем не менее, статистика по времени работы алгоритма на вычислительных мощностях сервера CSDB, сравнимых с персональным компьютером (Рис. 41Б), показывает что для большинства типичных задач время счета укладывается в несколько минут. Так как при минимуме ограничений и наличии большого числа замещаемых позиций в отдельных остатках расчёт может занимать длительное время, выполнение задания сервером не синхронизировано с пользовательской сессией. Если сессия все ещё открыта, пользователь получает результат непосредственно в ней, в противоположном случае он получает email-уведомление об окончании расчёта со ссылкой на результат на странице GRASS в Интернет. Результат содержит список структурных гипотез, отсортированных по метрикам соответствия. Каждая из них сопровождается наложением смоделированного спектра ЯМР $^{13}C$ на экспериментальный и ссылками на инструменты для дальнейшей работы с гипотезой: отнесением сигналов её одно- и двумерных спектров ЯМР и предсказанием молекулярной геометрии для оценки ЯЭО двумя способами – с помощью сервиса Sweet-II [164] и с помощью собственной разработки, описанной в разделе 3.3.4.

По результатам направленного исследования способов ЯМР-моделирования углеводов, выполненного автором [166], можно заключить, что общехимические подходы к автоматизации связи «структура - спектр» неприменимы даже к простым углеводам. Из существующих специализированных углеводных подходов только CASPER [286] обладает достаточной предсказательной силой. Тем не менее, использование CASPER принципиально ограничивается небольшим набором структурных особенностей, характерных для биогликанов млекопитающих. Многие компоненты структуры, поддерживаемые в GRASS, не поддерживаются в CASPER: фуранозы; высшие сахара, кроме нейраминовой

кислоты; пиранозы, встречающиеся в гликанах прокариот (напр., бациллозамин, 6-дезокситалоза и т.д.); полиолы; аминокислоты; жирные кислоты и сфингоиды; остатки фосфорной кислоты; амидные и сложноэфирные связи между остатками. Для сравнения предсказательной силы GRASS и CASPER такие структуры были исключены из выборки, после чего предсказания двумя подходами были сравнены по количеству попаданий правильной структуры в рейтинг. Для сравнения были использованы одинаковые ограничения и выбран режим 3, так как CASPER не имеет опции обзорной глубины поиска, присутствующей в других режимах. В среднем, для структур, поддерживаемых обоими методами, GRASS показал превосходство, особенно для сложных случаев (Рис. 43). В частности, две модельные структуры, представленные на Рис. 42А и Рис. 42В, не попали в десятку лучших гипотез, предсказанных CASPER, а модельная структура на Рис. 42Б не поддерживается итератором CASPER, потому заведомо не могла попасть в рейтинг. Достоверность сравнения для гепта- и октасахаридов меньше, чем для остальных структур, из-за малого размера выборки структур, поддерживаемых обеими программами.

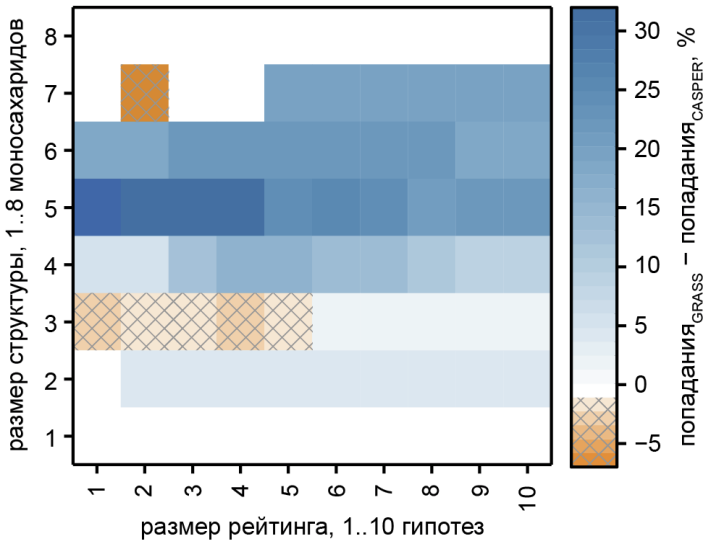



Рис. 43. Сравнение предсказательной силы GRASS и CASPER. Плотность тона соответствует разнице в проценте попаданий правильной структуры в список лучших гипотез (синие тона – превосходство GRASS, оранжевые тона – превосходство CASPER, белый цвет – одинаковые результаты)

Созданный инструмент генерирования и оценки качества структурных гипотез в гликохимии востребован при установлении первичной структуры при-

родных углеводов и их производных. Этот процесс не является полностью автоматическим, так как требует от исследователя осмысления результатов и строгого доказательства полученных ответов, но тем не менее существенно снижает трудозатраты и требования к квалификации исследователя. Более подробно с алгоритмом и инструментом GRASS можно ознакомиться в публикации [120].

### *3.4.3. Анализ распределения структурных особенностей*

Статистический анализ позволяет получить знания, неявно содержащиеся в базах данных. Систематическое сравнение гликомов различных таксономических групп и выявление связей «структура – таксономия» позволило создать базис для хемотаксономической классификации организмов, использующей специфичность синтезируемых ими углеводов, что особенно востребовано для микроорганизмов, так как их иммунологические свойства часто определяются углеводными антигенами [287]. Для решения этой задачи был создан инструмент статистического анализа содержимого базы CSDB, позволяющий изучить распределение встречаемости небольших структурных фрагментов в углеводах заданных таксономических групп на уровне доменов, типов, классов, родов, видов и подвидов/штаммов. Инструмент имеет веб-интерфейс и различные фильтры, позволяющие ограничивать анализ по следующим параметрам:

- таксоны, содержащие сравниваемые структуры (от царств до видов);
- размер фрагмента (мономер или димер) и его разветвлённость (количество заместителей);
- положение фрагмента в структурах (терминальное, на восстанавливающем конце, любое);
- уникальность фрагмента для выбранной таксогруппы более высокого ранга (например, поиск фрагментов, уникальных для структур определённого рода среди всех структур в типе или царстве, включающем данный род);
- наличие во фрагментах неопределённых аномерных или абсолютных конфигураций или размеров циклов;
- наличие во фрагментах агликонов, моновалентных заместителей и прочих неуглеводных компонентов структуры;
- считать ли разными фрагменты, отличающиеся только аномерной конфигурацией остатков.

Предлагаемые области применения созданных алгоритмов и программ включают поиск характерных углеводных маркеров в пределах таксонов, в частности, специфических антигенных детерминантов, а также исследования активностей гликозилтрансфераз в различных таксономических группах.

Созданный инструмент был применён для сравнительного анализа моно- (Рис. 44) и дисахаридных (Рис. 45) строительных блоков в гликанах бактерий, растений и грибов, а также для выявления наиболее распространённых димеров с неуглеводными остатками в растительных гликозидах (Рис. 46) и структурных фрагментов, уникальных для каждой большой таксономической группы патогенных микроорганизмов.

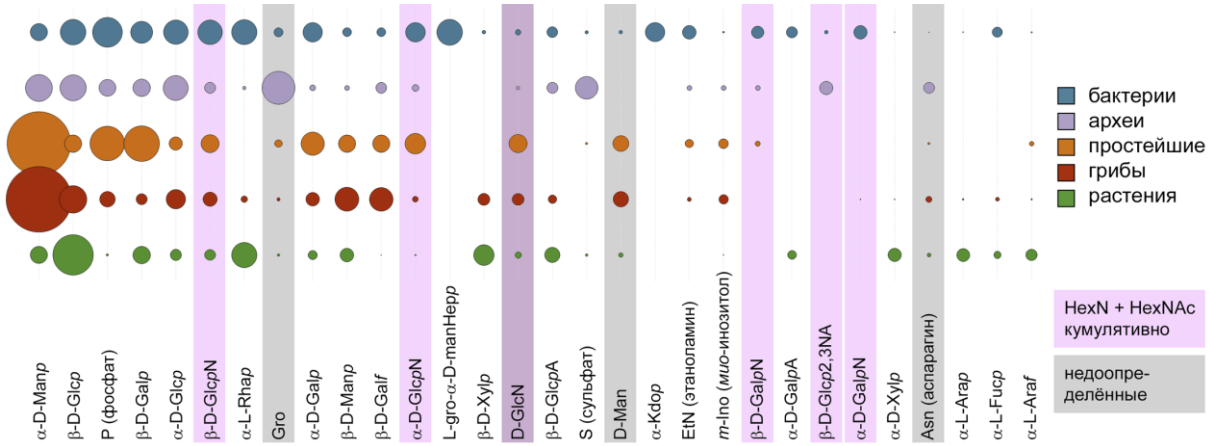



Рис. 44. 30 самых распространённых мономерных остатков в углеводных структурах из организмов пяти таксономических групп. Площадь кругов соответствует усреднённой частоте встречаемости мономера в домене, приведённой к числу структур. Остатки 2-аминосахаров включают ацетилированную и неацетилированную формы и выделены лиловым; недоопределенные остатки с неизвестной конфигурацией (аномерная, абсолютная, размер цикла) выделены серым.

Частоты встречаемости фрагментов получены путём нормализации числа мономеров и димеров на общее число структур из организмов, относящихся к соответствующему домену; неуглеводные моновалентные остатки (метанол, уксусная кислота) не принимались во внимание по причине их распространённости и неспецифичности. Детальное описание настроек и результатов статистического анализа представлено в публикации [31]. На Рис. 44 приведены 30 наиболее распространённых мономерных остатков, характерных для углеводных структур из организмов, принадлежащих к шести доменам (бактерии, археи, простейшие, грибы, растения, водоросли), представленным в CSDB. Распределение мономеров подтверждает, что бактериальные гликаны наиболее разнообразны по мономерному составу. Благодаря этому разнообразию бактерии занимают множество различных экологических ниш и выдерживают давление отбора, вызванное кон-

куренцией и иммунной системой организма-хозяина. К наиболее распространённым мономерам бактерий относятся: условное обозначение неустановленного липида А (этот «псевдомономер» на Рис. 44 не представлен), D-галактуроновая кислота, L-рамноза, L-глицеро-D-манногептоза и 3-дезокси-D-маннооктулозоновая кислота (Kdo). Последние два компонента структуры присутствуют в липополисахаридах грамотрицательных бактерий и отсутствуют в углеводах организмов из других доменов. Благодаря созданному инструменту, подобные «очевидные», но не доказанные знания были строго подтверждены статистически [26].

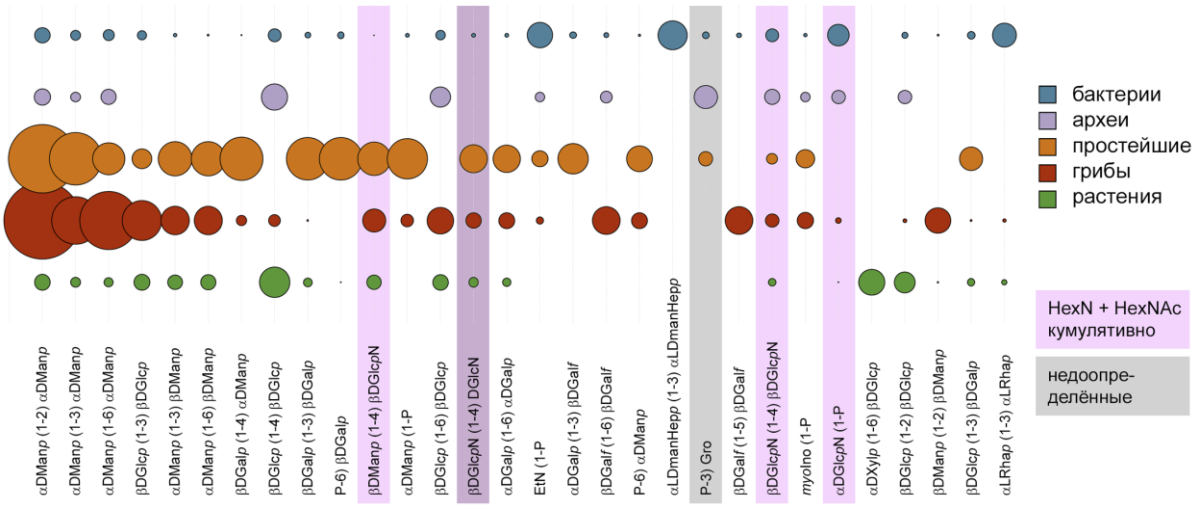



Рис. 45. 30 самых распространённых димеров в углеводных структурах из организмов пяти таксономических групп. Площадь кругов соответствует усреднённой частоте встречаемости димера в домене, приведённой к числу структур. Димеры, содержащие 2-аминосахара, включают ацетилированную и неацетилированную формы, они выделены лиловым; димеры, содержащие недоопределенные остатки с неизвестной конфигурацией (аномерная, абсолютная, размер цикла), выделены серым. Водоросли объединены с растениями (они считаются отдельным доменом в CSDB, но в настоящий момент уникальных дисахаридов для них не выявлено по причине недостаточной представленности структур в базе). Димеры, содержащие полиолы, чаще всего являются аналитическими артефактами[a] и не включены в рассмотрение.

В отличие от млекопитающих, синтезирующих всего два модифицированных моносахарида, отсутствующих у прокариот (2-ацетамидо-2-дезокси-6-О-

[a] Восстанавливающие концы олигомеров после расщепления полисахаридов по Смиту.

сульфо-β-D-глюкопиранозу и 5-гликолил-α-нейраминовую кислоту), для многих таксономических групп в пределах царства бактерий в результате проведённого анализа были выявлены уникальные эпитопы [26]. Современные вакцины часто используют уникальность углеводов, экспонированных на поверхности патогенов, в том числе на уровне моносахаридов, поэтому анализ уникальности гликоэпитопов – необходимый шаг для систематического поиска кандидатов в антибактериальные вакцины. Отдельные моносахариды характерны для своей таксономической группы только в определённом положении, напр., терминальные остатки 5-ацетил-α-нейраминовой кислоты в гликоконъюгатах млекопитающих являются медиаторами межклеточного взаимодействия или рецепторами патогенов [288], а их экспозиция на поверхности бактерий может маскировать бактерии для иммунной системы высшего организма. Другим характерным примером стало выявление того, что остатки глюкозы, часто встречающиеся на невосстанавливающих концах бактериальных сахаридов, никогда не занимают это положение в поверхностных O- и N-гликанах млекопитающих. Это согласуется с представлениями об эволюционной адаптации бактерий и важности терминальной глюкозы для адгезии и проникновения бактерий в эпителий, показанной на примере родов *Salmonella* и *Pseudomonas* [289].

Модификации моносахаридов в полимерах клеточной мембраны являются фактором устойчивости бактерий в разнообразным условиям среды. Например, ацетилирование аминогрупп играет ключевую роль в способности аминосахаров формировать водородные связи и нести заряд [290]. Анализ модификаций показал, что свободные аминогруппы значительно чаще встречаются в углеводах энтеробактерий, чем в других таксономических группах. По-видимому, резистентность по отношению к лизосомам организма-хозяина зависит от гидрофобности и заряда на аминогруппах аналогично тому, как это было продемонстрировано другими авторами для глюкозамина в Грам-положительных бактериях [291]. Гетерогенные в пределах вида или нестехиометрические O-ацетильные (а в случае актинобактерий – и O-метильные) группы предположительно маскируют бактериальные эпитопы за счёт изменения конформации, как это было продемонстрировано Фуско и коллегами для менингококков [292]. Модификации моносахаридов остатками этаноламина, метанола, муравьиной и пировиноградной кислот,

характерны для прокариот (в отдельных случаях - также для грибов и растений[a]), но отсутствуют у животных. С учётом того, что такие модификации не всегда кодируются в геноме, было сделано (и впоследствии подтверждено – см. раздел 2.6) предположение, что кластеризация таксонов по этим признакам может лучше соответствовать поражаемым организмам, органам и тканям, чем классическая филогенетика.

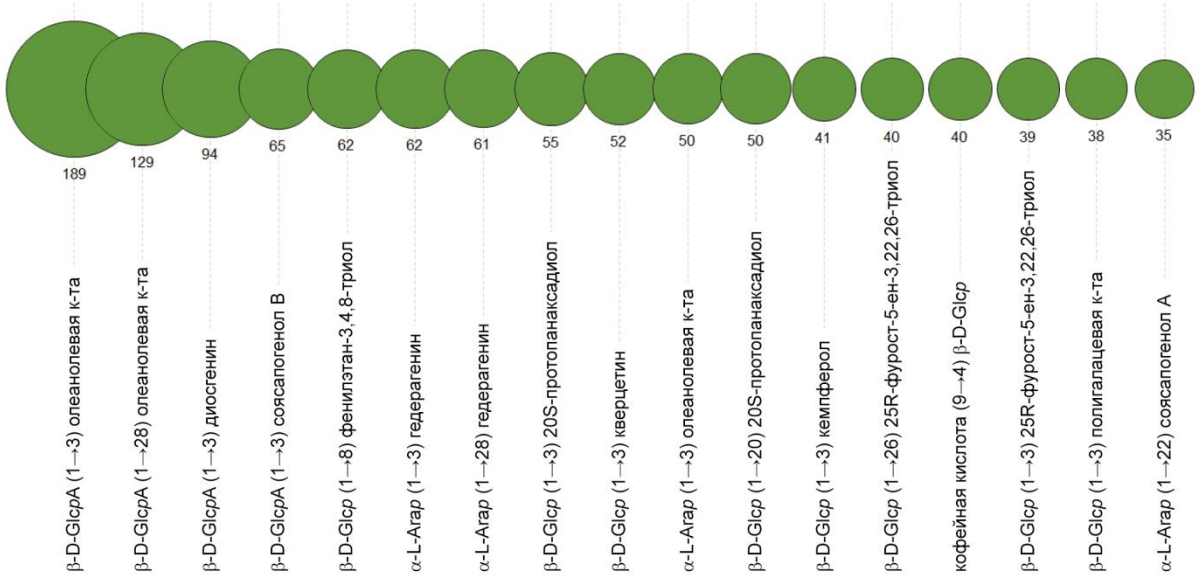



Рис. 46. 17 самых распространённых агликонсодержащих димеров в растительных гликозидах. Числа соответствуют абсолютному уровню встречаемости. Димеры, содержащие аминокислоты, в большинстве случаев соответствуют сайтам гликозилирования белков, и потому не включены в анализ.

Димерные фрагменты, уникальные для какой-либо таксономической группы, могут отражать особенности взаимодействий входящих в неё организмов с окружающей средой, реализуемых за счёт активностей специфических гликозилтрансфераз, в то время как растительные гликозиды, содержащие различные агликоны тритерпеновой, стероидной, флавоноидной и фенольной природы, обладают противораковой и другой биологической активностью и представляют значительный интерес с медицинской точки зрения. Растительные и прокариотические углеводы демонстрируют большее разнообразие гликозидных связей по сравнению с другими доменами. Доменоспецифичные ферменты син-

[a] Метанол – в составе простых и сложных метиловых эфиров моносахаридов в пектинах или глюкуроноксиланах в растениях и грибах; пировиноградная кислота – в полисахаридах водорослей.

тезируют фрагменты структуры, уникальность которых в растениях связана с положением замещения остатка-субстрата (например, α-L-рамнопиранозил-2-β-D-глюкопираноза, Рис. 45), а не с уникальными мономерными строительными блоками (как, например, дигептозные фрагменты в структурах бактериального кора). Анализ распространённости димеров и их распределения по классам патогенных микроорганизмов с последующим сравнением с пулом известных гликозилтрансфераз человека позволяет выявить уникальные микробные гликозилтрансферазы как потенциальные мишени антибиотиков. Аналогичное сравнение гликомов растений и фитопатогенных бактерий открывает возможности для избирательного химического подавления биосинтеза гликанов в бактериях без нанесения ущерба защищаемым сельскохозяйственным культурам.

Изучение разнообразия бактериальных гликомов востребовано в контексте создания автоматического синтезатора произвольного гликана из заданной группы организмов. Для оценки выборки необходимых строительных блоков (защищённых моносахаридов) было проведено сравнение разнообразия гликомов бактерий с разнообразием гликомов млекопитающих. При сравнении использовались следующие параметры:

- размер структуры или её повторяющегося звена; признак полимерности;
- многоантенность и плотность точек разветвления;
- плотность заряда и распределение типов заряженных групп;
- распределение моносахаридов в структурах и отдельно - на невосстанавливающих концах;
- распределение мономерных и димерных фрагментов, уникальных для изучаемой группы организмов;
- распределение модификаций моносахаридов (включая фосфорилирование и О-ацетилирование);
- распределение размеров углеродного скелета мономеров и способов циклизации;
- распределение доноров и акцепторов гликозилтрансфераз;
- распределение типов связей (активностей гликозилтрансфераз);

- изученность таксонов (количество публикаций и известных углеводных структур);

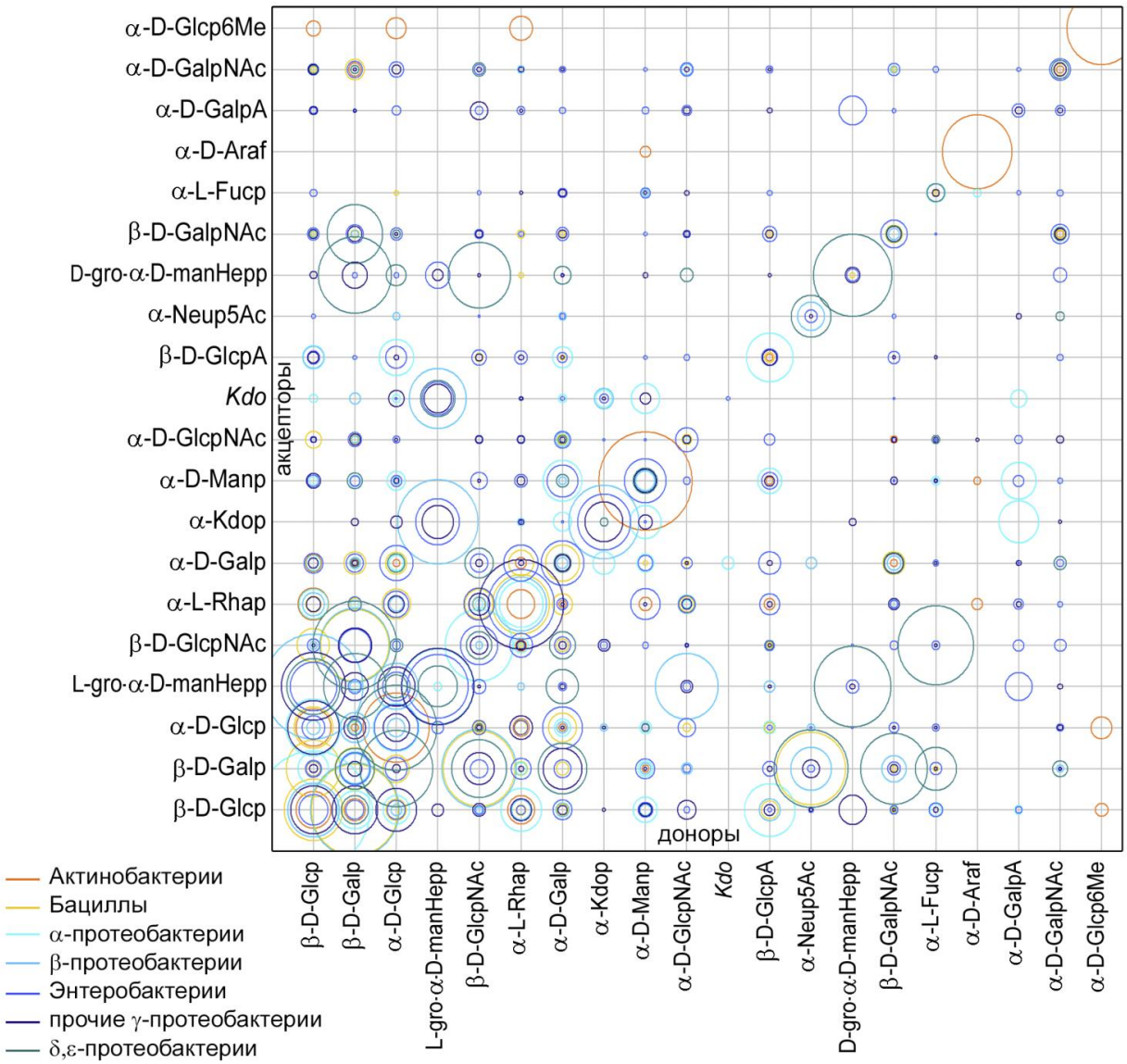



Рис. 47. Доноры и субстраты гликозилтрансфераз бактерий. Диаметр круга соответствует распространённости димерных фрагментов в структурах бактерий из таксономических групп, закодированных цветом. Компонент структуры с неопределённой аномерной конфигурацией и размером цикла (Kdo) приведён *курсивом*.

Одним из способов получения биогликанов является ферментативный синтез в бактериях с модифицированным геномом. На Рис. 47 представлен анализ доступности доноров и субстратов гликозилтрансфераз в бактериях различных таксономических групп. Эта информация востребована для выбора микроорганизмов и их ферментов, способных направленно синтезировать углеводные структуры, которые в настоящее время находят применение в биотехнологии и

медицине. Кроме того, карты встречаемости углеводных димеров позволяют выявить структуры, отсутствующие в гликомах тех или иных организмов, что, в свою очередь, позволяет осуществлять отбор микроорганизмов для экспрессии рекомбинантных гликозилтрансфераз с заданной активностью для целенаправленной наработки заданных продуктов.

С результатами сравнительного анализа доступности доноров и субстратов, а также других параметров углеводных структур, можно ознакомиться в публикации [26]. Аналогичный анализ созданными инструментами был проведён и для гликанов млекопитающих, полученных из базы GlycomeDB [72]. Для отдельных димерных фрагментов млекопитающих (D-GlcNAc(β1-3)D-GalNAc, D-GlcNAc(β1-4)D-Man, D-GlcNAc(β1-6)D-Man, D-GlcNS(α1-4)L-Ido в литературе А) не охарактеризованы гликозилтрансферазы, в том числе для последнего дисахарида, ассоциированного с амавротической идиотией Зандхофа [293] и обнаруженного более, чем в 500 записях GlycomeDB о гликанах и гликоконъюгатах гоминидов. Напротив, димерные фрагменты, отсутствующие в млекопитающих, но присутствующие в бактериях, являются кандидатами для проверки иммунноформирующего действия. Отдельные такие фрагменты уже использованы в ранее разработанных вакцинах, например, компоненты структуры капсульного полисахарида *Salmonella pneumonia* D-Glc(α1-2)D-Gal и D-Glc(β1-4)D-Gal [30].

### *3.4.4. Углеводная фенетика*

Инструмент таксономической кластеризации предназначен для выявления групп таксонов, объединённых по признаку схожести или различия синтезируемых ими углеводов. Он впервые позволил гликобиологам группировать организмы произвольных таксономических групп и рангов на основании схожести в активностях гликозилтрансфераз. Используемый подход рассчитывает встречаемость моно- и димерных фрагментов в углеводных структурах, присутствующих в организмах из заданных таксогрупп и удовлетворяющих фильтрам отбора. На основании сравнения полученных паттернов встречаемости методом Хэмминга [294] генерируются матрицы схожести для наборов структур, принадлежащих к конкретным таксонам. Далее таксоны нормализуются по степени изученности (с учётом общего количества опубликованных для них структур) и кластеризуются в родственные группы по характерным структурным признакам (см. блок-схему на Рис. 48).

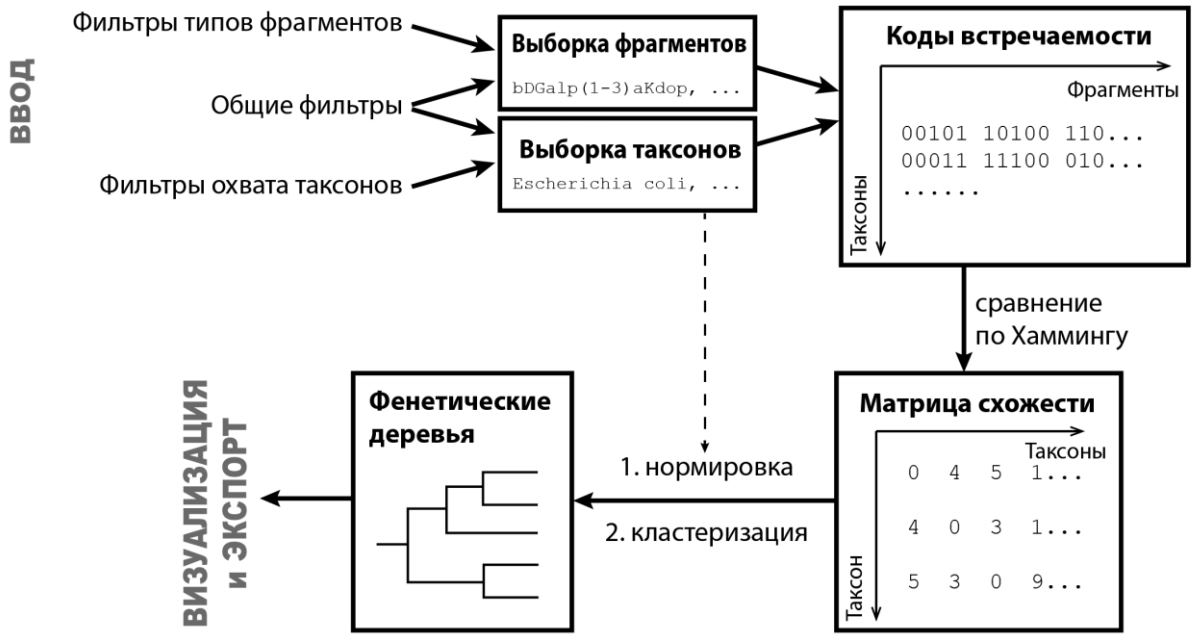



Рис. 48. Схема работы инструмента таксономической кластеризации (Taxon clustering).

Были протестированы шесть существующих алгоритмов кластеризации, хорошо зарекомендовавших себя в задачах группирования биологических видов. Их недостатки и преимущества применительно к задачам биологической кластеризации, основанной на родстве биомолекул, синтезируемых ферментативным аппаратом клеток, подробно изложены в публикации [31]. Инструмент кластери-

зации снабжён веб-интерфейсом и интегрирован в базу CSDB. Результаты кластеризации представляются в виде фенетических деревьев со значащей (дендрограммы) или незначащей (кладограммы) длиной ветвей и могут быть экспортированы в популярные филогенетические форматы Newick [295] и Nexus [296].

Данный инструмент может быть использован для выявления ферментативных активностей, вовлеченных в синтез и процессинг углеводов: организмы, синтезирующие гликаны со сходными дисахаридными фрагментами, должны обладать гликозилтрансферазами со сходными активностями. Предполагается, что подобный подход ускорит исследования углевод-активных ферментов, экспериментальное подтверждение функций которых сталкивается со значительными трудностями. В первую очередь это относится к ферментам бактерий, которые представляют особый интерес с биотехнологической точки зрения и углеводные структуры которых наиболее полно представлены в CSDB. Для иллюстрации качества кластеризации таксонов приведена дендрограмма (Рис. 49), построенная на основании анализа встречаемости димерных фрагментов в гликанах организмов, принадлежащих к родам, наиболее представленным в базе CSDB.

Кладограмма на Рис. 50А содержит результаты анализа 33 наиболее изученных видов бактерий в контексте общности их гликомов. Такой анализ позволяет выявить взаимосвязи между таксонами, не выявляемые генетически. Его результаты частично (от 44% до 53% в зависимости от алгоритма кластеризации) совпадают с классическим «деревом жизни», построенным на основании последовательностей консервативных субъединиц рибосомальной РНК (Рис. 50Б). Известно, что филогенетические деревья на основе полных геномных последовательностей и рРНК проявляют значительное сходство [297]. Различия в генах углевод-активных ферментов отражает различия в геномах, но лишь до определенной степени. В ходе эволюции давление отбора по-разному воздействует на различные гены, а следовательно, организмы, чьи геномы сильно различаются, могут, тем не менее, обладать сходными фенотипическими чертами. Более того, из-за горизонтального переноса генетической информации, характерного для прокариот, классические филогенетические деревья не всегда

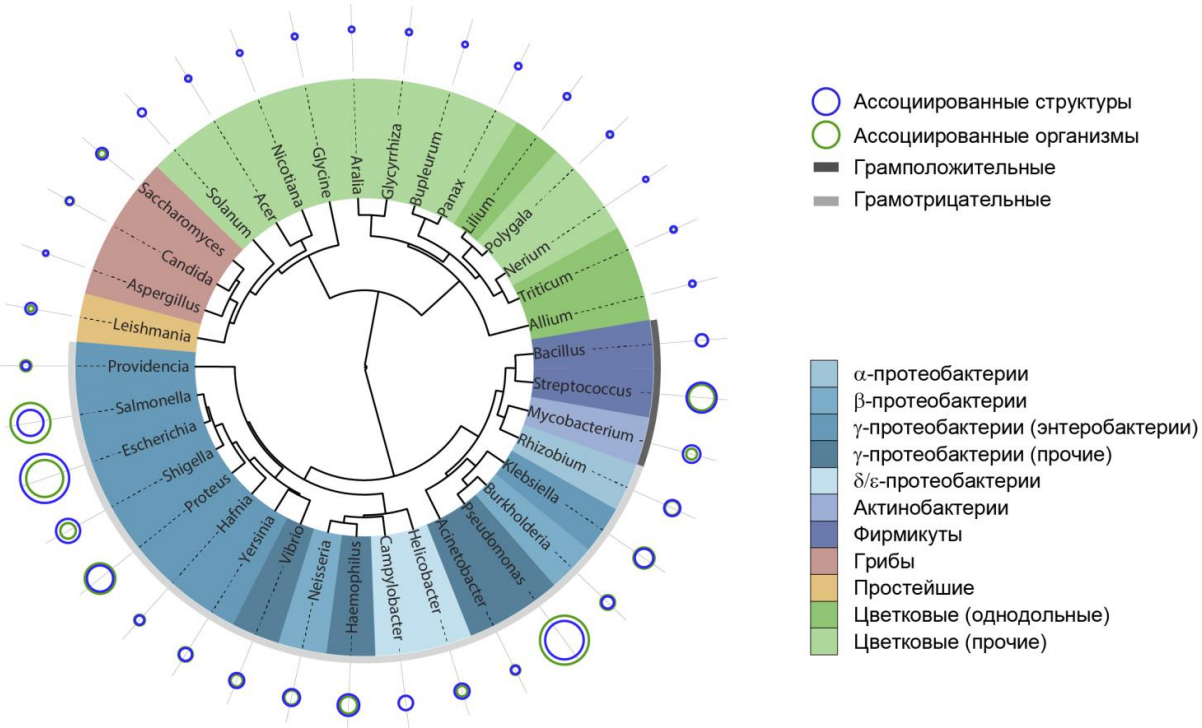



Рис. 49. Фенетическое дерево, построенное на основании результатов кластеризации наиболее представленных в CSDB родов. Оттенками синего показаны различные таксономические группы бактерий, оттенками зелёного – группы растений, красным и оранжевым – грибы и простейшие, соответственно. Цвет внешней дуги для бактерий отражает реакцию по Граму. Размер кругов соответствует нормализованной представленности данного рода в CSDB с точки зрения организмов (зелёный) и структур (синий). В тех случаях, когда зелёный круг не виден, его размер совпадает с синим.

соответствуют реальной фенетике бактерий [298, 299]. Таким образом, различия между бактериальными фенетическими деревьями, построенными на основании последовательностей сахаридов и рРНК, могут отражать тот факт, что углеводы являются основным инструментом взаимодействия бактерий с окружающей средой, продуцирование конкретных углеводных структур соответствует определённой среде обитания, и этот фактор является более значимым в контексте родства бактерий, чем генетический фактор. Так, *Neisseria gonorrheae* и *Haemophilus ducreyi*, расположенные достаточно далеко друг от друга на «дереве жизни» рРНК, но обладающие сходными гликанами, вызывают заболевания, передающиеся половым путем, обитают в мочеполовых путях человека и, очевидно, сталкиваются с близкими факторами среды.

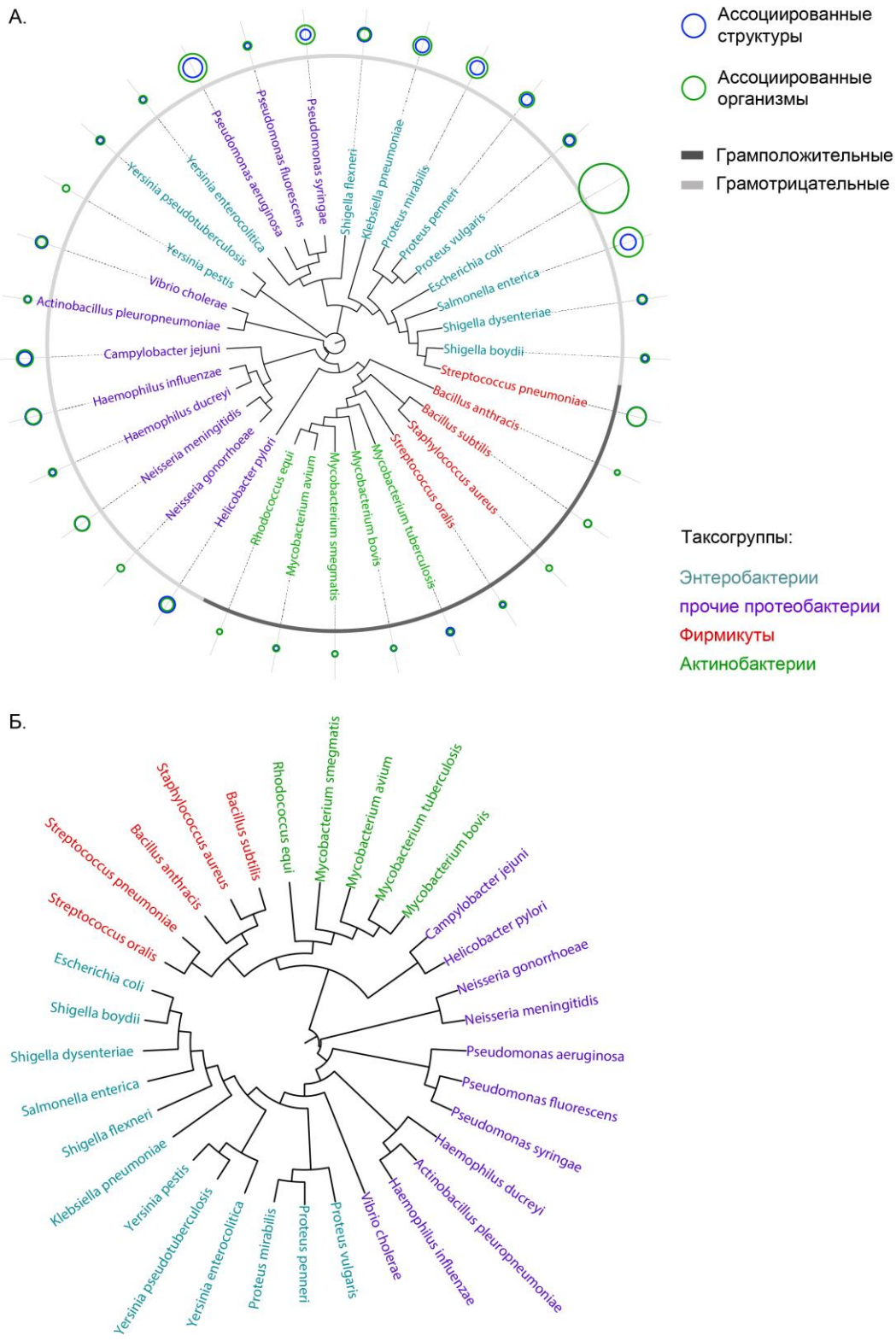



Рис. 50. Круговые фенетические деревья наиболее изученных бактериальных видов. Фирмикуты показаны красным, актинобактерии – зелёным, энтеробактерии – голубым, прочие протеобактерии – фиолетовым. Использован алгоритм кластеризации Biological Neighbor Joining. А. Кладограмма распределения димеров, присутствующих в биогликанах и содержащих моносахариды, агликоны и моновалентные остатки. Б. Кладограмма, построенная на основании последовательностей рибосомальной РНК.

Близость видов на углеводном фенетическом дереве отражает сходство их жизнедеятельности и позволяет прогнозировать дальнейшие направления экспериментальных исследований в существующих условиях недостаточной изученности механизмов бактериального патогенеза на молекулярном уровне.

### *3.5. Взаимодействие с другими проектами*

Развитая информационная среда характеризуется возможностью применять поисковые стратегии и алгоритмы получения нового знания, реализованные в одних проектах, к фактическим данным из других проектов и комбинировать инструменты, представленные разными разработчиками. Так как ни один проект не содержит всех возможных данных и инструментов, подобная интеграция чрезвычайно важна для ориентации в огромном массиве накопленных гликохимических данных. Интеграция в гликомике исторически отставала от таковой в геномике и протеомике, так как существующие проекты были изолированы друг от друга из-за недостаточной стандартизации форматов и протоколов [40].

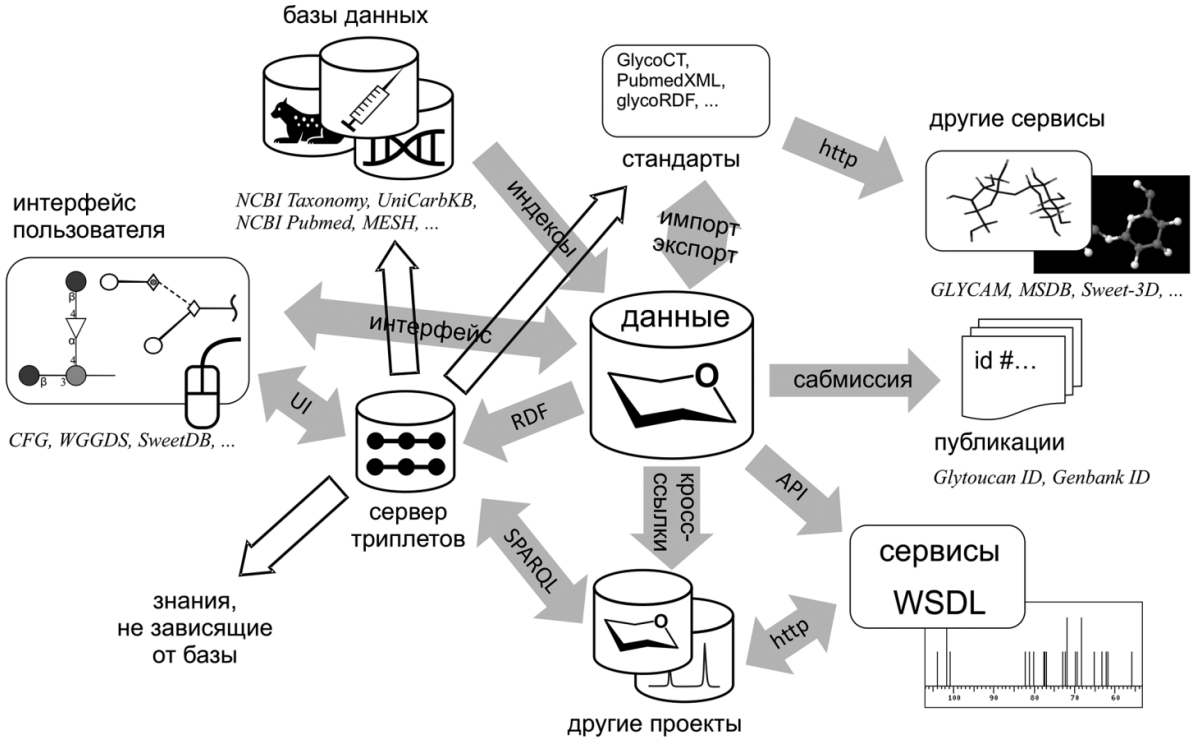



Рис. 51. Интеграция CSDB с другими проектами. Названия отдельных проектов показаны курсивом. Серые стрелки обозначают протоколы обмена данными.

Первая попытка связать углеводные проекты, обеспечив использование двух баз данных в инструментах поиска и прогнозирования, реализованных в каждой из них, была предпринята с участием автора более 10 лет назад [255]. В рамках этого модельного исследования были разработаны основные правила, позволяющие автоматическое и прозрачное для пользователя взаимодействие между программами гликохимии и гликобиологии. В дальнейшем они были усовершенствованы, дополнены моделью Resource Description Framework (см. ниже)

и использованы во многих других проектах. В той мере, в которой эти наработки реализованы в CSDB, они проиллюстрированы на Рис. 51. и включают:

1. Использование перманентных идентификаторов записей, не меняющихся при обновлении баз данных (в CSDB эту роль выполняет CSDB ID).
2. Возможность надёжной идентификации сахарида или гликоконъюгата независимо от базы, в которой он находится. Для этой цели предполагается использовать углеводный репозиторий GlyTouCan, разработанный коллабораторами в 2015 году [69, 300]. Ссылки на идентификаторы GlyTouCan автоматически генерируются при любом выводе структур в CSDB, когда это возможно (приблизительно для 50% структур в CSDB). Вопрос о включении этих идентификаторов в публикации в обязательном порядке (по аналогии с идентификаторами GenBank [34] в геномике) требует поддержки со стороны научных издательств и с участием автора лоббируется Консорциумом по Гликоинформатике в настоящее время.
3. Возможность перевода информации о структуре между основными углеводными языками (нотациями). Как минимум, интегрируемые проекты должны уметь транслировать структуры на один и тот же язык и с одного и того же языка. CSDB позволяет импортировать структуры в нотации GlycoCT и экспортировать структурную и сопутствующую информацию в нотациях GlycoCT, Glyde-II, LinUCS, SNFG, Sweet-DB (extended IUPAC), SMILES, WURCS, GLYCAM, DCI XML, PubMed XML. Каждая из них предназначена для взаимодействия со своим классом веб-сервисов, баз данных или программного обеспечения.
4. Использование стандартных индексов в существующих проектах для той информации, которая присутствует в других базах. CSDB имеет ссылки на записи в GlycomeDB (структуры [72]), Glytoucan (структуры [48]), MSDB (моносахариды [101]), NCBI Pubmed (публикации [36]), DOI (публикации [301]), NCBI Taxonomy (организмы [260]). В ближайшем будущем планируется аналогичная интеграция с международ-

ным каталогом заболеваний ICD-11 [37, 302] (заболевания, ткани и органы) и базой терминов MeSH [303] (методы анализа биогликанов).

5. Стандартные протоколы обмена данными, как в контексте логики обработки данных (SOAP, REST) [304], так и в контексте низкоуровневых стандартов (http, XML, и т.д.).
6. Стандартизированные, сертифицированные, опубликованные и признанные научным сообществом средства ввода и редактирования углеводных структур и обеспечение их совместимости с логикой каждого из использующих их проектов. В CSDB для этой цели адаптированы Java-апплет GlycanBuilder (разработка объединённого коллектива из нескольких европейских институтов [138]), браузерное приложение SugarSketcher (разработка коллабораторов из Швейцарского института биоинформатики [305][a]), браузерное приложение Structure Wizard (собственная разработка) и библиотека распространённых структурных фрагментов.
7. Наличие точек входа, работающих по стандартным протоколам информатики и их углеводным расширениям в каждом из проектов (автоматический программный интерфейс, API) и их формальное описание на языке WSDL[b] для того, чтобы другие проекты могли посылать запросы и получать ответы в автоматическом режиме, без необходимости подстраиваться под формат данных CSDB.

Интеграция проектов, связанных с химией и биологией углеводов в единое информационное пространство курируется международным консорциумом по гликоинформатике (GLIC), членом которого является и автор диссертации.

Новейшим направлением обработки данных в естественных науках и получения неявно заданных знаний является так называемая семантическая паутина [306, 307], представляющая данные в модели Resource Description Framework (RDF) [43] в виде триплетов *объект-предикат-субъект*. Такие данные позволяют интегрировать знания из разных проектов автоматически и допускают рас-

[a] https://github.com/alodavide/sugarSketcher

[b] https://www.w3.org/TR/wsdl/

пределённые запросы к базам данных с минимальным знанием форматов и интерфейсов каждой из них. Адаптация модели RDF к науке об углеводах была начата в сотрудничестве с другими группами относительно недавно [300, 308], но уже достигнуты значительные результаты.

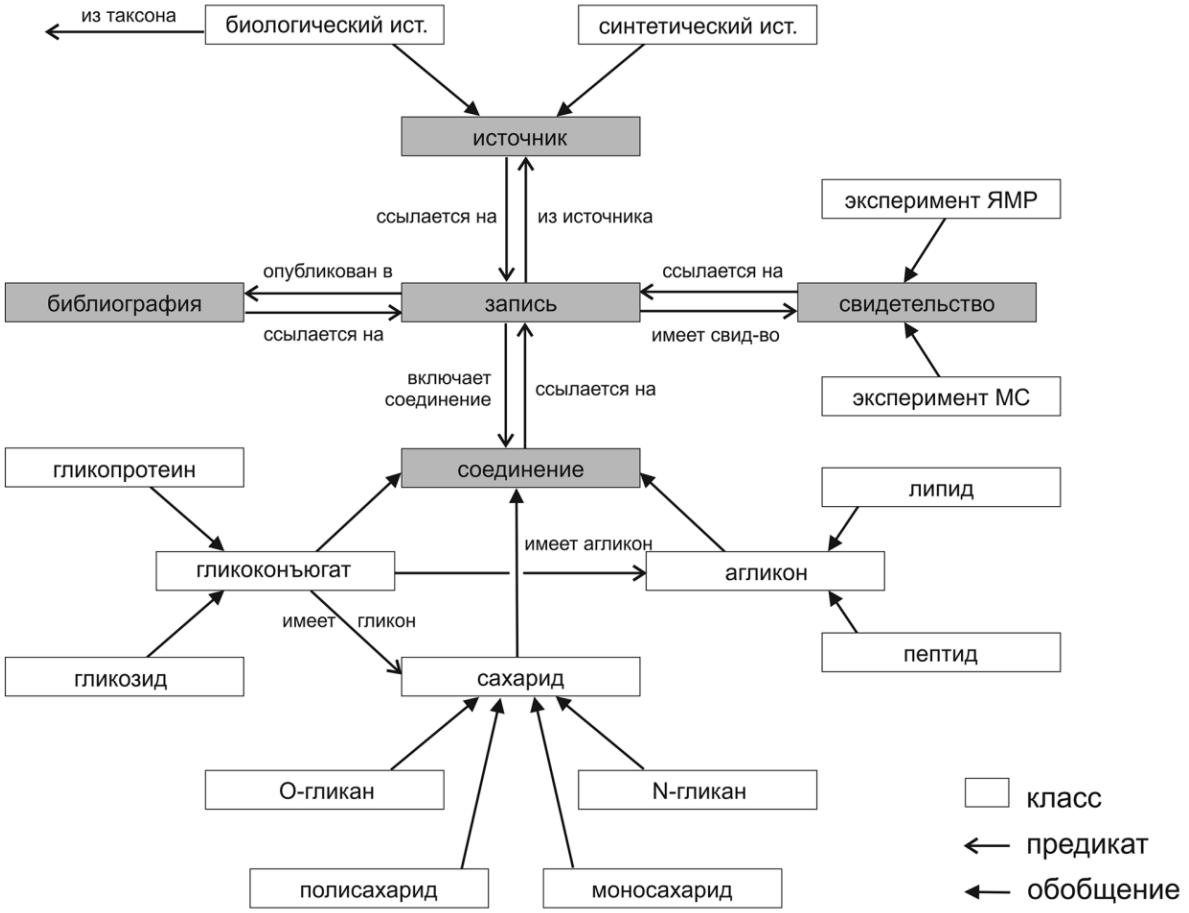



Рис. 52. Диаграмма ядра онтологии GlycoRDF. Названия классов и предикатов соответствуют синтаксису онтологии с точностью до перевода на русский язык. Пять основных классов обозначены серым. Показаны только крупнейшие подклассы и отдельные предикаты. Стрелка (предикат) соединяет субъект с объектом в RDF-триплете. Треугольная стрелка обозначает отношение «является подклассом».

Наилучшим образом мощь этого подхода можно продемонстрировать следующим модельным примером. Предположим, необходимо найти белок-носитель для произвольного гликана из CSDB. Это нельзя сделать напрямую, так как большая часть записей в CSDB не связана ссылками с белковыми базами. В то же время записи в CSDB имеют ссылки на идентификаторы в GlyTouCan. Как GlyTouCan, так и ещё одна база, UniCarbKB, могут экспортировать структуры в формате GlycoCT. Наконец, записи в UniCarbKB имеют ссылки на записи в бел-

ковой базе Uniprot. Эти факты позволяют предположить, что возможно сопоставить идентификаторы CSDB и UniCarbKB, используя GlyTouCan, и получить идентификаторы Uniprot из UniCarbKB для каждого идентификатора CSDB. Конечно, в каждом случае возможно неуниверсальное решение, реализованное вручную для каждого конкретного объекта исследования. В универсальном контексте эта задача решается с помощью простейшей программы на языке SPARQL, которая может быть сформирована автоматически на основании запросов пользователей в интерфейсе любой из баз [309]. Других способов эффективно решать подобные задачи, особенно для большого числа объектов в скрининговых исследованиях, в настоящее время не существует.

Для обеспечения возможности реализации таких распределённых запросов каждый из участников должен представить свои данные в модели RDF с использованием одной и той же формальной онтологии. Онтология знаний об углеводах была впервые разработана в 2015 году в сотрудничестве с американской и японской группами, которые как и российская группа, впоследствии вошли в Консорциум по Гликоинформатике. Она включает 246 объектов в 130 классах и стандартизирует 367 взаимосвязей между ними. Наиболее крупные классы и ключевые взаимосвязи показаны на Рис. 54. Примеры использования и особенности онтологии GlycoRDF приведены в публикации [44]. Полное описание онтологии на языке OWL, сопутствующая документация и ссылка на интерактивную веб-визуализацию онтологии доступны на странице GlycoRDF в репозитории кода.[a]

Процесс перевода проприетарных форматов и реляционных баз в модель RDF в соответствии с выбранной онтологией называется “RDF-изация”. Следуя договорённостям, достигнутым в ходе конференций Biohackathon в 2012 и 2013 гг.[b], ведущие мировые проекты гликоинформатики (CSDB, Glycosciences.de [77], MSDB[102], GlycomeDB – ныне GlyTouCan [69], UniCarbKB [56], GlycoEpitope [93], GlycoNAVI [55], GlycoProtDB [97], а впоследствии и другие) RDF-изировали свои базы и предоставили данные для размещения в репозитории

[a] https://github.com/ReneRanzinger/GlycoRDF/wiki

[b] http://2012.biohackathon.org/ ; http://2013.biohackathon.org/

триплетов [45]. Репозиторий обновляется раз в два года на основании RDF-данных баз-участников и позволяет делать запросы на языке SPARQL. Технически репозиторий располагается на сервере Университета Сока, Япония и свободно доступен через Интернет любому проекту.

### *3.6. Техническая реализация проекта (экспериментальная часть)*

База данных CSDB представляет собой реляционную базу данных, управляемую СУБД MySQL 5.6.[a]

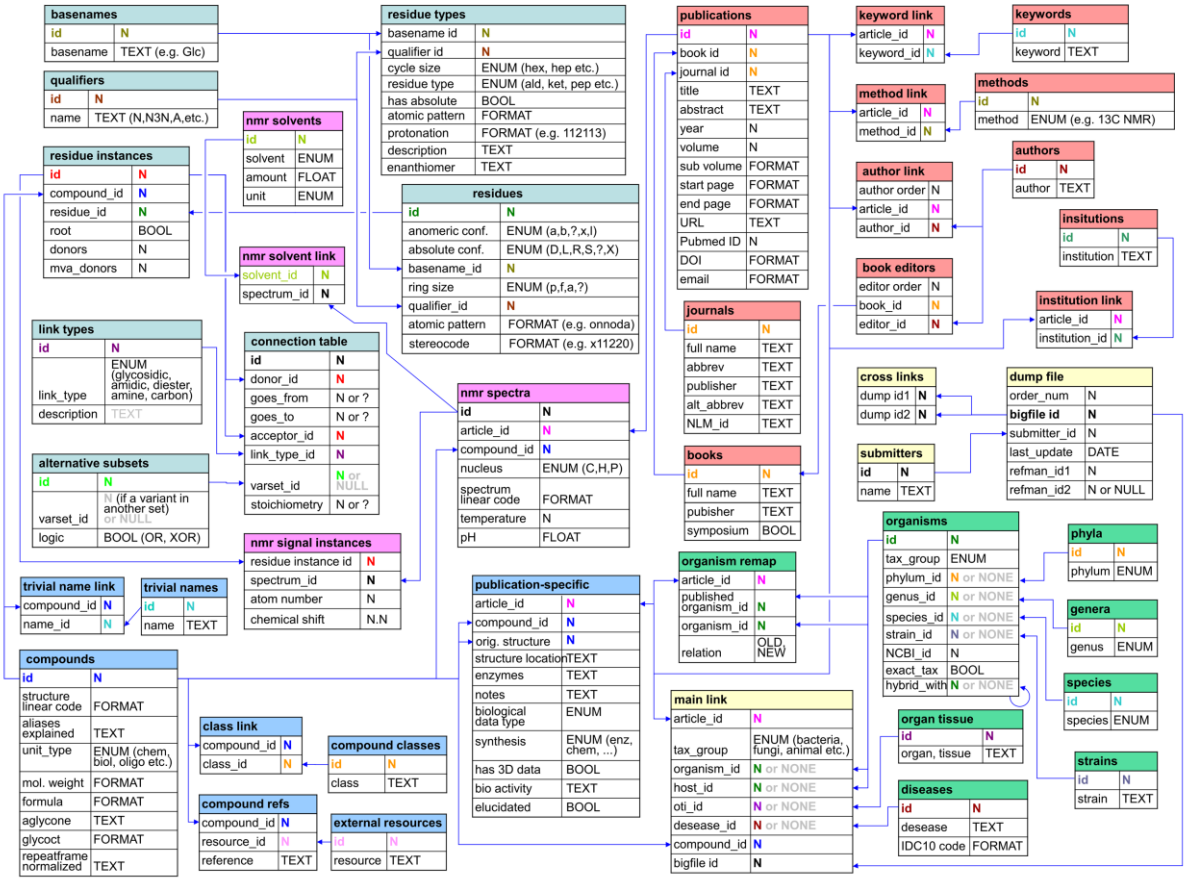



Рис. 53. Диаграмма отношений CSDB. Названия таблиц и полей приведены на английском языке для сохранения согласованности с базой данных. Типы данных: N – целое число, TEXT – произвольный текст, FORMAT – форматированный текст, ENUM – набор, BOOL – логический переключатель. Числовые индексы (N), имеющие одинаковый смысл, приведены одинаковым цветом. Более подробно с диаграммой можно ознакомиться на сайте проекта[b].

Взаимоотношения между данными, пришедшими из научных публикаций, и их индексами представлены на Рис. 53. Фон заголовков отражает группу данных, содержащихся в таблицах: ____ - биогликаны на уровне свойств молекулы, ____ - свойства моносахаридов и полная первичная структура, ____ - библио-

[a] http://www.mysql.com/

[b] http://csdb.glycoscience.ru/help/csdb_entities.pdf

графическая информация, ____ - биологическая привязка, ____ - данные ЯМР, ____ - взаимоотношения между данными из разных групп. ЯМР-спектроскопические и структурные данные, полученные статистической обработкой, предсказанием, конвертаций форматов и теоретическим прогнозированием собраны в отдельные таблицы, не представленные на Рис. 53. Модуль работы с гликозилтрансферазами реализован в виде отдельной базы данных, интегрированной с CSDB на уровне индексов (Рис. 54) и данных о молекулярной структуре.

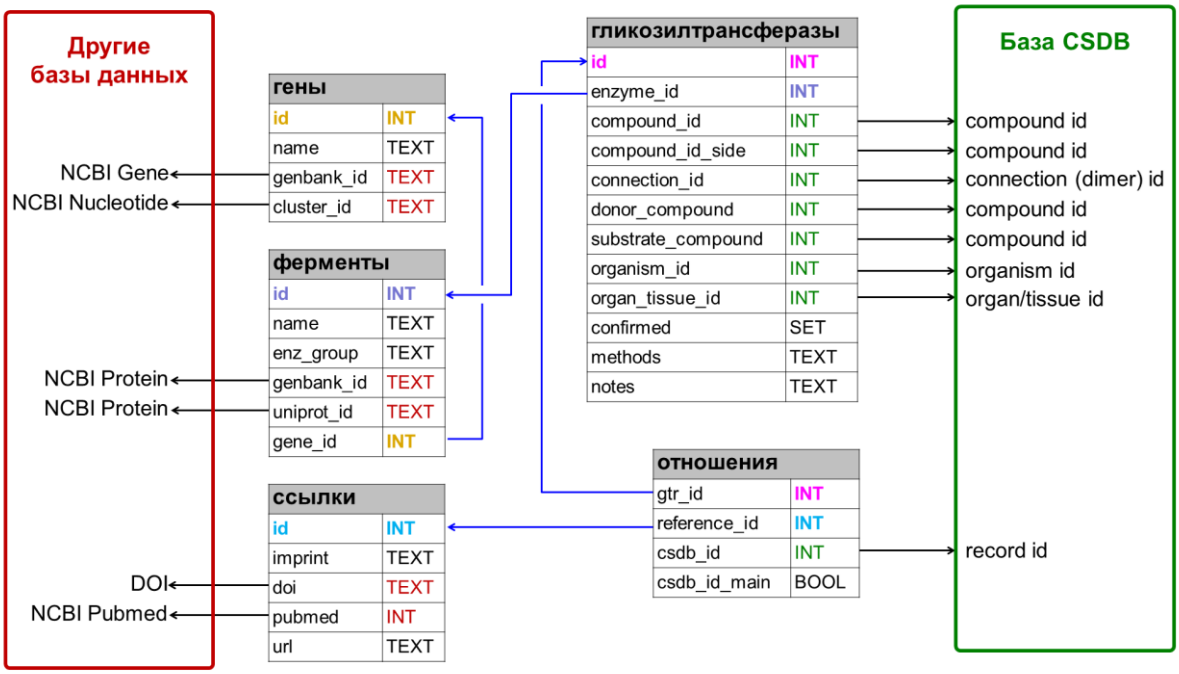



Рис. 54. Диаграмма отношений базы гликозилтрансфераз. Названия полей приведены на английском языке для сохранения согласованности с базой данных. Типы данных: INT – целое число, TEXT – текст, SET – множество, BOOL – логический переключатель.

Все данные, вносимые аннотаторами, хранятся и редактируются в текстовых дампах, которые одновременно служат резервными копиями. Формат дампа опубликован на сайте проекта[a]. Архитектура базы данных подробно описана в публикации автора [64].

Управляющие программы платформы CSDB написаны на языках PHP 5, SQL 5, R 3 и Python 3. Новые функции появляются несколько раз в год, все об-

[a] http://csdb.glycoscience.ru/database/index.html?help=dbdocs#dump

новления проходят тщательное тестирование. Взаимодействие с пользователем реализовано в виде веб-интерфейса с помощью статических и генерируемых на сервере страниц на DHTML, Javascript и JQuery. Интерфейс протестирован в современных версиях браузеров Mozilla Firefox и Google Chrome. Для отрисовки двумерных структурных формул используется пакет OpenBabel [310] для Python, для интерактивной визуализации трёхмерных моделей – апплет JSMol для веб-браузеров (упрощенная версия пакета химического моделирования Jmol [311]). Графический редактор углеводных структур в формате SNFG получен адаптацией программы SugarSketcher [305], предоставленной Швейцарским Институтом Биоинформатики.

В 2012-2018 гг. проект работал на арендуемом виртуальном Windows-сервере (VDS). Начиная с 2019-го года, как для расчетов, так и для обслуживания пользователей используется собственный физический сервер[a] под управлением Microsoft Windows Server 2016 и Internet Information Services 10.

Балансировка нагрузки при ресурсоёмких расчётах построена так, что наиболее требовательные к ресурсам задачи динамически получают наименьший приоритет (так как их результатов все равно приходится ждать). Результаты длительных вычислений предоставляются пользователю в виде веб-ссылок и email-уведомлений о завершении счёта.

Все функции платформы и базы данных, а также справочная система и техническая документация бесплатно доступны пользователям Интернета по адресу http://csdb.glycoscience.ru. В соответствии со сложившейся ситуацией в мировой науке пользовательский интерфейс, все материалы и документация на веб-сайте проекта приведены на английском языке для обеспечения возможности их использования учёными из любых стран и совместимости с оригинальными источниками данных (аннотированными публикациями).

Молекулярно-динамические расчёты моносахаридов проводились с использованием пакета TINKER [312], силового поля MM3-2001 [313] и 1-наносекундных траекторий при 1000 K. Для оптимизации собранных структур в потоковом режиме использовалась реализация молекулярной механики в сило-

[a] 2 CPU Intel Xeon 6144 (16 ядер), 3.5 ГГц, 96 Гб RAM, M.2 SSD RAID1, SSHD RAID1

вом поле MMFF94 [264] в библиотеке RDKit[a] для Python. Квантово-механическое расчёты проводились в программе Gaussian 09 [314] методами DFT GIAO на уровне теории B3LYP в базисе 6-311G++(2p,2p) с использованием модели растворителя COSMO [186] и в программе Priroda [217] в базисе PBE.

Кластеризация таксонов проводилась с помощью скриптов на языке R 3 [315]с использованием библиотеки Ape («анализ филогенетики и эволюции») [316] и известных алгоритмов UPGMA [317], Ward.D2 [318], BIONJ [319], fastME [320] и «полная связность». Фенетические деревья сравнивали методом Ние и соавторов [321] с помощью веб-сервиса Compare2Trees[b]. Для построения сложных дендрограмм использовался пакет iTOL [322] и данные, экспортированные из CSDB в формате Nexus [296].

Кроме литературных данных ЯМР, модельные объекты исследовались коллабораторами методами химии углеводов и ЯМР-спектроскопически автором диссертации. Спектры ЯМР снимались на приборах с рабочей частотой 300-600 МГц в растворах в $D_2O$ после двукратной лиофилизации и растворения. Температура образца выбиралась, исходя из положения остаточного сигнала HDO и составляла 25-65°C. Химические сдвиги калибровались по внутреннему стандарту $(CH_3)_2CO$ ($\delta_H$ 2.225, $\delta_C$ 31.45). Время смешивания в экспериментах со спин-локом и в экспериментах по наблюдению ЯОЭ выбиралось, исходя из молекулярного веса и предыдущих исследований и составляло от 60 до 200 мс. Спектры обрабатывались в программах Bruker TopSpin и MestreLabs Nova.

[a] http://www.rdkit.org/

[b] http://www.mas.ncl.ac.uk/~ntmwn/compare2trees/index.html

## 4. Использование разработок в гликохимии и гликобиологии

(обсуждение результатов в контексте научной области)

Научная значимость разработанной платформы CSDB, включающей собственно базу данных углеводных структур и различные надстройки, подтверждается её применением в различных исследованиях. В данном разделе перечислены типовые задачи, решаемые с помощью CSDB, и примеры её использования в гликохимических исследованиях. Сводные данные по использованию различных аспектов CSDB в гликохимии и гликобиологии приведены в Табл. 18. Ссылки даны только на обзоры, главы и статьи общего характера, а данные по конкретным исследованиям представлены кумулятивно.

Табл. 18. Использование наработок CSDB в других исследованиях.

| *Что использовано* | *Для чего использовано* | *Число цитирований и избранные ссылки* * |
|---|---|---|
| CSDB как проект (включая BCSDB и PFCSDB) | Интеграция с другими проектами гликоинформатики и с другими областями наук о жизни | 11 [323] |
| | Обзоры углеводных баз данных и инструментов гликоинформатики | 34 [33, 44, 51, 251, 324-333] |
| | Обзоры баз данных, удовлетворяющих определённым критериям (поддержка стандартов или технологий, использование конкретных типов данных и т.д.) | 3 [334, 335] |
| | Инициативы по автоматизации и «информатизации» анализа углеводов; создание углеводной семантической паутины | 6 [336, 337] |
| | Обзоры баз, применяемых в медицине, фармацевтике и микробиологии | 3 [253, 338] |
| | Исследование характеристик конкретных групп сахаридов и/или их разнообразия | 3 |
| | Дальнейшие разработки на основе CSDB | 18 |
| | Включение CSDB в NAR molecular biology database collection | 1 [250] |

| | | |
|---|---|---|
| Модель данных CSDB | Конкретные проекты (базы данных, онтологии, разработка лекарств) | 4 |
| | Обзоры информационных методов в гликологии и инструментов интерпретации экспериментальных данных | 5 [339-342] |
| Конкретные данные, извлечённые из CSDB | Исследование структуры конкретных бактериальных полисахаридов | 3 |
| | Обеспечение данными и валидация других компьютерных проектов | 12 |
| | Поиск общих мотивов в полисахаридах бактерий и выявление биологических повторяющихся звеньев | 2 |
| | Статистический анализ углеводных структур; анализ встречаемости и распределения структурных характеристик по таксонам | 5 [343] |
| | Моделирование геометрии углеводов и конформационный анализ | 1 |
| | Выявление ошибок в публикациях | 1 |
| | Поиск эпитопов; анализ разнообразия антиуглеводных антител; углеводные маркеры | 4 |
| | Идентификация, классификация и изучение работы ферментов в конкретных таксонах | 3 |
| | Прочее, в том числе введение в структуры углеводов из конкретных таксогрупп | 3 |
| Кумулятивные данные, полученные с помощью статистических инструментов CSDB | Построение углеводных фенетических деревьев и дальнейшие статистические исследования | 2 |
| | Обзоры разнообразия углеводов, в том числе в конкретных таксогруппах или в контексте биосинтеза гликанов | 12 [136, 331, 343] |
| | Обзоры разнообразия структурных особенностей углеводов в контексте их синтеза | 6 [344, 345] |
| | Обзоры разнообразия структурных особенностей углеводов в контексте их поддержки инструментами гликомики | 6 [51, 330, 346, 347] |

| | | |
|---|---|---|
| | Изучение эволюции и работы ферментов; гликозилирование белков | 4 [348] |
| | Объяснение кросс-реакций и связывания с бактериальными гликоэпитопами; распознавание конкретных антигенов и изучение иммунного ответа | 6 |
| | Использование данных о распределении конкретных структурных особенностей для изучения биогликанов определённых таксонов | 7 [349] |
| | Прочее | 5 |
| Статистические инструменты CSDB | Разработка других инструментов гликоинформатики | 4 [339] |
| | Методология интерпретации экспериментальных данных | 2 [350, 351] |
| | Исследования структуры и свойств конкретных сахаридов | 5 [352, 353] |
| | Исследования взаимодействия углеводов с белками | 5 [354] |
| | Прочее | 2 |
| Предсказание спектров $^{13}C$ углеводов и структур по спектрам; данные ЯМР, их компьютерный анализ и визуализация | Обзоры методов эмпирического ЯМР-моделирования углеводов и надстроек CSDB | 25 [166, 355-357] |
| | Обзоры методов квантовомеханического ЯМР-моделирования углеводов | 57 |
| | Дальнейшие разработки автора диссертации | 5 |
| | Дальнейшие разработки других авторов | 8 [286] |
| | ЯМР-мониторинг активности ферментов | 2 |
| | Выявление закономерностей структура-спектр в полисахаридах, отнесение сигналов конкретных биогликанов | 2 [351] |
| | Установление или характеризация структуры конкретных олиго- и полисахаридов | 6 |
| | Конформационный анализ и дизайн синтетических углеводов | 4 [342, 358] |
| | Автоматизация предсказания структуры углеводов | 2 |

| | | |
|---|---|---|
| | Прочее | 3 |
| Язык CSDB Linear | Поддержка CSDB Linear в конвертерах, базах данных, инструментах гликомики | 4 |
| | Обзоры углеводных языков (нотаций) | 8 [51, 135, 343, 346, 347, 359] |
| | Разработка собственных углеводных нотаций и других инструментов | 5 [145, 360] |
| Методы и результаты интеграции углеводных баз данных | Обзоры сложностей интеграции и/или существующих инициатив | 12 [33, 51, 308, 330, 331, 359] |
| | Дальнейшие разработки других компьютерных инструментов, стандартов, протоколов | 8 |
| Создание семантической паутины в химии и биологии | Интеграция баз данных и других проектов гликомики | 10 |
| | Теоретические модельные исследования, создание онтологий, алгоритмов | 11 [145, 361] |
| | Обзоры перспективных методов работы с информацией об углеводах | 13 [324, 326, 328, 335, 362-364] |
| | RDF-изация других баз данных | 10 |
| | Разработка других инструментов гликоинформатики | 8 |
| | Прочее | 2 |
| Онтология GlycoRDF | Разработка, RDF-изация и интеграция других углеводных баз | 9 |
| | Обзоры способов интеграции биохимических баз; углеводная семантика | 6 [20, 326, 335, 362] |
| | Теоретические работы и другие онтологии | 3 |
| Методология ви- | Поддержка SNFG в других проектах[a] | >150 |

[a] Веб-ссылки на указанные проекты приведены в специальном разделе сайта NCBI: https://www.ncbi.nlm.nih.gov/books/NBK310273/

| | | |
|---|---|---|
| зуализации углеводных структур (SNFG) | базы данных: UnicarbDB, UnicarbKB, MonosaccharideDB, KEGG Pathways, JCGGDB, ACGG-DB, GlyTouCan, CFG, Glycam-Web, GLYCOSCIENCES.de, GlycoStore, Glyco3D, GlycoPedia, Glycomics@ExPASy, SugarBindDB<br>программы: 3D-SNFG, Draw-Glycan, GlycanBuilder 2<br>инициативы по стандартизации: IUPAC, MIRAGE<br>журналы: *Glycobiology, Glycoconjugate Journal, Journal of Biological Chemistry, Journal of Cell Biology, Molecular and Cellular Proteomics, Carbohydrate Research* | |
| Выявленные недостатки других проектов | Проекты, связанные с курированием | 10<br>[52] |
| | Доработка инструментов анализа экспериментальных данных | 2<br>[350, 365] |

* Данные приведены по состоянию на середину 2018 г.

### *4.1 Примеры решения модельных задач*

Этот раздел содержит иллюстрированное пошаговое руководство использования базы CSDB на 12 примерах, моделирующих задачи, с которыми исследователь углеводов сталкивается в своей ежедневной научной практике. Задачи не претендуют на получение нового знания; они специально сформулированы в качестве обучающих моделей применения базы данных CSDB в реальных исследованиях. В этих примерах, количество найденных записей и их временные идентификаторы (все кроме перманентного идентификатора CSDB ID) могут отличаться от реальной ситуации на сайте проекта, так как база непрерывно пополняется новыми данными. Все поисковые примеры объясняют сложные составные запросы, так как простые одноступенчатые запросы интуитивно понятны и не требуют специальных разъяснений. В тексте **сжатым жирным шрифтом** выделены элементы интерфейса, а `моноширинным шрифтом` – вводимые пользователем данные. Раздел 4.1 можно рассматривать как учебник по использованию CSDB. Он подразумевает последовательное прочтение – первые примеры рассмотрены наиболее подробно, а последующие опираются на материал предыдущих. Примеры 1-6 связаны с функциями поиска данных по нескольких критериям, примеры 7-9– с исследованием связи «структура-спектр ЯМР»; примеры 10-12 – со статистическим поиском закономерностей «структура-таксономия»:

1. Изучить, как введение аминогруппы влияет на химические сдвиги в лактозном фрагменте.
2. Найти бактериальные углеводы, содержащие галактуроновую кислоту и ещё как минимум одну гексозу, структура которых опубликована после 2005 года в связи с антигенной активностью.
3. Найти соединения, полученные из растений рода *Паслён* и содержащие соланидиновый фрагмент.
4. Найти углеводы, кроме октозосодержащих, имеющие в спектре ЯМР $^{13}C$ характеристичный сигнал вблизи 34 м.д.
5. Найти публикации *Книреля* или *Шашкова (АС)* по бактериальным гликанам, включающим хиновоз-4-амин, амидированный любой N-ацетилированной аминокислотой.

6. Найти бактериальные структуры, построенные из любых ноноз одного типа (моносахариды или их гомополимеры).
7. Промоделировать спектры ЯМР 3-O-абеквозил-6-дезокси-β-D-манногептопиранозил-(D-рибит-1)-фосфата в водном растворе и оценить точность предсказания наименее достоверных сигналов.
8. Простым способом установить характер связывания и топологию полимерного фукоглюкана с дисахаридным повторяющимся звеном на основании одномерного спектра ЯМР $^{13}C$.
9. Предсказать структуру неустановленного олигомера, содержащего остатки бациллозамина, лизина и глюкуроновой кислоты, на основании данных ЯМР.
10. Изучить состав гликанов аспергиллов *Aspergillus oryzae* и *Aspergillus fumigatus* с особым вниманием к эпитопам, располагающимся на концах боковых цепей.
11. Установить, какие димерные фрагменты (включая сахара и агликоны) гликанов высших растений уникальны для рода *Lupinus*.
12. Получить статистические данные об изученности гликома протеобактерий.

Более подробно с вышеуказанными примерами можно ознакомиться на сайте проекта[a] и в монографии автора [366].

[a] http://csdb.glycoscience.ru/database/core/help.php?topic=examples

*4.1.1. Изучение влияния введения аминогруппы на химические сдвиги в лактозном фрагменте.*

Для решения этой задачи найдём структуры, содержащие дисахарид 4-O-β-D-галактопиранозил-2-амино-β-D-глюкопиранозу (как со свободной, так и с ацетилированной аминогруппой) ИЛИ 4-O-β-D-галактопиранозил-β-D-глюкопиранозу, для которых в CSDB присутствуют данные ЯМР.

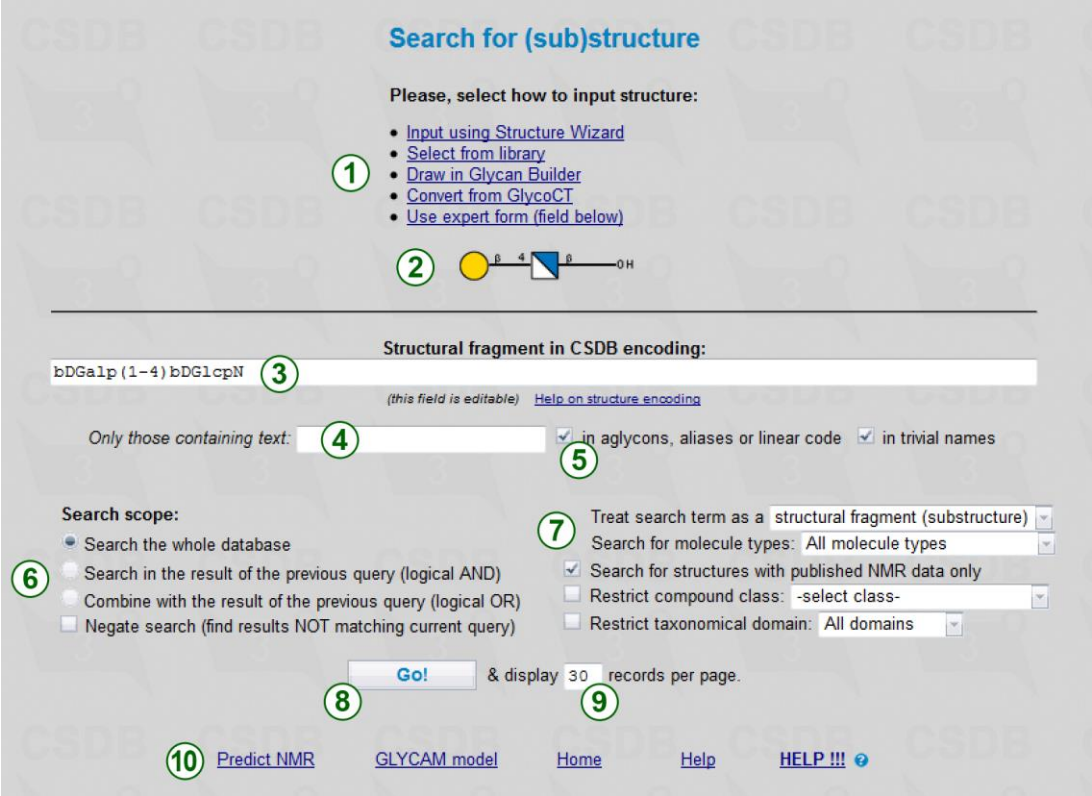



Рис. 55. Запрос на поиск по фрагменту структуры. В форму введён дисахаридный фрагмент (2,3): 4-O-β-D-галактопиранозил-2-амино-β-D-глюкопираноза.

Для проведения поиска выберем его тип в главном меню CSDB. В данном случае это поиск по фрагменту структуры - **(Sub)structure search**, его форма показана на Рис. 55. Мы помещаем одним из доступных способов поисковый запрос в структурное поле (3)[a] и/или свободнотекстовое поле (4), выбираем область поиска (6), устанавливаем дополнительные ограничения (7) и запускаем поиск (8). Введённый структурный фрагмент визуализируется в графической форме (2) в

[a] Здесь и далее числа в скобках указывают на элемент интерфейса, обозначенный этим числом в круге на рисунке, на который была последняя ссылка в предшествующем тексте. При смене рисунка, к которому относятся числа в скобках, ссылка на рисунок даётся заново.

формате SNFG (расширение нотации CFG, активно используемой гликобиологами). Если поисковый запрос редактировался вручную и содержит ошибки, из-за которых он не может быть проинтерпретирован, в области (2) показываются сообщения об ошибках.

Каждая поисковая форма (кроме поиска по идентификаторам, **ID search**) имеет селектор области поиска (6), позволяющий уточнять предыдущий поиск путём пересечения или объединения с новым запросом того же или другого типа. В частности, **Search in the result of the previous query** пересекает текущий запрос с предыдущим (логическое И), а **Combine with the result of the previous query** объединяет текущий запрос с предыдущим (логическое ИЛИ). Опция **Negate search** отрицает текущий запрос (логическое НЕ) и может быть использована как для поиска всех данных, кроме удовлетворяющих запросу, так и для исключения результатов текущего запроса из результатов предыдущего (логическое И НЕ, в сочетании с выбором **Search in the result of the previous query**).

Количество найденных записей, которые будут отображаться на одной веб странице, можно указать в поле (9). Форма имеет ссылки (10) на прочие структурные инструменты CSDB и справочную систему, имеющую как классический (**Help**), так и интерактивный режим (**HELP!!!**).

Ввести искомую структуру или её фрагмент можно несколькими способами (1):

- Мастер структур (**Structure wizard**) используется для сборки структуры с помощью визуальных операций мышью, таких как отметка опций или выбор вариантов из меню. Он не требует специальных знаний, кроме общей номенклатуры углеводов. В то же время отдельные редкие особенности структуры не могут быть созданы с помощью «мастера».
- Выбор из библиотеки (**Select from library**) позволяет выбрать распространённую структуру из библиотеки, используя для поиска её тривиальное название или устоявшуюся аббревиатуру. Структуры в библиотеке каталогизированы по классам и имеют возможность предпросмотра в графическом (SNFG) или псевдографическом (SweetDB) формате.
- Графический редактор (**Draw in Glycan Builder / Draw in Sugar Sketcher**), адаптированный из разработки коллабораторов Швейцарского Института Био-

информатики, позволяет нарисовать большинство распространённых структур непосредственно в формате SNFG.

- Транслятор GlycoCT (**Convert from GlycoCT**) позволяет перевести структуру из распространённой углеводной нотации GlycoCT condensed в родной формат CSDB Linear.
- Копирование предыдущего запроса (**Copy from the previous structural query**) с последующим редактированием доступно, если в пределах пользовательской сессии уже был другой структурный запрос. Этот способ полезен, если необходимо повторить запрос или внести лишь небольшие изменения.
- Прямой ввод структуры вручную (**Use expert form**) на языке CSDB Linear позволяет использовать полный спектр структурных особенностей, поддерживаемых базой, однако он требует знания семантики и синтаксиса нотация CSDB Linear (см. раздел 3.3.1).

Дисахарид, использованный в этом примере, не содержит специфических особенностей и может быть введён без использования экспертного режима. Для ознакомительных целей, используем здесь Glycan Builder для ввода 4-O-β-D-галактопиранозил-2-амино-β-D-глюкопиранозы. Для этого проследуйте по ссылке **Draw in Glycan Builder** (1) в окно редактора (Рис. 56).

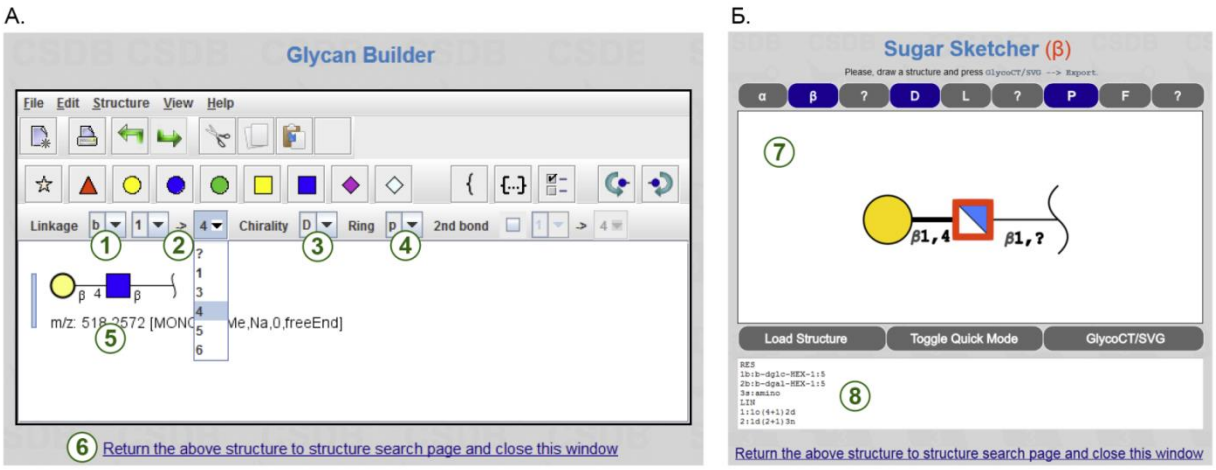



Рис. 56. Графические редакторы углеводных структур: А. Glycan Builder (2011), Б. SugarSketcher (2018).

Это Java-приложение позволяет сконструировать гликан с помощью интуитивно понятного интерфейса. Выбираемые с помощью пиктограмм остатки

моносахаридов добавляются в структуру, начиная с восстанавливающего конца. Для создания точки разветвления требуется кликнуть на произвольный остаток, и далее цепь будет расти от него. Элементы интерфейса позволяют выбрать аномерную (1; α, β или неизвестная) и абсолютную (3; D, L или неизвестная) конфигурацию текущего остатка, его способ циклизации (4; пираноза, фураноза, линейный или неизвестный) и способ связывания с акцептором (2; номера углеродных атомов в самом остатке и в его акцепторе). Результирующая структура (5) отображается в одном из трёх форматов (переключение в меню **View**): формат CFG Консорциума по Функциональной Гликомике (показан на рисунке); Оксфордский формат OUXF; текстовые обозначения моносахаридов. В настоящее время идёт замена приложения GlycanBuilder (Рис. 56А) на более современный SugarSketcher (Рис. 56Б), визуализирующий структуру в форматах SNFG (7) и GlycoCT (8) и не требующий от пользователя установленного Java-окружения. Несмотря на различия в интерфейсе, эти два приложения функционально эквивалентны. Нажатие на ссылку (6) передаёт структуру в форму **(Sub)structure search** (Рис. 55), из которой был вызван редактор, и закрывает окно.

2-амино-β-D-глюкопираноза редко встречается в неацетилированном виде, поэтому её нет в меню остатков GlycanBuilder, и мы использовали вместо него распространённую N-ацетилированную форму. После перехода обратно к форме структурного поиска, поисковый запрос содержит введённый фрагмент в формате CSDB Linear (3 на Рис. 55.): `bDGalp(1-4)[Ac(1-2)]bDGlcpN`, в котором N-ацетилглюкозамин представлен как два остатка: глюкозамина и уксусной кислоты. Для получения желаемой структуры, мы должны «деацетилировать» N-ацетилглюкозамин, т.е. вручную удалить `[Ac(1-2)]` из поля (3). При любом исправлении, синтаксис записи структуры проверяется автоматически и результат проверки отображается в формате SNFG в поле (2). Показанный на Рис. 55 запрос найдёт структуры, содержащие указанный фрагмент как с ацетилированной, так и со свободной аминогруппой.

Этот пример подразумевает следующие дополнительные параметры, большинство из которых установлено по умолчанию:

- поиск структур, содержащих указанный фрагмент, но не обязательно исчерпывающихся им (7; опция **Structural fragment**)

- поиск во всей базе данных без учёта предыдущих запросов (6)
- поиск структур всех типов (7; опция **All molecule types:** олигомеры, повторяющиеся звенья, циклические полимеры и т.д.)
- отсутствие ограничений на текстовые фрагменты в агликонах или тривиальных названиях (4)
- отсутствие ограничений на класс соединения и на домен его биологического источника (7)
- поскольку в этом примере нам нужны данные ЯМР, поставим отметку **Search for structures with published NMR data only** (7).

Нажатие на кнопку **Go!** (8) запускает поиск. В результате мы получим 44 структуры, содержащие указанный фрагмент. Теперь добавим к результатам структуры, содержащие аналогичный фрагмент без аминогруппы. Для этого можно нажать ссылку **New query** в конце страницы с результатами или воспользоваться главным меню для повторного открытия формы **Search for (sub)structure.**

Поскольку 4-O-β-D-галактопиранозил-β-D-глюкопираноза – это распространённый дисахарид «лактоза», мы можем выбрать его из библиотеки структур, нажав на **Select from library** (1). Окно библиотеки показано на Рис. 57А. Здесь выберем `lactose` (2) из раздела `Named saccharides` (1), и структура будет предварительно показана в форматах SweetDB (3) и SNFG (4). Нажатие на ссылку (5) передаёт выбранную структуру в форму поиска и закрывает библиотеку.

Форма поиска с введённым лактозным фрагментом `bDGalp(1-4)bDGlcp` (7) показана на Рис. 57Б. Другим приемлемым в данном случае способом ввода является копирование предыдущего запроса с помощью ссылки **Copy from previous query** (6) и последующие удаление аминогруппы непосредственно в поле (7).

Теперь мы готовы скомбинировать результаты двух запросов: текущего поиска 4-O-β-D-галактопиранозил-β-D-глюкопиранозы и предыдущего поиска 4-O-β-D-галактопиранозил-2-амино-β-D-глюкопиранозы. Мы проводим оба поиска только среди структур с опубликованными данными ЯМР (10). Число объектов, найденных в рамках предыдущего запроса показано (9) под блоком вариантов области поиска, а нажатие на **ID list** (9) выведет их список для справки. Выбрав область поиска **Combine with the result of the previous query** (8), мы объединяем ре-

зультаты текущего и предыдущего запроса. После запуска поиска нажатием на кнопку **Go!** (11) мы получаем первую страницу списка из 109 структур (Рис. 58).

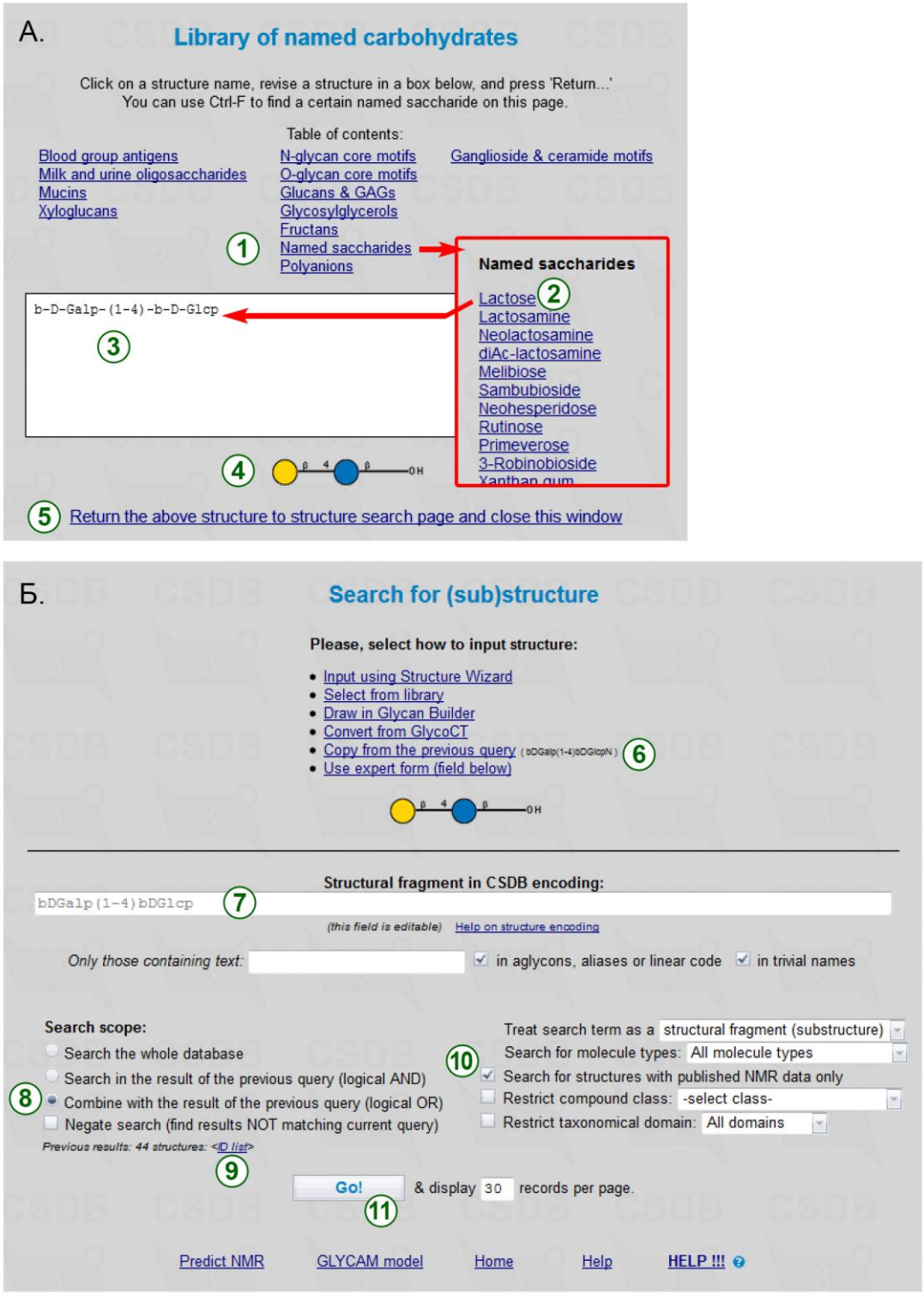



Рис. 57. А. Библиотека типовых структур. Б. Комбинирование структурного поиска (фрагмент, выбранный из библиотеки) с предыдущим запросом.

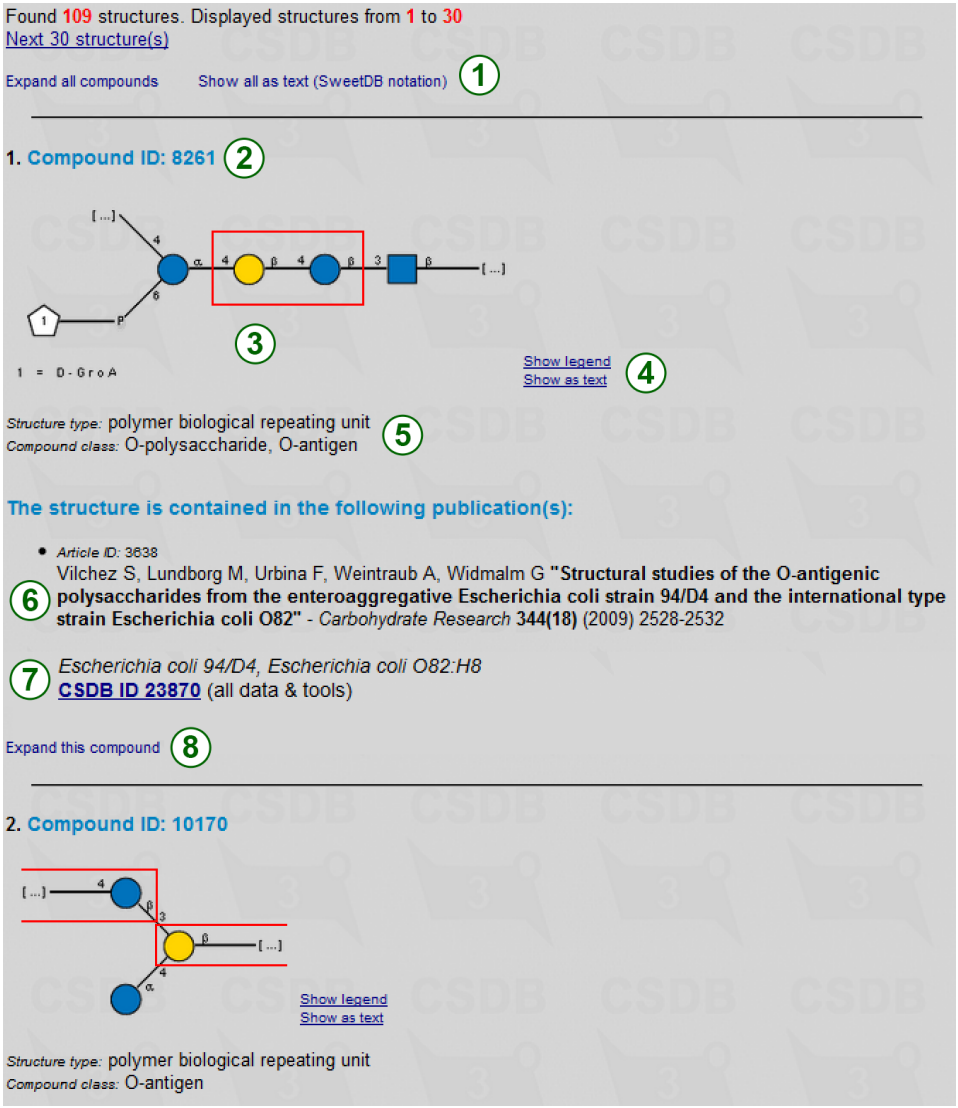



Рис. 58. Результат комбинированного структурного запроса в краткой форме (показан для первой, и частично для второй найденной структуры). Фрагменты структуры, соответствующие поисковому запросу, обведены красной рамкой.

По умолчанию, результаты показываются в сжатой форме, содержащей только важнейшие данные. Заголовок страницы (1) содержит число найденных объектов, средства навигации по страницам с результатами (**Next**... и **Previous**...), ссылку **Expand all**... для показа данных в развёрнутой форме и переключатель визуализации структур между псевдографическим SweetDB и графическим SNFG (**Show all as**...) форматами. Каждое найденное соединение представлено идентификатором (2), первичной структурой (3), пояснениями к формату визуализации и его переключателем (4), типом структурной единицы (напр., «биологическое повторяющееся звено») и аннотацией класса соединения (напр., «О-антиген»). Структуры сопровождаются списком источников, к которых они опубликованы (6); в пока-

занном примере в этом списке только одна статья. Уникальная комбинация структуры и статьи, в которой она установлена или обсуждается, в CSDB называется «запись». В отличие от прочих объектов, записи имеют перманентный идентификатор (CSDB ID), не меняющийся при обновлении базы. Его можно использовать для однозначной адресации данных в других базах и в последующих публикациях. Каждая перечисленная статья имеет ссылку (7) на соответствующую запись информацию о таксономии организмов, ассоциированных со структурой в этой записи. Нажатие на эту ссылку приводит к отображению всех данных, присутствующих в записи. Ссылка (8; **Expand this compound**) переключает режим отображения соединения с краткого на развёрнутый и обратно.

На Рис. 59 в развёрнутой форме показана запись 27282, содержащая второе из 109 найденных соединений. Эта форма отображает все (при наличии) или часть следующих данных:

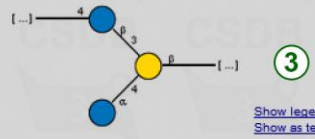
- перманентный идентификатор записи CSDB ID (1);
- авторов, название и выходные данные публикации (2);
- первичную структуру (3) и инструменты визуализации (4);
- таксономию (царство, тип, род, вид, серогруппу/штамм) организма, из которого выделена структура или в котором она встречается в соответствии с указанной публикацией (5);
- ссылки на организм или иной таксономический ранг в базе NCBI Taxonomy вместе с информацией о более поздней по сравнению с датой публикации переклассификации организма либо о переименовании таксона (5,12);
- расширенную библиографию (ссылки на публикацию в мировых библиографических базах, аффилиации авторов, их контакты и т.д.) (6), реферат (7) и ключевые слова (8);
- тип структурной единицы, аннотацию класса соединения и место, где эта структура обсуждается в пределах статьи (напр., номер рисунка) (9);
- методы, использованные в данной статье для установления структуры, модификации соединения или его дополнительного исследования (10);
- ссылки на структуру в других базах данных и ссылки на родственные записи в CSDB (11);

1. (CSDB ID: 27282) (1) Report data error

Pieretti G, Carillo S, Lindner B, Kim KK, Lee KC, Lee JS, Lanzetta R, Parrilli M, Corsaro MM

**Characterization of the core oligosaccharide and the O-antigen biological repeating unit from Halomonas stevensii lipopolysaccharide: the first case of O-antigen linked to the inner core** (2)

*Chemistry* **18(12)** (2012) 3729-3735




(3)

Show legend
Show as text (4)

***Halomonas stevensii***
(NCBI TaxID 502821, species name lookup)
(5)

*Taxonomic group:* bacteria / Proteobacteria *(Phylum: Proteobacteria)*

*NCBI PubMed ID:* 22334398
*Publication DOI:* 10.1002/chem.201102550
*Publisher:* Vch Verlagsgesellschaft
*Correspondence:* corsaro@unina.it
*Institutions:* Dipartimento di Chimica Organica e Biochimica, Universita Federico II di Napoli, Complesso Universitario Monte S. Angelo, Via Cintia 4, 80126 Napoli (Italia), Fax: (+39)081674393. (6)


A novel core structure among bacterial lipopolysaccharides (LPS) that belong to the genus Halomonas has been characterized. H. stevensii is a moderately halophilic microorganism, as are the majority of the Halomonadaceae. It brought to light the pathogenic potential of this genus. On account of their role in immune system elicitation, elucidation of LPS structure is the mandatory starting point for a deeper understanding of the interaction mechanisms between host and pathogen. In this paper we report the structure of the complete saccharidic portion of the LPS from H. stevensii. In contrast to the finding that the O-antigen is usually covalently linked to the outer core oligosaccharide, we could demonstrate that the O-polysaccharide of H. stevensii is linked to the inner core of an LPS. By means of high-performance anion-exchange chromatography with pulsed amperometric detection we were able to isolate the core decasaccharide as well as a tridecasaccharide constituted by the core region plus one O-repeating unit after alkaline degradation of the LPS. The structure was elucidated by one- and two-dimensional NMR spectroscopy, ESI Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry, and chemical analysis.

(7)

*lipopolysaccharide, LPS, NMR, O-antigen, oligosaccharide, structure, core, core oligosaccharide, O-polysaccharide, O polysaccharide, bacteria, biological repeating unit, alkaline degradation, anion-exchange chromatography, pulsed amperometric detection, Halomonas stevensii*
(8)

*Structure type:* polymer biological repeating unit
*Location inside paper:* p.3730, scheme 1
*Compound class:* O-antigen (9)

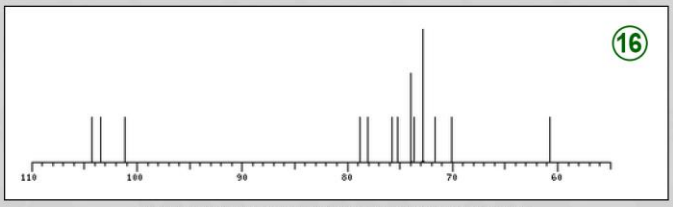
*Methods:* 1H NMR, 13C NMR, NMR-2D, methylation, GC-MS, sugar analysis, 31P NMR, ESI-FTICR-MS, GC, alkaline hydrolysis, de-O-acylation with hydrazine, NMR-1D, HPAEC-PAD (10)

*Related record ID(s):* 28574 (11)
*NCBI Taxonomy refs (TaxIDs):* 502821 (12)

*NMR conditions:* in D2O at 283 K (13)
*$^1$H NMR data:*

| Linkage | Residue | H1 | H2 | H3 | H4 | H5 | H6 |
|---|---|---|---|---|---|---|---|
| 3 | bDGlcp | 4.43 | 3.17 | 3.49 | 3.57 | 3.43 | 3.72<br>3.84 |
| 4 | aDGlcp | 4.73 | 3.35 | 3.57 | 3.29 | 3.99 | ? |
| | bDGalp | 4.36 | 3.40 | 3.59 | 3.83 | 3.61 | ? |

(14)



(16)

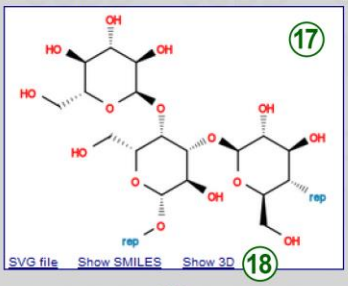
*The spectrum also has 2 signals at unknown positions (not plotted).*

*$^{13}$C NMR data:*

| Linkage | Residue | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|
| 3 | bDGlcp | 103.5 | 74.0 | 75.2 | 78.8 | 75.7 | 60.7 |
| 4 | aDGlcp | 101.2 | 72.8 | 73.6 | 70.1 | 72.7 | ? |
| | bDGalp | 104.3 | 71.7 | 72.9 | 78.1 | 73.9 | ? |

(15)

There is only one chemically distinct structure:



(17)

SVG file Show SMILES Show 3D (18)

Convert structure to GlycoCT condensed
Simulate NMR spectra (GODESS)
Show Sweet-II 3D model Generate 3D coords by GLYCAM (19)

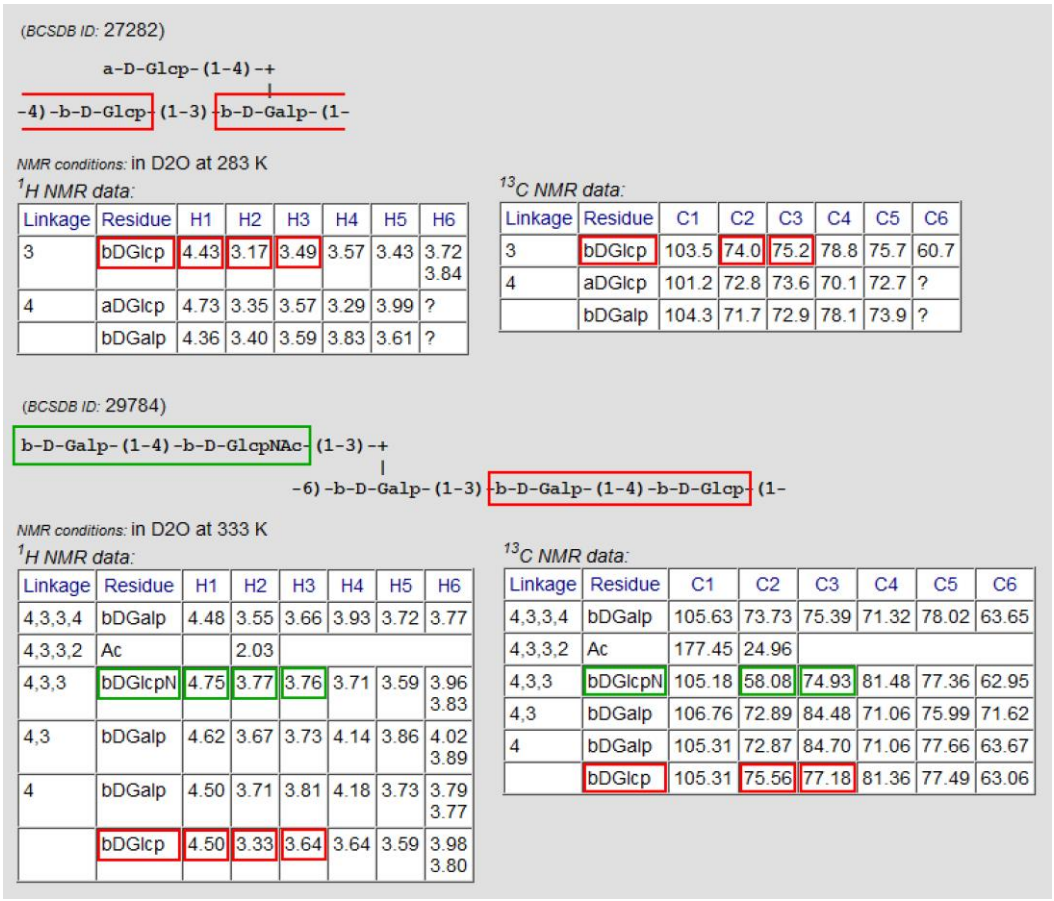
Collapse this record (20)

Рис. 59. Переход к записи CSDB (в полной форме).

- экспериментальные условия записи спектров ЯМР (температура, растворитель, pH) (13), таблицы отнесения сигналов в протонном (14) и углеродном (15) спектрах и схематическое представление спектра ЯМР $^{13}C$ для визуального сравнения (16);
- структурную формулу соединения (17). Если структура имеет неопределённости, ей может соответствовать более одной структурной формулы даже с учётом обобщения с помощью отрисовки простых связей вместо *стерео*-направленных. В этом случае страница записи содержит меню с выбором вариантов.
- инструменты работы со структурой в атомарном формате (18): экспорт рисунка, генерирование SMILES и запуск модуля 3D-моделирования и работы с атомными координатами;
- дополнительные инструменты работы со структурой (19): перевод в другие углеводные нотации, моделирование спектров ЯМР, отправка в другие проекты для стороннего 3D-моделирования.
- Ссылка (20) переключает развёрнутую форму отображения на краткую и обратно.

Для оценки эффекта введения аминогруппы на химические сдвиги, следует рассмотреть таблицы отнесения спектров ЯМР в записях найденных соединений. Рис. 60 демонстрирует эти таблицы из записей 27282 и 29784, первая из которых описывает нативный лактозный фрагмент (в данном случае занимающий части двух соседних повторяющихся звеньев полимера), а вторая – как нативный, так и нитрогенированный фрагменты. Обведённые рамками химические сдвиги сигналов глюкозы и глюкозамина позволяют выявить альфа- и бета-эффекты аминирования 4-β-галактозамещенной β-D-глюкозы непосредственно из таблиц (путём усреднения разности между «красными» и «зелёными» значениями) или опосредованно с помощью инструмента ЯМР-моделирования соединений из этих записей ((19) на Рис. 59). По данным этих записей для $^{13}C$ они составили –16.7 м.д. на α-углеродном атоме и –1.3 м.д. на β-углеродном атоме C3. Эффект на β-углеродном атоме C1 в данном случае нерепрезентативен, так как моносахариды с *глюко*-конфигурацией в приведённых структурах имеют разные остатки-акцепторы и типы связей с ними. Наличие ацетильной группы во втором

положении глюкозамина и отсутствие кислых моносахаридов исключает погрешности, связанные с разницей естественной кислотности раствора сахарида, несущего заряженные группы. Тем не менее в случае, когда в публикациях указан pH образца, записи, содержащие структуры со свободной аминогруппой глюкозамина, могут быть использованы для сравнения с учётом значения pH.



(BCSDB ID: 27282)

a-D-Glcp-(1-4)-+
|
-4)-b-D-Glcp-(1-3)-b-D-Galp-(1-

NMR conditions: in D2O at 283 K

$^{1}H$ NMR data:

| Linkage | Residue | H1 | H2 | H3 | H4 | H5 | H6 |
|---|---|---|---|---|---|---|---|
| 3 | bDGlcp | 4.43 | 3.17 | 3.49 | 3.57 | 3.43 | 3.72<br>3.84 |
| 4 | aDGlcp | 4.73 | 3.35 | 3.57 | 3.29 | 3.99 | ? |
| | bDGalp | 4.36 | 3.40 | 3.59 | 3.83 | 3.61 | ? |

$^{13}C$ NMR data:

| Linkage | Residue | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|
| 3 | bDGlcp | 103.5 | 74.0 | 75.2 | 78.8 | 75.7 | 60.7 |
| 4 | aDGlcp | 101.2 | 72.8 | 73.6 | 70.1 | 72.7 | ? |
| | bDGalp | 104.3 | 71.7 | 72.9 | 78.1 | 73.9 | ? |

(BCSDB ID: 29784)

b-D-Galp-(1-4)-b-D-GlcpNAc-(1-3)-+
|
-6)-b-D-Galp-(1-3)-b-D-Galp-(1-4)-b-D-Glcp-(1-

NMR conditions: in D2O at 333 K

$^{1}H$ NMR data:

| Linkage | Residue | H1 | H2 | H3 | H4 | H5 | H6 |
|---|---|---|---|---|---|---|---|
| 4,3,3,4 | bDGalp | 4.48 | 3.55 | 3.66 | 3.93 | 3.72 | 3.77 |
| 4,3,3,2 | Ac | | 2.03 | | | | |
| 4,3,3 | bDGlcpN | 4.75 | 3.77 | 3.76 | 3.71 | 3.59 | 3.96<br>3.83 |
| 4,3 | bDGalp | 4.62 | 3.67 | 3.73 | 4.14 | 3.86 | 4.02<br>3.89 |
| 4 | bDGalp | 4.50 | 3.71 | 3.81 | 4.18 | 3.73 | 3.79<br>3.77 |
| | bDGlcp | 4.50 | 3.33 | 3.64 | 3.64 | 3.59 | 3.98<br>3.80 |

$^{13}C$ NMR data:

| Linkage | Residue | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|
| 4,3,3,4 | bDGalp | 105.63 | 73.73 | 75.39 | 71.32 | 78.02 | 63.65 |
| 4,3,3,2 | Ac | 177.45 | 24.96 | | | | |
| 4,3,3 | bDGlcpN | 105.18 | 58.08 | 74.93 | 81.48 | 77.36 | 62.95 |
| 4,3 | bDGalp | 106.76 | 72.89 | 84.48 | 71.06 | 75.99 | 71.62 |
| 4 | bDGalp | 105.31 | 72.87 | 84.70 | 71.06 | 77.66 | 63.67 |
| | bDGlcp | 105.31 | 75.56 | 77.18 | 81.36 | 77.49 | 63.06 |

Рис. 60. Части записей CSDB 27282 и 29784, содержащие отнесение сигналов ЯМР. Химические сдвиги, значительно отличающиеся в остатках глюкозамина и глюкозы в лактозном фрагменте, обведены рамками.

*4.1.2. Поиск бактериальных углеводов, содержащих галактуроновую кислоту и ещё как минимум одну гексозу, структура которых опубликована после 2005 года в связи с антигенной активностью.*

Эта задача может быть решена с помощью поиска по составу (**Composition search** в главном меню). Мономерный состав, полученный из экспериментов по хроматографии и/или масс-спектрометрии, зачастую ограничивает данные, известные о природном сахариде. Поисковая форма, показанная на Рис. 61, описывает частичный моносахаридный состав, включающий неизвестную гексозу и галактуроновую кислоту. Каждому мономеру соответствует один пронумерованный ряд. Селектор (1) позволяет выбрать мономер из нескольких десятков наиболее распространённых остатков или их классов. Для ввода остатка, отсутствующего в списке, воспользуйтесь вариантом **Show all residues**, который открывает окно с полным списком остатков, поддерживаемых CSDB. Выбранные остатки и их минимальное количество (2) в структурной единице отображаются в области предпросмотра (3). Кнопки (4) **Add unit (+)** и **Remove unit** (-) увеличивают или уменьшают количество разнородных остатков в желаемом составе.

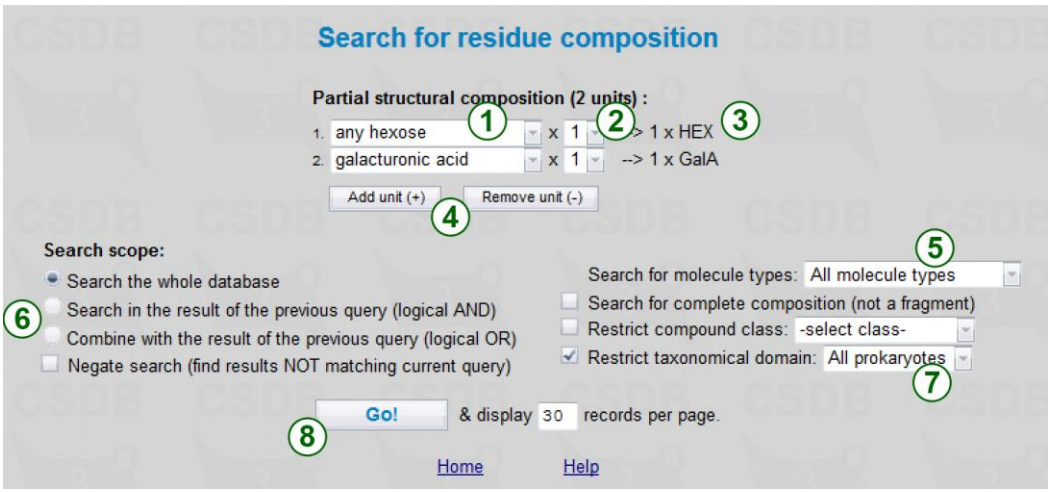



Рис. 61. Запрос на поиск по составу.

В этом запросе мы будем искать структуры, содержащие указанные мономеры, среди молекул всех типов (5), во всей базе данных (6) и без ограничений на класс (7). Отсутствие отметки **Search for complete composition** (7) указывает на то, что искомые структуры могут кроме указанных содержать и другие мономеры. Поскольку данный пример связан именно с бактериальными сахаридами, ограничим область поиска прокариотическими структурами, выбрав `All Pro-`

`karyotes` в селекторе **Restrict taxonomical domain** (7). В результаты попадёт также какое-то количество структур из домена архей. Нажатие на кнопку **Go!** (8) запускает поиск и возвращает 738 структур, состав которых включает как минимум два указанных моносахарида.

Теперь уточним результаты с учётом библиографических данных. Форма поиска по библиографии (**Bibliography search** в главном меню) позволит выбрать из предыдущих результатов только те структуры, которые были опубликованы после 2005 года с условием, что в названии или ключевых словах публикации встречается термин `antigen` (Рис. 62).

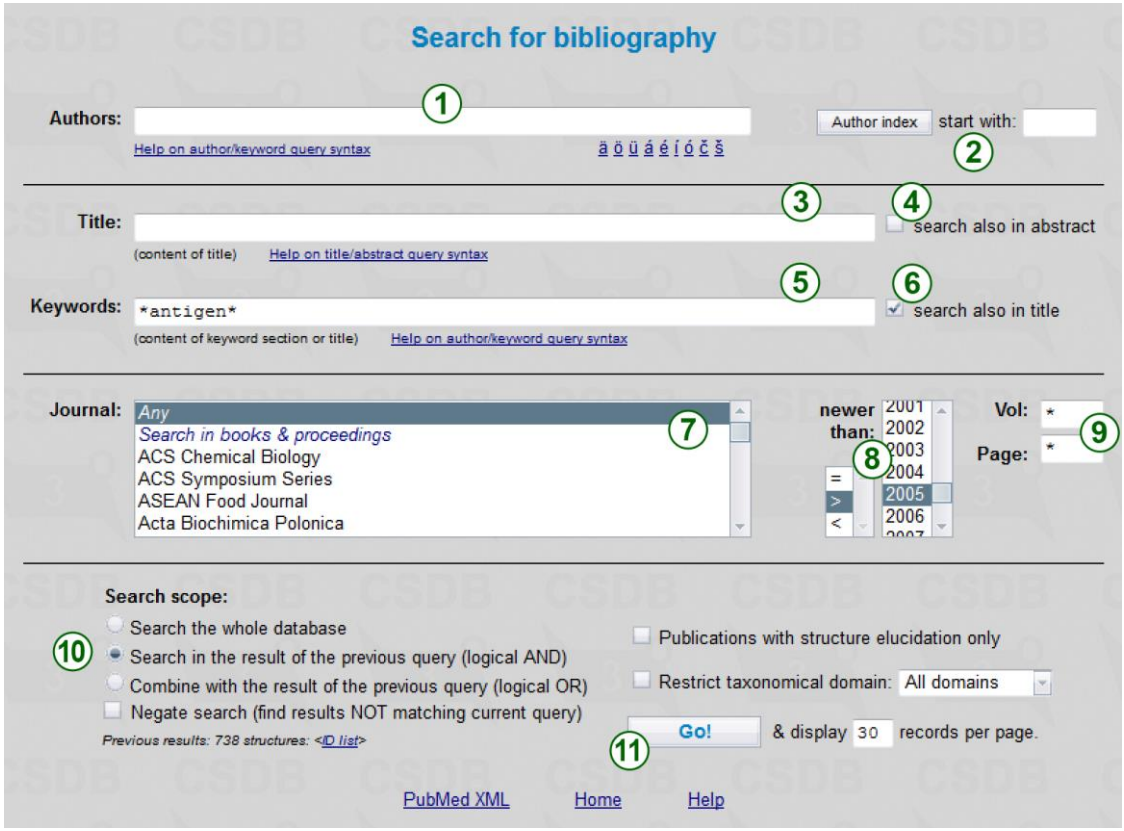



Рис. 62. Уточнение результатов с помощью запроса на поиск по библиографии.

В рамках поиска по библиографии возможно создание сложных запросов, включающих выходные данные и мета-информацию о публикациях. Ни один из критериев не является обязательным. Поисковая форма позволяет указать:

- авторов - непосредственно (1; **Authors**) или путём выбора из авторского указателя (2);
- название статьи или его часть (3; **Title**); при установленной отметке **search also in abstract** (4) поиск текста будет проводиться также в рефератах статей;

- ключевые слова публикации (5; **Keywords**) с возможностью проверки также и названия статьи (6);
- название журнала путём выбора из списка (7; **Journal**);
- год публикации (8) и способ его интерпретации (**<** до, **>** после, **=** строго в этот год);
- том и страницы (9; **Vol** и **Page**);

Поля (1), (3) и (5) поддерживают язык запросов, который кроме терминов может содержать логические операции AND, OR и NOT, задание порядка операций с помощью скобок, группировку терминов с помощью кавычек и символы-заменители (* и ?). По умолчанию поиск чувствителен к регистру и нечувствителен к символам с акцентами (1), т.е. нет необходимости вводить умляуты и другие национальные символы.

Для поиска статей содержащих среди ключевых слов или в заголовке термин antigen в любой форме (antigens, antigenic, O-antigen и т.д.), введём в поле (5) `*antigen*` и поставим отметку (6). Символ-заменитель «*» означает произвольный набор символов. Выбрав опцию «после» (**>**) и год 2005 в селекторах (8), мы ограничим поиск данными, опубликованными после 2005 года. Для пересечения текущего библиографического запроса с предыдущим структурным запросом следует выбрать область поиска **Search in the results of the previous query** (10) и нажать на кнопку **Go!** (11).

Результат содержит 73 публикации, удовлетворяющие критериям поиска и описывающие любую из структур, найденную предыдущим запросом. Одна из них приведена на Рис. 63.

В отличие от структурного поиска, возвращающего объекты типа «соединение», библиографический поиск возвращает список объектов типа «публикация». В данном случае (пересечение запросов разного типа) второй запрос вернёт публикации, каждая из которых содержит одну или более структур, хотя бы одна из которых соответствует предыдущему запросу по составу.

Для каждой публикации отображается её идентификатор (1), авторы, название и выходные данные (2), реферат (3), ключевые слова (4) и список соединений, обсуждаемых в статье. Каждый элемент этого списка содержит идентификатор гликана (5), его первичную структуру (6), переключатель формата

отображения между SweetDB и SNFG (7), организм или таксон, в связи с которым эта структура обсуждается в статье, и ссылку на соответствующую запись CSDB (8), включающие все доступные данные.

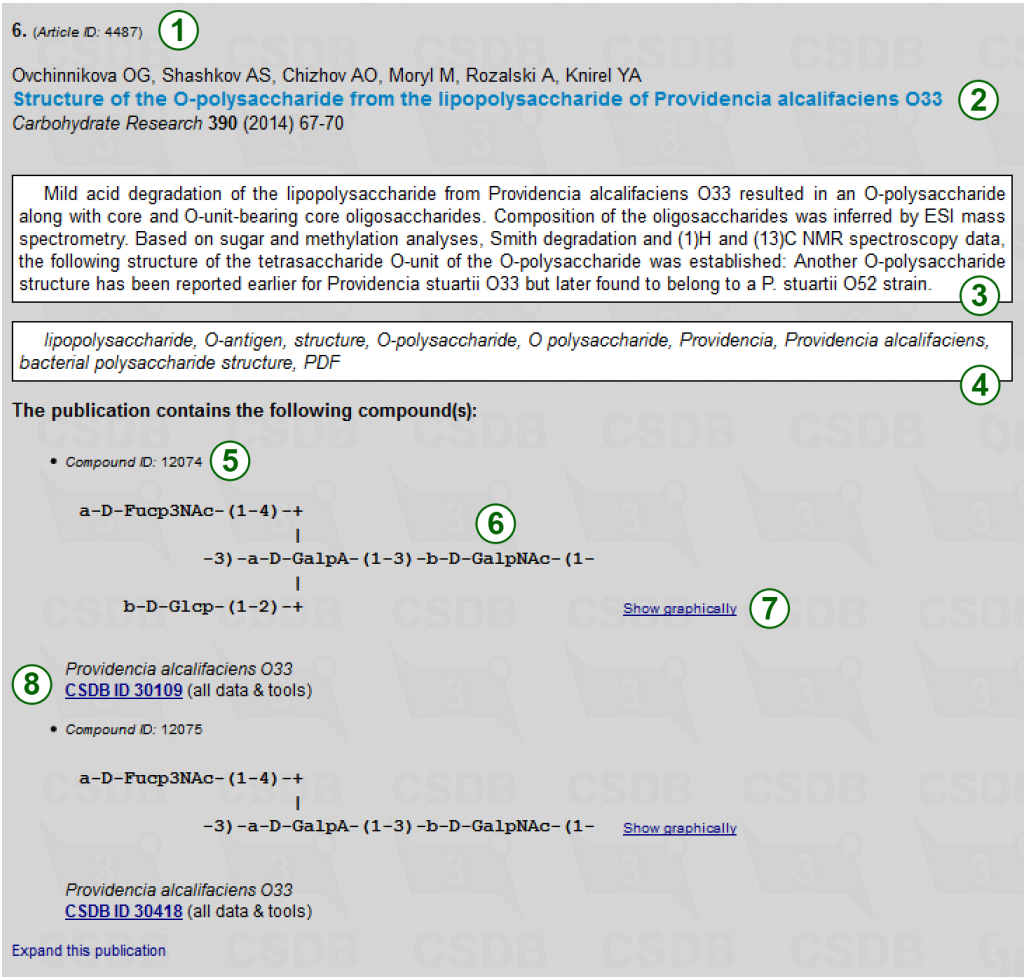



Рис. 63. Результат комбинированного библиографического запроса в краткой форме. Показаны данные только по одной публикации; структуры отображены в формате SweetDB.

Оба соединения, описанные в публикации на Рис. 63 имеют состав (1 x HEX + 1 x GalA + ...). Однако, статья включается в результаты поиска, если хотя бы одно обсуждаемое в ней соединение удовлетворяет критериям структурного поиска, при этом она может включать и другие гликаны. Из реферата следует, что роль первой структуры – О-антиген бактерии *Providencia struartii*. Для более детального исследования следует перейти к полной записи, и оттуда – к полному тексту статьи во внешней библиографической базе данных или на сайте издательства.

*4.1.3. Поиск соланидиносодержащих гликоконъюгатов, выделенных из растений рода Паслён.*

Решение этой задачи лучше всего начать с поиска всех организмов, принадлежащих определённому таксономическому роду (в данном случае *Паслён – Solanum*). Для этого предназначен поиск по таксономии (Рис. 64), доступный в главном меню как **Taxonomy search**.

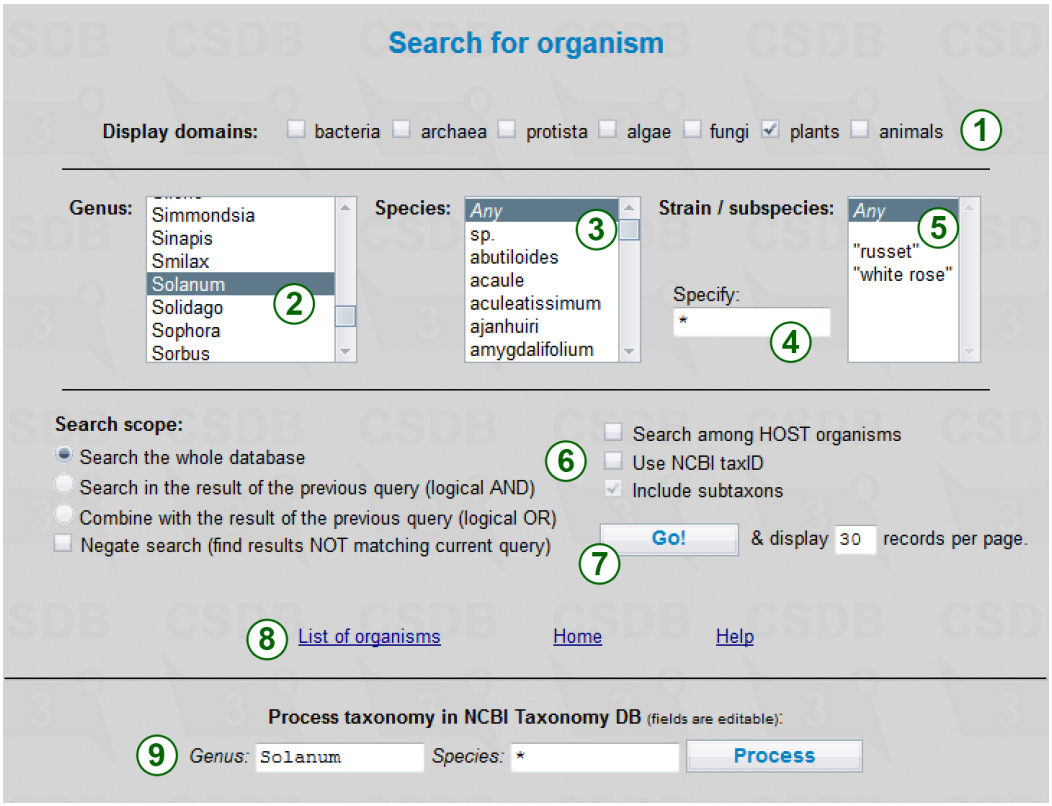



Рис. 64. Запрос на поиск по таксономии.

Группа отметок **Display domains** (1) позволяет выбрать домены, для которых будут отображаться рода, чтобы не загромождать список заведомо неподходящими названиями. В данном примере это царство растений (**plants**). По умолчанию выбраны все домены, кроме царства животных, которые в соответствии с декларированным покрытием CSDB присутствуют в базе лишь в контексте организмов-хозяев микроорганизмов, но не непосредственных носителей биогликанов.

При выборе рода (2; **Genus**) списки видов (3; **Species**) и более мелких таксономических рангов (5) обновляются в соответствии с выбранным родом. Для задания рода *Паслён (лат. Solanum)* выбираем `Solanum` из списка родов (2) и `Any` из списка видов (3), что означает поиск всех видов, принадлежащих роду, а

также организмов этого рода с неустановленным видом и гибридов. Меньшие ранги (подвид, мутант, штамм) и серогруппы (в случае бактерий) можно ввести в поле (4) или выбрать из доступных для данного рода с помощью селектора (5). Вариант `Any` означает отсутствие ограничений, как и символ `*` в поле (4).

Мы проводим поиск по всей базе (6); по умолчанию будут найдены организмы, относящиеся как к самому указанному таксону, так и к его субтаксонам. Для поиска по родам и видам это единственный оправданный подход, поэтому отметку **Include taxonomic children** можно снять, только если вместо селекторов для идентификации таксонов используется идентификатор в базе NCBI Taxonomy (отметка и поле ввода **NCBI TaxID**).

Среди прочих возможностей стоит отметить фильтр на поиск по организмам-носителям (**Search among HOST organisms**), возвращающий при комбинировании со структурным поиском объекты, выделенные не из самого организма, а из инфицирующих его микроорганизмов или симбионтов. Ссылка (8; **List of organisms**) возвращает полный список организмов, присутствующих в базе CSDB, а область (9) позволяет получить дополнительные данные (напр., положение на дереве жизни) о выбранном таксоне из базы данных NCBI Taxonomy.

Нажатие на кнопку **Go!** (7) запускает поиск и возвращает 48 организмов рода *Паслён*. Для выбора из продуцируемых ими гликоконъюгатов только тех, которые содержат стероидный алкалоид соланидин, воспользуемся уточнением запроса с помощью структурного поиска (Рис. 65).

Соланидин не является углеводным остатком, поэтому может присутствовать в структурах в CSDB только как агликон или как нестандартный заместитель. Для поиска среди этих компонентов структуры введём `solanidine` в текстовое поле (1) и установим отметку **in aglycons, aliases or linear code** (2). Этот поиск будет включать структурные единицы всех типов (3) без дополнительных ограничений и будет проводиться среди результатов предыдущего таксономического запроса, о чем свидетельствует выбор опции **Search in the results of the previous query** (4). Нажатие на кнопку **Go!** (5) запускает поиск и возвращает 16 гликоконъюгатов соланидина, выделенных из растений рода *Паслён*. Вид страницы с результатами (объектами типа «структура») аналогичен рассмотренному в примере 4.1.1.

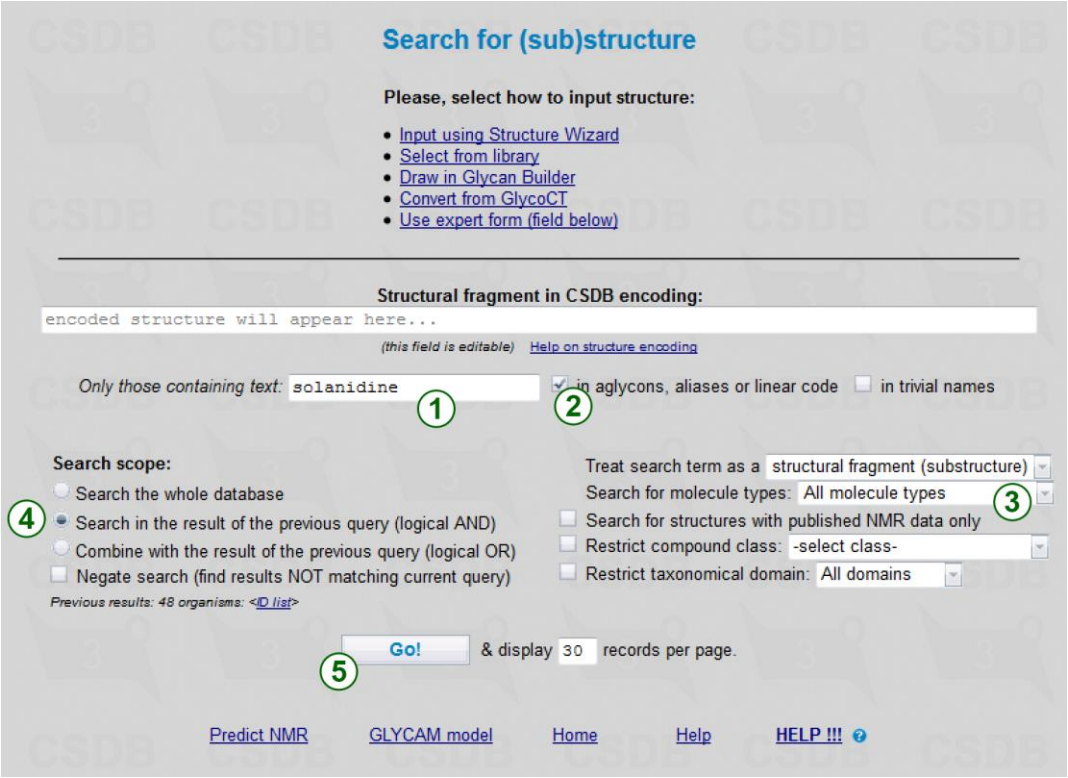



Рис. 65. Уточнение результатов с помощью запроса на поиск по агликону.

*4.1.4. Поиск углеводов, кроме октозосодержащих, имеющих в спектре ЯМР $^{13}C$ характеристичный сигнал вблизи 34 м.д.*

Эта задача подразумевает поиск фрагментов структуры, имеющих характерные признаки в спектрах ЯМР, выявленные экспериментально. Чаще всего сигнал в области 30-40 м.д. свидетельствует о дезоксигенированном атоме углерода. Однако большая распространённость структур бактериального кора, содержащих 3-дезокси-D-маннокт-2-улозоновую кислоту (Kdo), имеющую сигнал в этой области, будет затруднять выявление необычных остатков среди множества результатов. Поэтому мы будем искать только те структуры, не содержащие октозы, среди которых Kdo и родственные остатки представляют абсолютное большинство. Первым шагом станет поиск спектров ЯМР с сигналом, близким к указанному, и гликанов, для которых опубликованы такие спектры. Для этого предназначена опция **NMR signal search** в главном меню, отображающая форму поиска данных ЯМР (Рис. 66).

Эта форма позволяет выбрать углерод или протоны в качестве ядра (1; **Nucleus**) и ввести интересующие химические сдвиги сигналов (3; **Chemical shifts**) в произвольном порядке и с любой точностью, разделённые пробелами или переносами строк. Порог схожести (2; **Threshold**) определяет, насколько далеки части искомых спектров могут быть от указанных сигналов. В данном примере эта часть состоит из единственного сигнала. Для определения схожести между введённым спектром и спектром из базы данных, из большего спектра выделяется наиболее близкий подспектр с тем же числом сигналов, что и в меньшем спектре, после чего сигналы сортируются, попарно вычитаются, и абсолютные значения отклонений усредняются. Схожесть оценивается как обратная величина этого среднего отклонения. Два зарезервированных крайних значения составляют 0 (полное отсутствие схожести) и 1000 (полное совпадение спектров). Разумные значения порогов схожести по умолчанию составляют 1.2 для спектров $^{13}C$ и 5.0 для спектров $^{1}H$. Эти значения обеспечивают поиск спектров в пределах отклонений, типичных для спектроскопии ЯМР углеводов, но в то же время не приводят к избыточному количеству плохо подходящих результатов.

Мы проводим поиск по всей базе (4). Отметка **Signals within a single residue** (5) по умолчанию установлена. Она означает, что все перечисленные сигналы

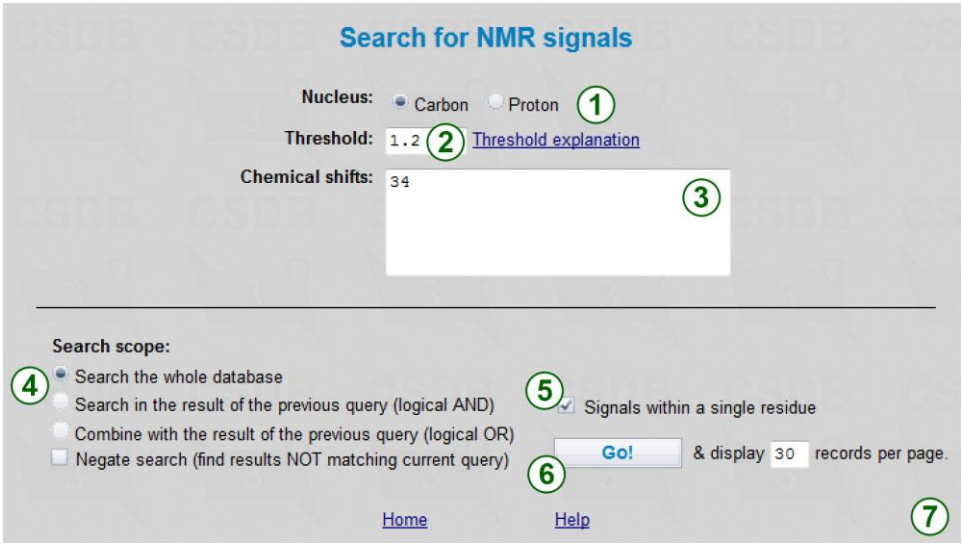



Рис. 66. Запрос на поиск по сигналам ЯМР.

должны принадлежать одному и тому же остатку в структуре. Поскольку введён всего один сигнал, это не имеет значения, но в остальных случаях может существенно ограничить область поиска и ускорить его, исключив результаты, в которых разные сигналы относятся к разным остаткам. Нажатие на кнопку **Go!** (6) запускает поиск и возвращает 133 соединения, список которых отсортирован по схожести спектров с указанными сигналами. В случае поиска только одного сигнала, схожесть равна обратной величине от разницы указанного химического сдвига и положения ближайшего к этому значению сигнала в спектре.

Рис. 67 демонстрирует один из элементов результирующего списка структур. Для каждого найденного соединения отображается его идентификатор (1), первичная структура (2) и переключатель форматов (3; SweetDB и SNFG) и тип структурной единицы. Далее следует метрика схожести искомых и найденных сигналов и все спектры ЯМР на интересующем ядре, присутствующие в базе для этого соединения. Спектры содержат ссылку на статью, в которой они опубликованы (5), условия эксперимента (6), отнесение сигналов (7) и схематическое изображение (8; для $^{13}$C). Если спектров несколько, для вычисления схожести используется усреднение. Сигналы, соответствующие поисковому запросу, подсвечены жёлтым в таблице отнесения. Как и в случае структурного поиска, структуры сопровождаются выходными данными статей, в которых они опубликованы (9), ссылками на соответствующие записи CSDB (10) и информацией об организмах, ассоциированных со структурой в этих записях.

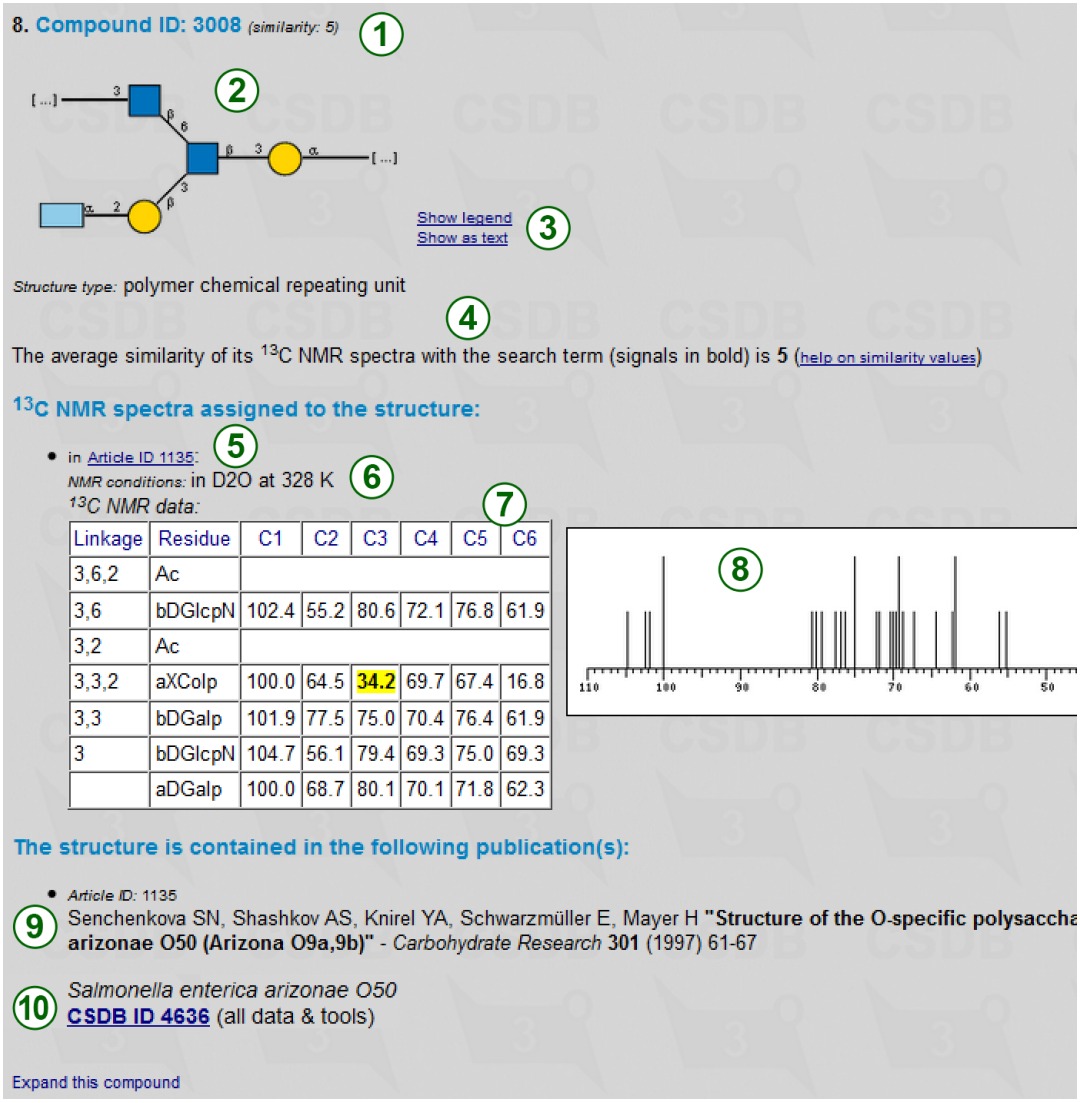



Рис. 67. Результат ЯМР-спектроскопического запроса в краткой форме (показаны данные по одного соединения).

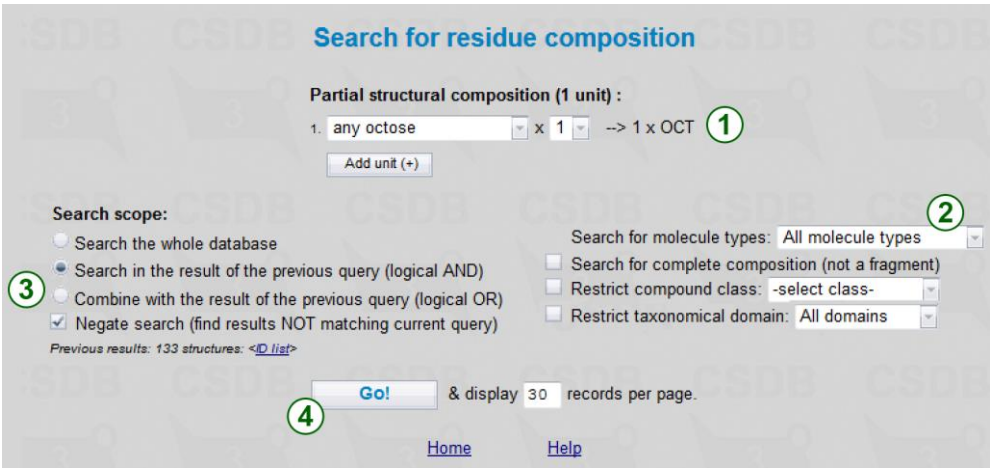



Рис. 68. Исключение из предыдущего результата октозосодержащих структур с помощью уточняющего запроса на поиск по составу.

Для исключения из списка результатов октозосодержащих структур применим поиск по составу (Рис. 68). В этой форме следует выбрать класс `Any octose` (любая октоза) в качестве единственного критерия (1); другие ограничения не требуются (2). Чтобы произошло вычитание результатов текущего запроса из результатов предыдущего, требуется выбрать область поиска как **Search in the result of the previous query** и поставить отметку **Negate search** (3), тем самым применив логическую операцию «И НЕ» (AND NOT). Нажав на кнопку **Go!** (4), мы получим список структур, не содержащих октозы и имеющих хотя бы один сигнал, близкий к 34 м.д., в спектре ЯМР $^{13}$C. Этот список содержит 71 структуру, в которых такой сигнал обеспечивается неуглеводными остатками либо моносахаридами, отличными от Kdo (3-дезокси-D-манноктулозоновой кислоты), Ko (D-глицеро-D-талоктулозоновой кислоты) и им подобных, например, 3,6-дидезоксигексозами.

Поскольку каждая поисковая форма предлагает несколько критериев поиска, использовать логическое отрицание стоит с осторожностью – чётко понимая, что именно будет отрицаться. Например, использование отрицания с двумя критериями (частичный состав, включающий октозу (1), и таксономический домен *Прокариоты* – последняя опция в блоке (2)) не имеет смысла, так как с точки зрения булевской логики условие [НЕ (октозосодержащий И прокариотический)] – это то же самое, что [(НЕ октозосодержащий) ИЛИ (НЕ прокариотический)]. Такое условие, очевидно, вернёт результаты, не имеющие практической ценности, так как в них попадут структуры, содержащие октозы, но выделенные не из прокариот. Если требуется отфильтровать результаты, оставив из них только прокариотические, и в то же время требуется сложный запрос с отрицанием, следует использовать структурный поиск ещё раз (в режиме пересечения результатов с результатами предыдущего запроса), указав в качестве структуры `ANY` (любая) и в качестве домена `All prokaryotes`.

*4.1.5. Поиск публикаций Книрель или Шашкова (АС) по бактериальным гликанам, включающим хиновоз-4-амин, амидированный любой N-ацетилированной аминокислотой.*

Решение задачи поиска публикаций одного из двух авторов, идентифицируемых по фамилии или фамилии и инициалам, должно позволить найти только те статьи, где обсуждаются гликаны, имеющие указанные компоненты структуры. Как и ранее, для поиска публикаций воспользуемся библиографическим поиском (Рис. 69).

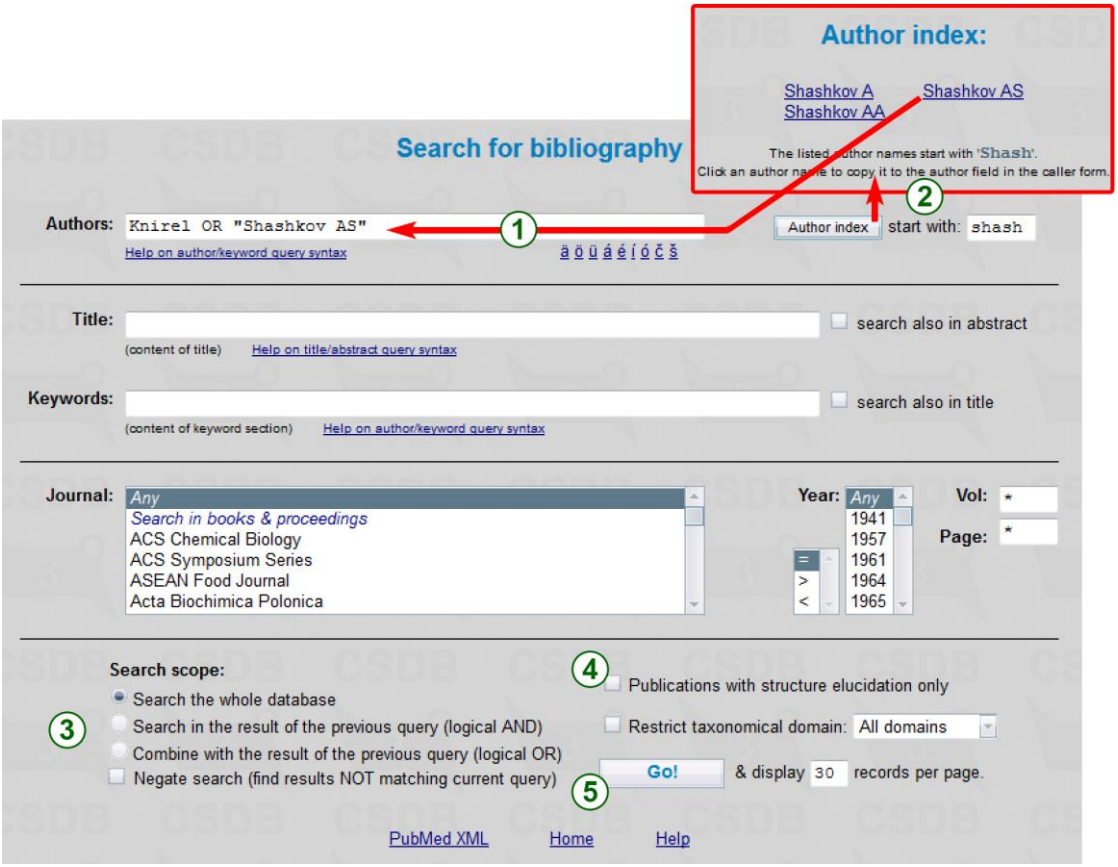



Рис. 69. Запрос на поиск по библиографии с использованием индекса авторов.

Запрос по авторам `Knirel OR "Shashkov AS"` (1; **Authors**) найдёт материалы, опубликованные любым из авторов (или обоими, в числе прочих соавторов). Одна фамилия указана без инициалов, а другая объединена с конкретными инициалами с помощью кавычек, позволяющих включать в термин пробелы. Для помещения правильно написанной фамилии и инициалов конкретного автора в поисковый запрос предназначен авторский указатель. Чтобы воспользоваться им, наберите начальные буквы фамилии автора (хотя бы две) в поле (2) и нажмите на кнопку **Author index**. В появившемся окне, перечисляющем всех авто-

ров с подходящими фамилиями, выберите интересующего автора, и он будет скопирован в поисковый запрос. Проведём поиск по всей базе (3). Отметка **Publications with structure elucidation only** (4) остаётся неотмеченной, чтобы найти не только статьи, где структуры были впервые установлены, но и главы, обзоры и публикации, где эти гликаны обсуждаются. Остальные параметры библиографического поиска описаны в примере 4.1.2. Нажатие на кнопку **Go!** (5) запустит поиск и вернёт список из 770 публикаций.

Для того, чтобы найти среди этих публикаций те, которые описывают структуры, содержащие 4-хиновозамин, амидированный любой N-ацетилированной аминокислотой, используем поиск по фрагменту структуры и введём поисковый запрос с помощью «мастера» структур (**Structure wizard**). Форма «мастера» показана на Рис. 70.

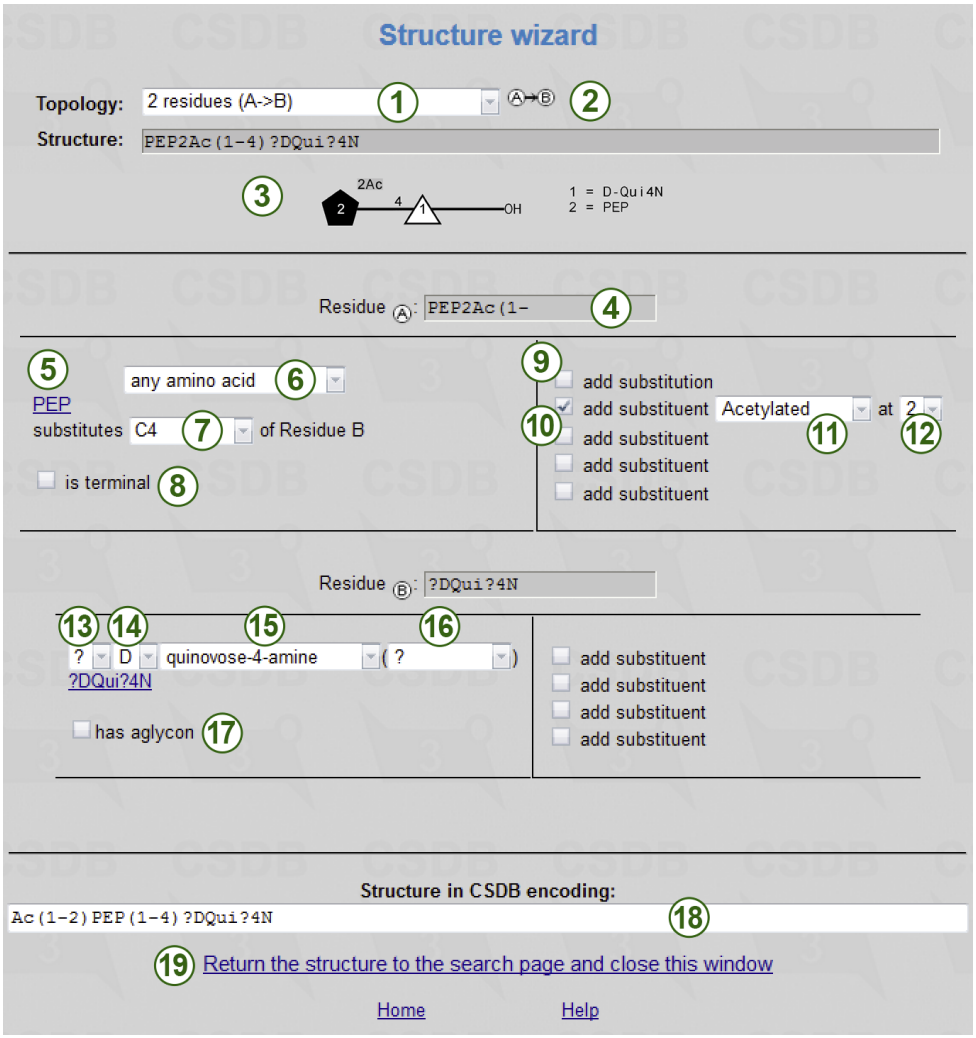



Рис. 70. Визуальный редактор углеводных структур («мастер»).

При использовании «мастера» необходимо выбрать топологию структурного фрагмента и задать свойства каждого из мономерных остатков. Селектор **Topology** (1) предоставляет возможность выбора топологий фрагмента размером до четырёх остатков и имеет область предпросмотра (2). Когда топология выбрана, ниже отображается соответствующее число секций для редактирования её узлов (остатков), в нашем случае это две секции, поскольку выбрана топология «дисахарид». Поле **Structure** (3) содержит текущее состояние структуры в формате IUPAC condensed и ниже - её предпросмотр в формате SNFG. Заголовки секций (4) показывают положение остатка в топологии (напр., `Residue A`) и его название вместе с конфигурациями и химическими модификациями.

Левая часть секции включает селектор остатка (6,15), позволяющий выбрать один из нескольких десятков наиболее распространённых остатков или их классов. В случае, если нужный остаток отсутствует в списке, воспользуйтесь первым элементом **Show all residues** для открытия полного списка остатков. В данном примере, в качестве остатка **A** (донор в дисахаридной топологии) выбран класс `any amino acid` («любая аминокислота», (6)), а в качестве остатка **B** (акцептор в дисахаридной топологии) – остаток `4-aminoquinovose` (4-хиновозамин, (15)). В зависимости от выбранного остатка могут присутствовать селекторы конфигураций: аномерной (13; “α”, “β” или “?”), абсолютной (14; “D”, “L”,“R”, “S” или “?”) и способа циклизации (пираноза, фураноза, открытая форма, полиол, неизвестно). Ссылка (5) показывает полное имя остатка и ведёт к соответствующему элементу полного списка. Селектор связи (7) позволяет выбрать позицию в остатке-акцепторе, с которым текущий остаток образует связь с отщеплением воды. В данном примере, это позиция C4, что означает связь аномерного атома остатка **A** с положением 4 остатка **B**. Аномерным атомом считается C1 для альдоз и C2 для кетоз. В случае если выбранная позиция не может быть замещена в акцепторе в соответствии с его функционализацией или размером углеродного скелета, область предпросмотра (3) будет содержать соответствующее сообщение об ошибке вместо структуры. У остатка, находящегося на восстанавливающем конце фрагмента (в нашем случае это остаток **B**), селектор связи отсутствует, и вместо него присутствует отметка **Has aglycon** (17) и селектор (на рисунке скрыт), позволяющий выбрать один из распространённых агли-

конов в качестве корневого остатка. Отметка **is terminal** (8) для терминальных узлов выбранной топологии позволяет в явном виде указать, что остаток должен находиться на одном из невосстанавливающих концов структуры, т.е. не иметь гликозил-доноров.

Правая половина секций предназначена для добавления химических модификаций остатков. Она используется для указания позиции, в которой остаток образует связь с чем-то за пределами фрагмента (9; **Add substitution**), или для добавления моновалентных заместителей (10; **Add substituent**) в определённом положении (11; **at**). Чаще всего, как и в этом примере, это ацетильные группы. Так как аминогруппа в аминокислотах находится в положении 2, выбираем `Acetylated` at `2` (10,11).

Результирующая структура показана в нотации CSDB Linear в конце страницы (18; **Structure in CSDB encoding**). При нажатии на ссылку **Return the structure to the structure search page** (19) эта строка передаётся в форму структурного поиска и окно «мастера» закрывается.

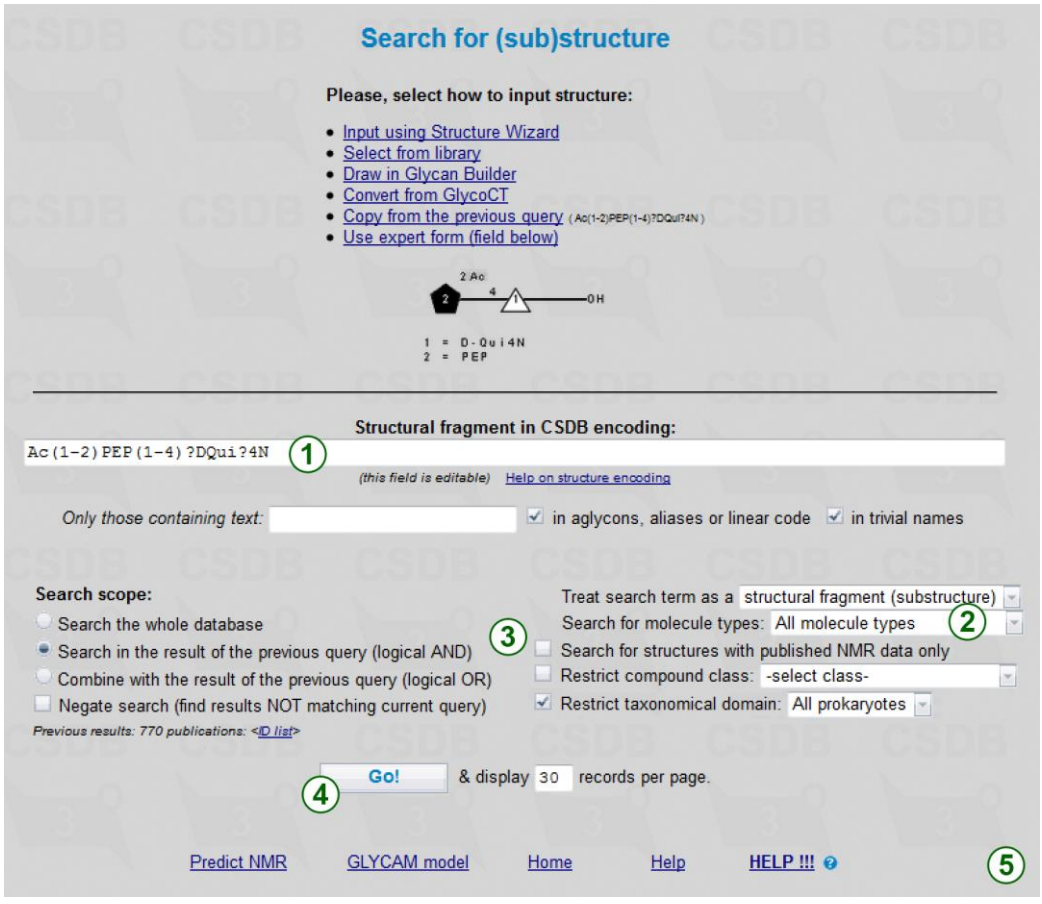



Рис. 71. Уточнение результатов с помощью запроса на поиск по фрагменту структуры.

Поисковая форма с введённым таким способом термином (1) показана на Рис. 71. В блоке ограничений (2) присутствует отметка **Restrict taxonomical domain** и поиск ограничен гликанами прокариот (`All prokaryotes`) в соответствии с исходной задачей. Для пересечения с найденными ранее публикациями, поиск проводится в предыдущих результатах (3; **Search in the results of the previous query**). Нажатие на кнопку **Go!** (4) возвращает 11 гликанов, содержащих указанный димерный фрагмент и опубликованных в статьях Книреля или Шашкова (АС). Формат вывода результатов аналогичен примеру 4.1.1.

*4.1.6. Поиск бактериальных структур, построенных из любых ноноз одного типа (моносахариды или их гомополимеры).*

Этот пример демонстрирует использование опции поиска по точному, а не частичному составу с помощью формы **Composition search** (Рис. 72). Здесь в качестве единственного компонента структуры выбран класс **any nonose** (любая нонноза, (1)), встречающаяся в структурной единице строго один раз (2). Установленная отметка **Search for complete composition (not a fragment)** (4) говорит о том, что структуры должны исчерпываться указанным составом (3), т.е. не иметь других компонентов. Так как выбран только один мономер, это означает, что будут найдены моносахаридные структуры с девятью углеродными атомами и гомополимеры, состоящие из ноноз. Моновалентные неуглеводные заместители не учитываются при подсчёте атомов и остатков, что позволяет найти в том числе и структуры, ацетилированные любым способом. Так как условия поиска включают принадлежность к бактериям, в качестве ограничения на таксономический домен выбрано `All prokaryotes` (5).

Нажатие на кнопку **Go!** (6) возвращает 6 бактериальных гликанов и их производных, состоящих только из ацетилированных нонулозоновых кислот или прочих ноноз одного типа.

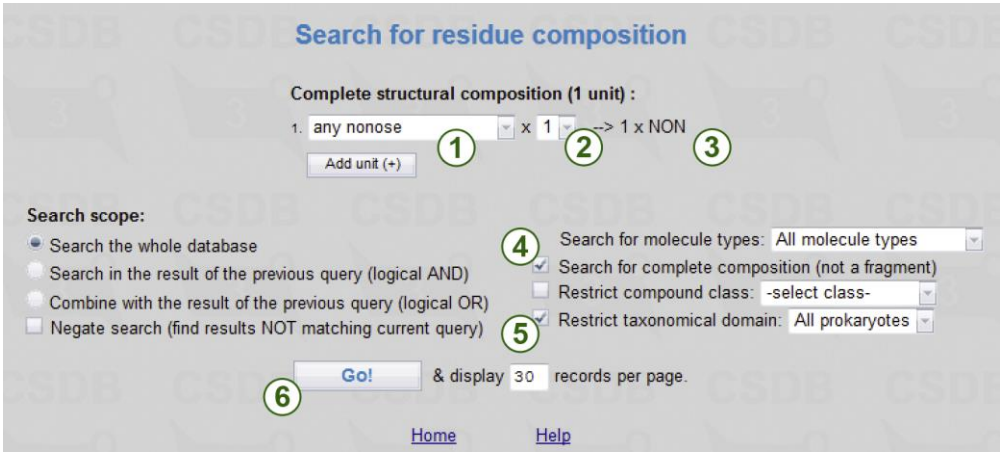



Рис. 72. Запрос на поиск по составу.

*4.1.7. Моделирование спектров ЯМР 3-О-абеквозил-6-дезокси-β-D-манногептопиранозил-(D-рибит-1)-фосфата в водном растворе и оценка точности предсказания наименее достоверных сигналов.*

Более 9200 спектров ЯМР, содержащихся в CSDB, позволяют точно предсказывать химические сдвиги статистическим методом (см. раздел 3.4.1). Интерфейс этого расчётного модуля (Рис. 73) доступен по ссылке **NMR simulation** в меню **Extras**.

Наибольшую точность демонстрирует расчёт спектральных параметров образцов в водных растворах. Возможнось ЯМР-моделирования в других растворителях можно оценить, используя текущее распределение хранящихся в базе спектров по растворителям. Оно доступно по ссылке **Coverage** рядом с селектором растворителя (6).

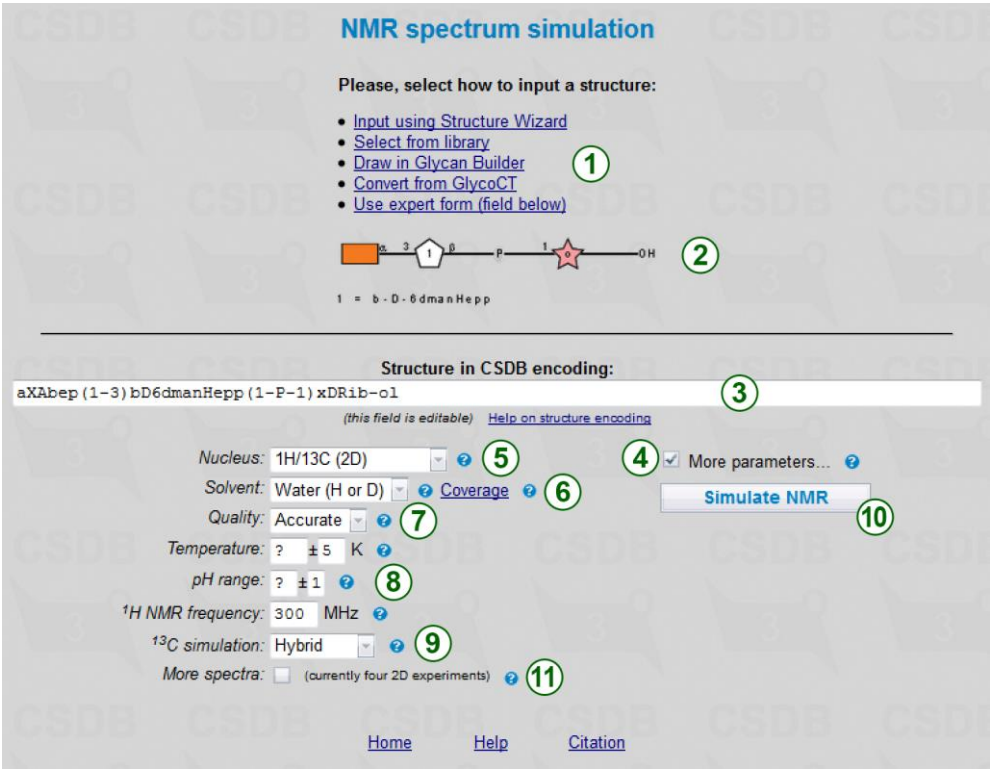



Рис. 73. Введённая структура и параметры ЯМР-моделирования в режиме «Больше параметров».

Простые способы ввода (1) и визуализации (2) структуры аналогичны использованным в форме структурного поиска и рассмотрены в предыдущих примерах, поэтому здесь мы используем экспертный режим – прямой ввод структуры 3-О-абеквозил-6-дезокси-β-D-манногептопиранозил-(D-рибит-1)-фосфата в

нотации CSDB Linear в поле **Structure in CSDB encoding** (3): `aXA-bep(1-3)bD6dmanHepp(1-P-1)xDRib-ol`. Эта структура выбрана для демонстрации возможности корректной работы с необычными углеводными остатками.

По умолчанию, пользователю предлагается проконтролировать минимальный набор параметров: ядро (углерод, протон или оба (5)) и растворитель (6). Полный набор параметров, показанный на Рис. 73, отображается после установки отметки «Больше параметров» (4; **More parameters**).

Выберем ядро $^{1}H/^{13}C$ (5; **Nucleus**) для предсказания гомо- и гетероядерных спиновых корреляций, `Water` в качестве растворителя (6; **Solvent)** и режим точного моделирования `accurate` (7; **Quality**). Будут использованы данные, полученные как в обычной, так и в тяжёлой воде. Кроме режима точного моделирования, возможны варианты быстрого (`fast`) и полного (`extreme`) моделирования, отличающиеся соотношением качества и времени счёта. Если в результате моделирования появляются непредсказанные сигналы (знаки вопроса в таблицах отнесения), имеет смысл изменить режим в сторону повышения качества и перезапустить моделирование ещё раз (10). В режиме `extreme` расчёт может занять до 15 минут. Кроме растворителя (7), возможны дополнительные ограничения на выборку данных с учётом особых экспериментальных условий (8): температуры (**Temperature**) и кислотности среды (**pH range**). Значение частоты спектрометра влияет только на размеры кросс-пиков в двумерных спектрах с учётом предсказанных констант спин-спинового взаимодействия (КССВ).

В соответствии с выбранным методом `Hybrid` (9; **$^{13}$C simulation**) для углеродных химических сдвигов будут использованы как эмпирический, так и статистический подходы, и результаты будут «гибридизованы» (см. раздел 3.4.1). Гибридный режим доступен только если в качестве растворителя выбрана вода. В остальных случаях для углерода и во всех случаях для протонов используется статистический подход к ЯМР-моделированию.

Нажатие на кнопку **Simulate NMR** (10) запускает счёт и приводит к отчёту о результатах, часть которого показана на Рис. 74. В нем содержится копия структуры, краткий обзор сделанных шагов, таблицы отнесения $^{1}H$ и $^{13}C$, цветовой код

остатков, схематическая модель спектра ЯМР $^{13}C$, полученная тремя разными методами, двумерные спектры и дополнительные инструменты.

В таблицах отнесения сигналов каждый остаток представлен одним рядом. Остатки идентифицируются генеалогией, считая от корневого остатка в структурной единице (1; колонка **Linkage**), и названием в нотации CSDB Linear (2; колонка **Residue**). Генеалогия (последовательность позиций замещения) – это путь, который нужно пройти от корневого остатка, чтобы прийти к интересующей точке. Например, в дисахариде Ac(1-2)-β-D-Glc*p*N-(1-3)-α-L-Rha*p* остаток уксусной кислоты имеет генеалогию «3,2». Корневой остаток рамнозы (восстанавливающий конец в олигомере) имеет пустую генеалогию («»). Цветной квадрат около генеалогии – это цветовой код остатка, используемый для раскрашивания двумерных спектров в соответствии с отнесением сигналов.

Колонки **C1..C9**[a] (4) содержат химические сдвиги и сопутствующую информацию. В эмпирической модели (не показана) химические сдвиги дополнены значениями для незамещённых мономеров и применёнными эффектами замещения в м.д., а достоверность предсказания имеет общее значение для всего остатка и вынесена в отдельную колонку **Trust.**

В статистической модели (Рис. 74А.) химические сдвиги дополнены оценкой возможной ошибки в м.д., метрикой достоверности моделирования (в %), числом записей из базы данных, которые были использованы при усреднении, и ссылкой на отчёт о проведённых обобщениях. Нажатие на число записей открывает список идентификаторов этих записей (6, в данном примере список состоит из единственного идентификатора), которые позволяют проследить источники данных по конкретному атому до оригинальных публикаций (7). В этих записях данные, использованные при моделировании этого атома, подсвечены жёлтым.

Метрика достоверности, усреднённая по всем атомам остатка, дополнительно показывается в колонке **Trust** (3). Её значения варьируются от 0 до 100% и показаны на цветном фоне в диапазоне цвета от красного (0%) до зелёного (100%). Общая достоверность модели, представленной в таблице отнесения, выражена в процентах (5) снизу от таблицы. Оценка уровня ошибки модели в м.д.

[a] Для протонных таблиц отнесения - **H1..H9**.

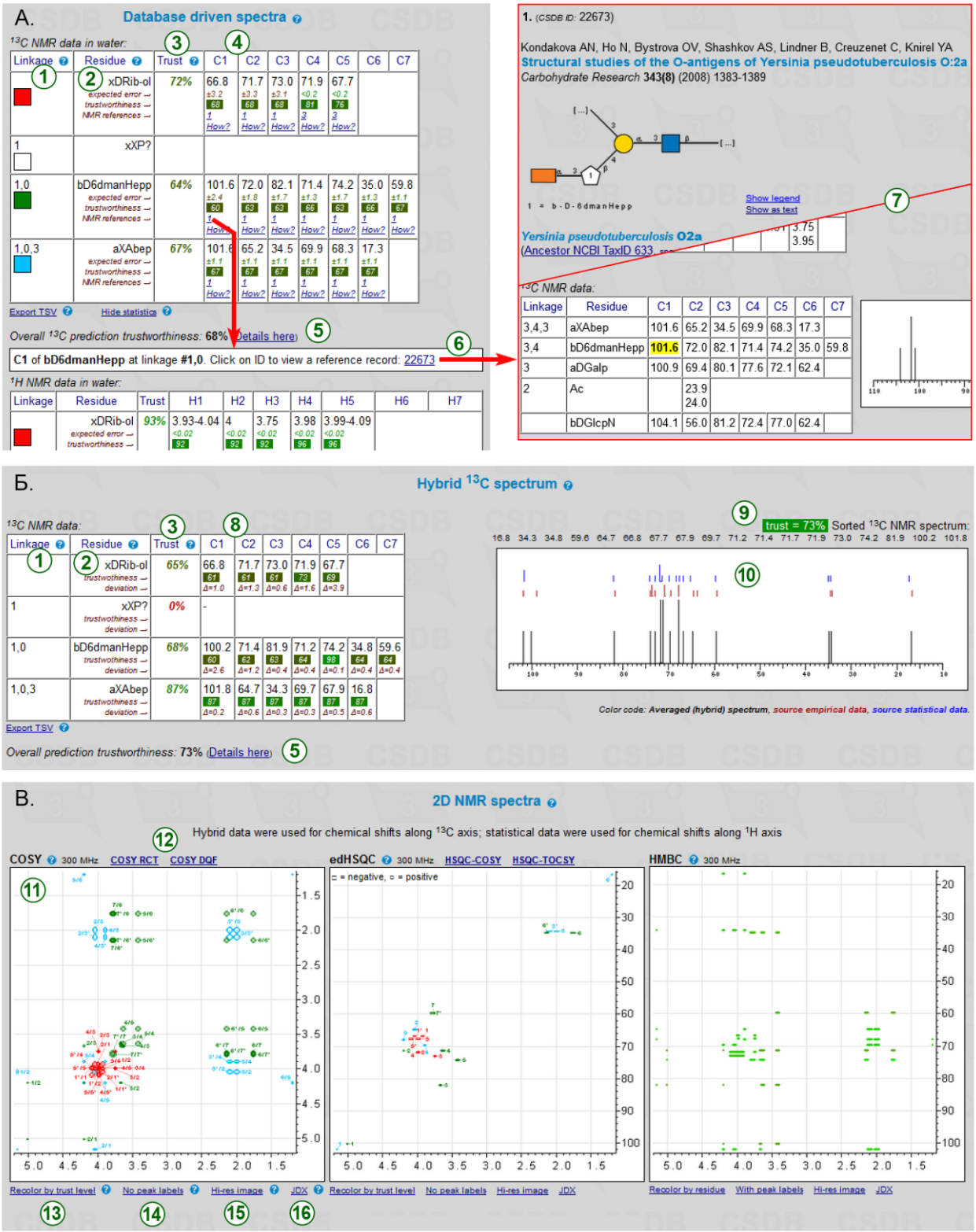



Рис. 74. Результат ЯМР моделирования. А. Отнесение спектра ЯМР $^{13}$C, полученное статистическим методом, цветовой код остатков и отслеживание данных до исходной публикации. Б. Спектр ЯМР $^{13}$C, предсказанный гибридным методом. В. Двумерные спектры ЯМР (показано три спектра: COSY и HSQC в режиме отнесения сигналов с подписями, HMBC в режиме оценки достоверности без подписей).

основана на взаимосвязи точности предсказания с уровнем достоверности, формализованного с помощью линейной регрессии, и показана отдельно для каждого атома цветным шрифтом в диапазоне цвета от красного (>5.0 м.д. для $^{13}C$, грубая оценка) до зелёного (<0.2 м.д. для $^{13}C$, высокая точность).

Гибридная модель (Рис. 74Б) смешивает эмпирические и статистические данные и, как правило, предоставляет наиболее точные результаты. В гибридной таблице отнесения химические сдвиги дополнены производной метрикой достоверности и разностью между эмпирической и статистической моделями (Δ, в м.д.) (8). Референсные значения, полученные обоими подходами, добавлены на схематическое изображение спектра (10): красные линии = эмпирическая модель, синие линии = статистическая модель. Над спектром показана его общая производная достоверность и отсортированный список сигналов для копирования в виде текста. Любая таблица укомплектована ссылкой **Export TSV** для её переноса в Microsoft Excel или другой табличный процессор.

Если моделирование проводилось для ядер $^{1}H$ или $^{1}H/^{13}C$ , полученные химические сдвиги используются для отрисовки моделей двумерных спектров (Рис. 74В). В зависимости от выбранных ядер и состояния отметки **More NMR experiments** генерируется от двух до восьми спектров спиновых корреляций, наиболее популярных в гликохимии (COSY, TOCSY, HSQC, HMBC и их аналоги с переносом когерентности). Моделирование NOESY / ROESY, подразумевающее совершенно иной подход и множество конформационных расчётов, является вопросом ближайшего будущего. В качестве примера на Рис. 74В показано три спектра: **COSY**, **edHSQC** и **HMBC** (11). Ссылки (12) заменяют спектр на его аналоги, получаемые из родственных экспериментов.

Дополнительные возможности предоставляются инструментами, расположенными под спектром. Режим отображения (13) определяет, как будут раскрашены сигналы – в соответствии с отнесением или в соответствии с достоверностью. В первом режиме, цвет сигнала соответствует цветовому коду остатка (в данном примере, COSY и edHSQC), во втором – достоверности моделирования сигнала, в диапазоне от красного (0%) до зелёного (100%) (в данном примере HMBC).

Переключатель подписей сигналов (14) включает и выключает текстовые подписи. В режиме отнесения, подписи содержат номера углеродных атомов, таким образом цвет (=остаток) и подпись (=положение) вместе идентифицируют любую точку углеродного скелета любого остатка и присоединённые к ней протоны. Спектры любых корреляций, кроме прямых, могут содержать сложные многоатомные подписи, в том числе двухцветные, если корреляция связывает атомы из разных остатков (напр., **H1/C3** в HMBC). В режиме оценки достоверности подписи содержат метрику достоверности в процентах.

Ссылка **Hi-res image** (15) или нажатие на спектр открывает отдельное окно с изображением спектра в полиграфическом разрешении для копирования в документы. Если какие-либо из сигналов не были предсказаны (т.е. имеют знак вопроса вместо химического сдвига в таблице отнесения), они перечисляются рядом со спектром, в котором они должны были присутствовать. Ссылка **JDX** (16) экспортирует спектр в формате Jcamp-DX [275] для дальнейшей работы в специализированных программах обработки спектров ЯМР (MestreLabs MestreNova [276] [a], ACD/Labs NMR viewer[b], Bruker TopSpin[c] и т.д.). При этом пользователь может либо сохранить файл (**Download file**), либо передать его на сайт ChemInfo[d] (**Live view NMR**) для простейшей интерактивной обработки и визуального сравнения с другими спектрами.

В данном примере, как эмпирический, так и статистический методы сообщили о низкой достоверности (60%) моделирования сигнала C1 β-6-дезокси-манногептопиранозы. Это связано в том числе с тем, что теоретические эффекты фосфорилирования моносахаридов изучены плохо, так как результирующий химический сдвиг зависит от кислотности среды, а статистических данных недостаточно для однозначного вывода – была использована всего одна запись в базе CSDB. Нажав на `1` (число использованных записей) в колонке **C1** ряда `bD6dmanHepp`, мы можем изучить, откуда появилось значение 101.6 м.д. и какие изменения структура претерпела в процессе обобщения, сделавшего воз-

[a] http://mestrelab.com/software/mnova/

[b] https://www.acdlabs.com/products/spectrus/workbooks/nmr/whatsnew.php

[c] https://www.bruker.com/products/mr/nmr/nmr-software/software/topspin/overview.html

[d] http://www.cheminfo.org/

можным использование этой записи. Список обобщений в порядке увеличения их веса для моделируемого атома C1 гептозы отображается при нажатии на ссылку **How?** около химического сдвига и объясняет низкую метрику достоверности тем, что за неимением записей, строго подходящих в смысле похожести структур, в процессе обобщений была удалена фосфатная группа.

*4.1.8. Установление характера связывания и топологии полимерного фукоглюкана с дисахаридным повторяющимся звеном на основании одномерного спектра ЯМР $^{13}C$.*

Решение этой задачи показывает, как обладая большей частью информации о строении гликополимера, узнать недостающие данные без «ручной» интерпретации спектров ЯМР. Допустим, хроматографией продуктов гидролиза полимера был установлен мономерный состав и полимер был идентифицирован как фукоглюкан с соотношением компонентов 1:1. Одномерный спектр ЯМР $^{13}C$ содержит 12 сигналов, что соответствует регулярному полимеру с дисахаридным повторяющимся звеном. Невыясненными остаются аномерные и абсолютные конфигурации, размеры циклов, позиции замещения, и топология повторяющегося звена. Для классического инструментального установления недостающих структурных параметров требуется набор двумерных спектров ЯМР и квалифицированный эксперт для их интерпретации. Мы сможем установить эти параметры в полуавтоматическом режиме, пользуясь лишь данными ГХ и одномерным спектром ЯМР $^{13}C$. Для этого понадобится выбрать наиболее подходящую структурную гипотезу из списка, генерируемого модулем предсказания структуры (пункт **Elucidation from NMR** в меню **Extras**). По умолчанию, интерфейс этого модуля показывается в развёрнутом режиме, но для данного примера, подразумевающего распространённую структуру, полный набор возможностей не требуется, и можно упростить ввод известных параметров, нажав на ссылку **Hide** в группе **Advanced options.** В результате мы получим вид окна, представленный на Рис. 75. Верхняя половина рисунка содержит введённые известные параметры структуры, а нижняя - результат ранжирования структурных гипотез, который появляется после завершения счета.

Каждый ряд соответствует данным об одном мономерном остатке. По умолчанию выбрана D-глюкопираноза (`D; glucose; pyranose`). Так как мы проводим поиск в режиме Widespread, подразумевающем перебор только распространённых остатков и их конфигураций, для глюкозы в любом случае будет учитываться только D-конфигурация и пиранозная форма. Поэтому можно оставить данные о мономере в первом ряду без изменений. Для того, чтобы задать дисахаридную структурную единицу, нажмём на ссылку **Add residue** (1) для до-

бавления второго остатка и в появившемся втором ряду выберем в колонке **Residue** остаток `fucose` из группы `deoxyhexoses` с помощью селектора (5). Колонки с аномерной (3; **α / β**) и абсолютной (4; **D / L**) конфигурациями и размером

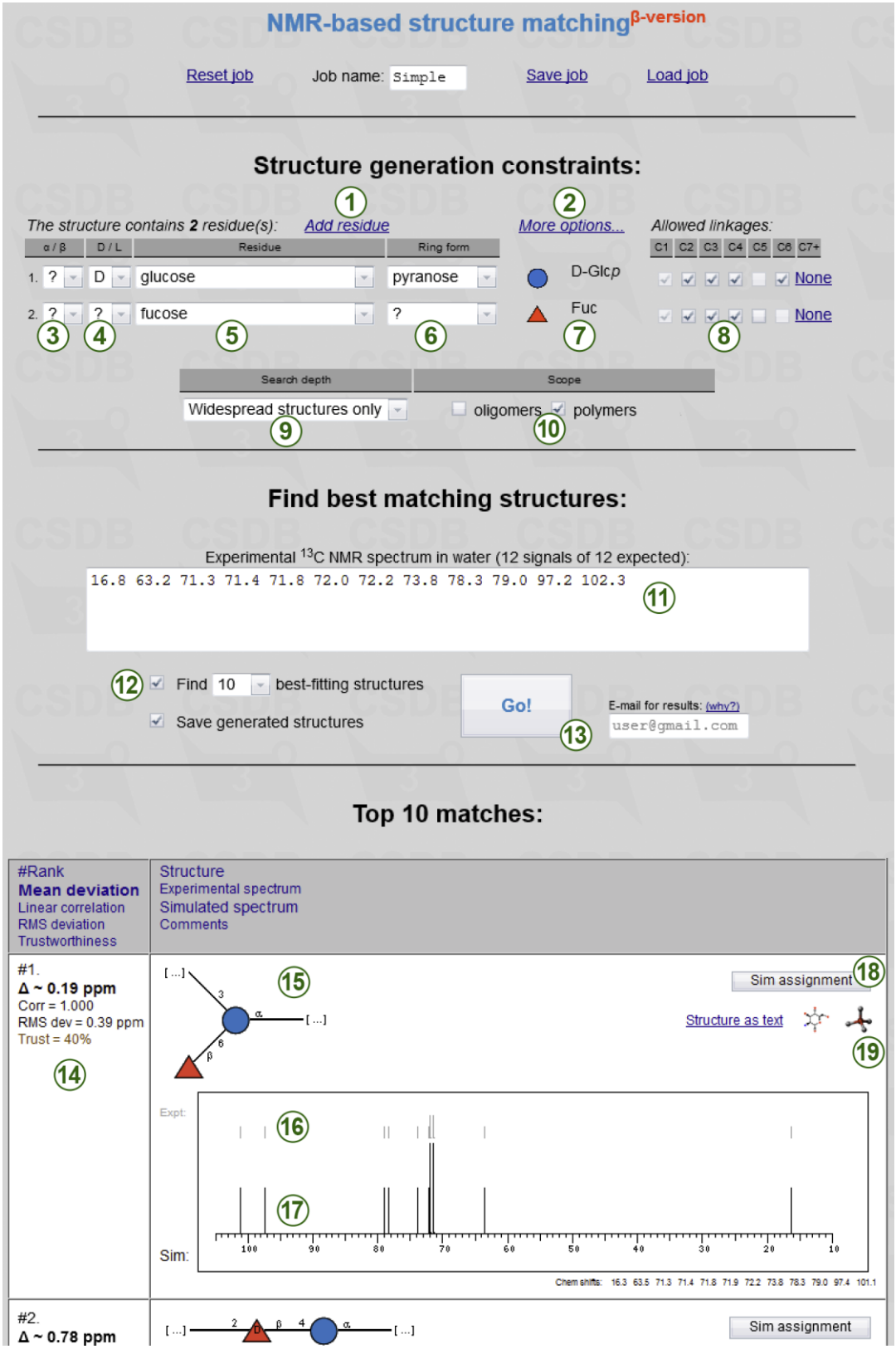



Рис. 75. Параметры поиска структурных гипотез в краткой форме (верхняя половина рисунка) и результаты ранжирования (нижняя половина, показаны частично).

цикла (6; **Ring form**) по умолчанию содержат значения «неизвестно» (знак вопроса) и не требуют модификации. Данные о выбранном мономерном составе предпросматриваются в области (7).

Блок отметок **Allowed linkages** (8) позволяет перечислить, какие позиции в каждом остатке могут быть замещены, а какие нет. По умолчанию отмечены все позиции, в которых образование эфирной, сложноэфирной или амидной связи химически возможно. Некоторые позиции (например, аномерные центры остатков в полимерах) заблокированы в отмеченном состоянии. Химически запрещённые положения (в данном примере - C6 в 6-дезоксигексозах и C5 в альдопиранозах) заблокированы в неотмеченном состоянии.

Ниже находится блок опций общего характера. В простом режиме он содержит только селектор глубины поиска (9; **Search depth**) и признак полимерности (10; **Scope**). В первом выберем `Widespread structures only`, чтобы ограничить поиск распространёнными структурами, а во втором снимем отметку oligomers, так как известно, что структура полимерная.

Поле ввода спектра (11) принимает химические сдвиги в произвольном порядке. Сигналы кратной интегральной интенсивности следует вводить несколько раз. По умолчанию толерантность к пропущенным и лишним сигналам составляет один сигнал. В заголовке поля (11) показано число сигналов, ожидаемое на основании заданного мономерного состава, и его соответствие числу сигналов во введённом спектре. Отметки (12) говорят о том, что мы ищем 10 наиболее подходящих структурных гипотез и сохраняем все генерируемые структуры для последующего анализа и контроля полноты перебора. Нажатие на кнопку **Go!** (13) запускает поиск и ранжирование гипотез.

В рассматриваемом случае структурная единица не велика, и с учётом глубины поиска только среди распространённых конфигураций остатков исчерпывающий набор химически разрешённых комбинаций неизвестных параметров составляет всего 240 структур. В частности, он не содержит фураноз, полиолов и сахаров в открытой форме, а также глюкозы в L-конфигурации. Аномерные конфигурации и абсолютная конфигурация фукозы могут принимать любые значения. Перебор и последующее уточнение спектров 240 структур занимает около 7 минут, их которых 99% времени приходится на уточнение спектров. В более

сложных задачах уточнению подлежат лишь 500 наилучших структурных гипотез. Время счета позволяет дождаться результатов прямо в браузере, не пользуясь отложенным уведомлением по электронной почте.

Результаты представляют собой таблицу **Top matches**, в которой каждый ряд соответствует одной структурной гипотезе. Ряды отсортированы в порядке соответствия между предсказанным и экспериментальным спектром. На Рис. 75 показан первый ряд (наилучшая гипотеза) и часть второго. Левая колонка (14) содержит несколько числовых характеристик гипотезы:

- номер гипотезы в списке (**#**);
- метрику соответствия (**Δ**), учитывающую отклонение сигналов от эксперимента, достоверность моделирования, наличие лишних или отсутствие существующих сигналов;
- коэффициент линейной корреляции между моделью и экспериментом (**Corr**);
- среднеквадратичное отклонение модели от эксперимента (**RMS dev**);
- оценку достоверности моделирования (**Trust**).

Правая колонка содержит описания гипотез и инструменты работы с ними: предсказанные структуры в формате SNFG (15), промоделированный спектр ЯМР $^{13}$C (16) и выровненный по той же шкале экспериментальный спектр (17), кнопку запуска моделирования всех одно- и двумерных спектров и таблиц отнесения для предсказанной структуры (18) и дополнительные инструменты визуализации (19): переключение в формат SweetDB, генерирование атомарной структурной формулы и запуск модуля 3D моделирования пространственной структуры для оценки протон-протонных контактов и последующей валидации гипотезы с помощью экспериментального ЯЭО.

В данном примере метрика соответствия лучшей гипотезы (1,3-α-глюкан с β-L-фукозой в боковой цепи, присоединённой по C6; Δ=0.19 м.д.) значительно отличается от таковой для второй гипотезы (линейный гликополимер с 1,2-β-D-фукозой и 1,4-α-D-глюкозой в основной цепи, Δ=0.78 м.д.). Это значит, что для выбора правильной гипотезы из списка не требуется дополнительных экспериментов или соображений о вероятности появления тех или иных структурных

признаков, т.е. гипотезу, предсказанную как лучшую можно считать однозначным ответом.

Таким образом на основании данных о моносахаридном составе (глюкоза и фукоза) гликополимера, его экспериментального спектра ЯМР $^{13}C$ и предположения об отсутствии редких структурных особенностей расчётным методом установлены:

- α-аномерная конфигурация глюкозы и β-аномерная конфигурация фукозы;
- L-конфигурация фукозы;
- позиции связывания: у глюкозы 1,3 – основная цепь, 6 – боковая цепь; фукоза терминальна (связь образована только в положении 1);
- топология и последовательность остатков (основная цепь – линейный гомоглюкан, боковая цепь - остаток фукозы, точка разветвления на каждом остатке глюкозы).

*4.1.9. Предсказание структуры неустановленного олигомера, содержащего бациллозамин, лизин и глюкуроновую кислоту, на основании данных ЯМР.*

В этом примере возможности установления структуры по известным данным простых экспериментов рассмотрены более подробно. Здесь мы предполагаем, что в наличии имеются данные хроматографии продуктов гидролиза (мономерный состав), представления о молекулярном весе (признак олигомерности), экспериментальный спектр ЯМР $^{13}C$, информация об отсутствии фураноз (нет характеристичных сигналов в области 85-95 м.д.) и некоторые общие представления о биосинтезе гликанов. Для полного установления структуры требуется определить недостающие аномерные и абсолютные конфигурации, позиции замещения и последовательность остатков.

Так как подбор структурных ограничений и сам расчёт может, в зависимости от задачи, занять значительное время, модуль работы со структурными гипотезами позволяет сохранить и загрузить как задание, так и задание вместе с результатами работы, если они уже получены. Для этого в верхней части окна предусмотрены инструменты работы с заданиями (Рис. 76). **Reset job** (1) возвращает интерфейс к исходному состоянию. Поле **Job name** (2) позволяет отфильтровать задания по началу имени файла.

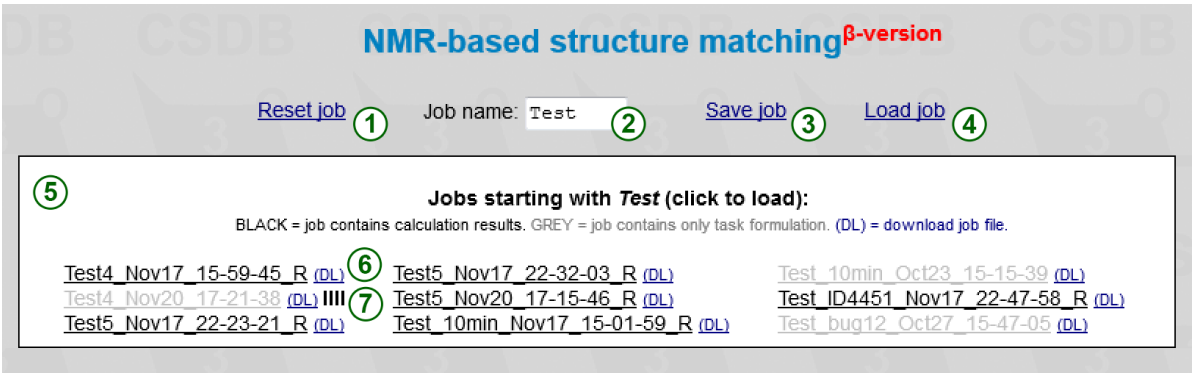



Рис. 76. Инструменты работы с заданиями GRASS.

**Save job** (3) сохраняет текущее задание (и результат расчёта при его наличии) в файл на сервере. Имя файла формируется из префикса, указанного в поле (2) и текущих даты и времени. Каждый раз при запуске расчёта используется новый штамп времени и файл пересохраняется автоматически. Если сохранить задание во время расчёта в пользовательской сессии, сессия закрывается и интер-

фейс переключается на работу с копией задания с новым штампом времени. Тем не менее, запущенный расчёт продолжается на сервере в фоновом режиме и его результаты можно в будущем загрузить с помощью ссылки **Load job** (4).

**Load job** (4) показывает все недавние задания (5), имена которых начинаются с префикса, указанного в поле (2), в данном примере используется префикс `test`. Задания хранятся на сервере в течение месяца, поэтому для повторного использования рекомендуется скачивать их и сохранять у себя на компьютере. Чтобы загрузить задание, достаточно нажать на его имя в списке (5). Чёрным показаны файлы, содержащие задание и его результат, серым – содержащие только задание. Для скачивания файла предназначена ссылка **(DL)** (6) напротив его имени. Напротив заданий, которые выполняются в данный момент, присутствует анимированный индикатор выполнения (7). Если загрузить такое активное задание, его статус начнёт обновляться в пользовательской сессии и результаты будут отображены в браузере после завершения счёта.

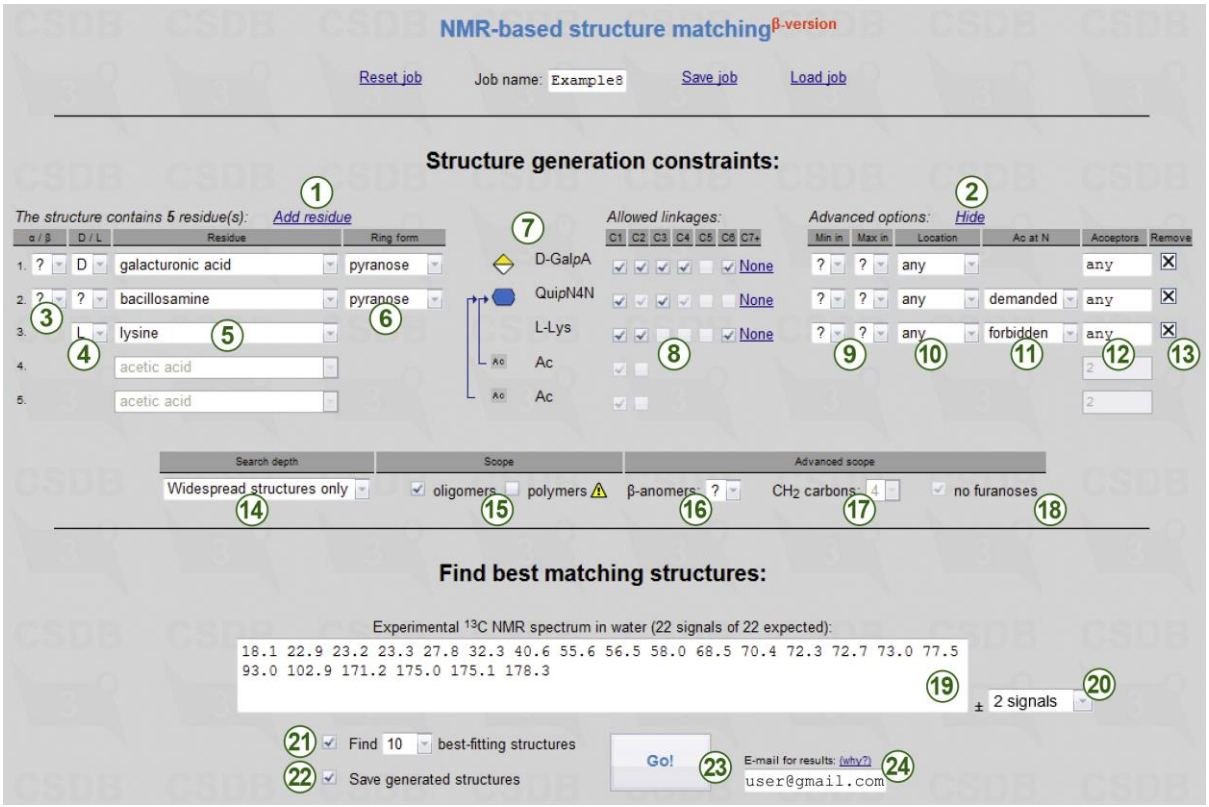



Рис. 77. Параметры поиска структурных гипотез в полной форме.

Ограничения на структуру (**Structure generation constraints**) определяют, сколько гипотез будет перебираться и сравниваться с экспериментальными дан-

ными. Форма модуля предсказания структуры, заполненная входными данными, соответствующими задаче, показана на Рис. 77.

В большинстве случаев ввод противоречивых данных заблокирован. Некоторые комбинации введённых параметров могут вызывать подозрения в некорректности, однако полностью исключать их нельзя. Такие ситуации отслеживаются модулем ввода и в соответствующих местах пользователю показывается значок предупреждения (жёлтый треугольник) с описанием проблемы.

Как и в предыдущем примере, воспользуемся ссылкой **Add residue** (1) для добавления остатков. В данном примере пять рядов соответствуют пяти остаткам, включая уксусную кислоту, ацетилирующую аминогруппы бациллозамина. Пиктограммы креста (13) предназначены для удаления остатков из состава.

Селекторы типа остатка (5) позволяют задать конкретный остаток или его класс. Список содержит несколько десятков наиболее распространённых остатков, и если нужный остаток отсутствует в списке, необходимо воспользоваться самым первым вариантом – **Show all residues.** В зависимости от выбранного остатка могут быть показаны дополнительные опции: аномерной конфигурации (3) для моносахаридов в циклической форме, абсолютной конфигурации (4) для оптически активных остатков, способа циклизации (6) для всех моносахаридов. Каждая из этих конфигураций может остаться неопределённой (знак вопроса). Если известно, что в структуре присутствует несколько одинаковых остатков в одинаковых конфигурациях, эти остатки должны быть указаны несколько раз. Область предпросмотра (7) подытоживает заданные ограничения на мономерный состав (пиктограммы SNFG), конфигурации мономеров (текстовые имена) и частичную последовательность (стрелки; если она известна).

Отметки в блоке **Allowed linkages** (8) разрешают или запрещают замещение определённых положений каждого остатка при генерировании структур. В соответствии с другими выбранными ограничениями (типы и положение остатков, ацетилирование, частичные последовательности) какие-то из этих отметок могут быть заблокированы в том или ином состоянии. Отмеченные положения будут связываться с другими остатками, если связь может быть образована с отщеплением воды. Для высших сахаров, колонка **C7+** соответствует любому из положений хвостовой части остатка, начиная с C7. Отметка, соответствующая исходя-

щей связи (чаще всего **C1** в альдозах и неуглеводных остатках и **C2** в кетозах) должна быть отмечена, если только вы не хотите расположить остаток строго на восстанавливающем конце.

Селекторы **Min** и **Max** (9) определяют минимальное и максимальное количество заместителей (степень разветвлённости) остатка, кроме акцептора гликозидной связи. Значение по умолчанию `any` означает отсутствие ограничений, значение «–» (прочерк) зарезервировано для остатков, которые могут находиться только в терминальном положении (напр., в связи с отсутствием замещаемых функциональных групп). Селектор **Location** (10) ограничивает возможное расположение остатка в структуре и может принимать следующие значения: `any` (любое), `terminal` (терминальное), `reducing` (на восстанавливающем конце), и `not reducing` (любое, кроме восстанавливающего).

Селектор **N-acetylation** (11) определяет, должны ли аминогруппы быть ацетилированными (`demanded`) или свободными (`forbidden`) или допустимы оба варианта (`allowed`). Поскольку для бациллозамина был выбран вариант `demanded`, два остатка уксусной кислоты были добавлены автоматически, и отметки в положениях, несущих аминогруппы (C2 и C4), были заблокированы как отмеченные (8). Явное указание акцепторов в поле **Acceptors** (12) позволяет задать частичную последовательность, если структурные фрагменты известны из экспериментов по химической деградации структуры и анализу продуктов. Числа, которые можно выбрать, нажав на это поле – это порядковые номера, располагающиеся в начале рядов. Так, бациллозамин имеет порядковый номер #2 в мономерном составе, и остатки уксусной кислоты жёстко связаны с ним, как с акцептором. Такие частичные последовательности визуализируются в области предпросмотра (7) синими стрелками. Для описания не полностью определённых последовательностей акцепторы могут включать более одного остатка (в значении «один из»), в том числе самого себя, если основная цепь повторяющегося звена полимера состоит из единственного остатка.

Хотя абсолютные конфигурации остатков не известны, мы предполагаем, что они соответствуют распространённым в биогликанах энантиомерам. Это предположение фиксируется выбором варианта `Widespread` в селекторе **Search depth** (14), который также исключит из рассмотрения редкие остатки (если часть

мономерного состава представлена классами) и размеры циклов, сложноэфирные и амидные связи между моносахаридами, сильноразветвлённые структуры и громоздкие боковые цепи в полимерах. В большинстве случаев молекулярный вес соединения приблизительно известен, поэтому установим олигомерность (15, **Scope**) отметкой `oligomers`.

Прочие ограничения уровня структурной единицы в целом, выясняемые на основании одномерных спектров ЯМР, включают общее число β-аномеров (16, **β-anomers**), общее число $CH_2$-групп (17, **$CH_2$ carbons**) и признак отсутствия фураноз (18, **no furanoses**). Если выбранный мономерный состав и конфигурации не подразумевают вариативности этих параметров, селекторы заблокированы в единственно возможном состоянии (напр., в данном примере любая возможная структура содержит строго 4 вторичных углеродных атома).

Экспериментальные химические сдвиги ЯМР $^{13}C$ следует ввести в поле (19), дублируя сигналы двойной интегральной интенсивности. Достаточная точность – 0.1 м.д., порядок произвольный. Если число сигналов не соответствует диапазону, следующему из выбранного мономерного состава (или единственному значению, если состав не содержит классов остатков, которые могут иметь разный размер углеродного скелета), оно подсвечивается в заголовке поля (19). Максимально допустимое отклонение размера спектра указано в селекторе (20); следует помнить, что при несовпадении числа сигналов с ожидаемым точность предсказания уменьшается (см. раздел 3.4.2).

Воспользуемся размером рейтинга по умолчанию (21, **Find N best fitting structures**), но сохраним все сгенерированные структуры для последующего анализа (22, **Save generated structures**). Кроме браузера, результаты будут сохранены на сервере и ссылка для их просмотра будет отправлена по указанному электронному адресу (25, **E-mail for results**). Это может быть полезно, если время счета превысит время пользовательской сессии браузера.

Нажатие на кнопку **Go!** (25) запускает перебор структур в рамках заданных ограничений и черновое сравнение их эмпирических спектральных моделей с экспериментальными данными (первая фаза). Во время второй фазы, лучшие результаты моделируются статистическим методом. Статус процесс счета непрерывно обновляется в браузере (Рис. 78). Поле (1) знакомит с информацией

о текущем шаге, а поле (2) - с общим числом обработанных структур. Если требуется закрыть браузер, имеет смысл сохранить ссылку на будущие результаты (4). Когда результаты готовы, они могут быть загружены по этой ссылке, по ссылке из электронного письма или с помощью инструмента **Load job** (см. выше) с использованием имени задания и штампа времени. Если пользовательская сессия ещё не истекла, результаты загружаются в браузер автоматически.

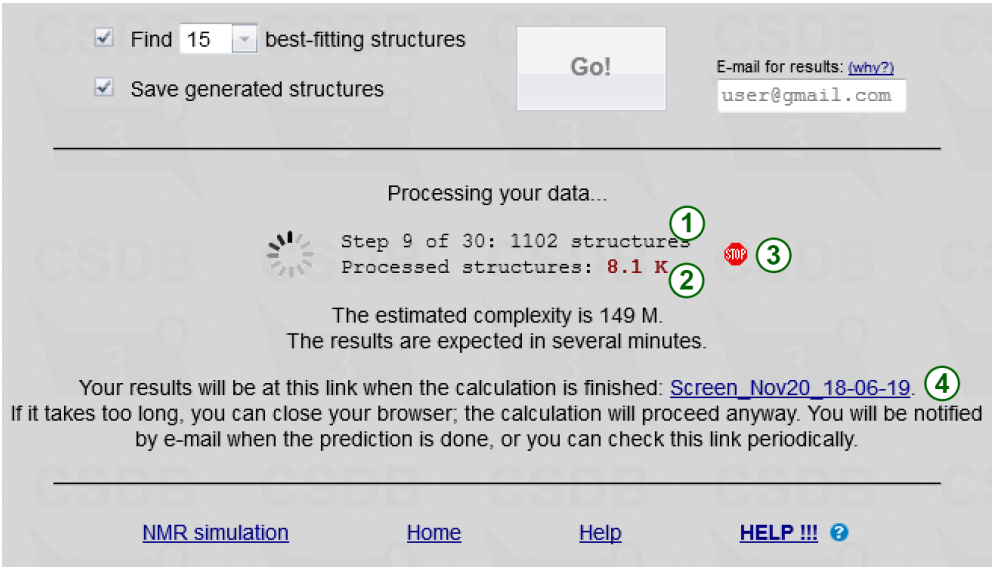



Рис. 78. Контроль процесса счета GRASS.

В зависимости от ограничений и природы остатков, выработка гипотез может занять значительное время. В среднем обработка 5-6 полностью определённых моносахаридов или 2-3 моносахаридов с неопределённостями (классы, неизвестные конфигурации) занимает от получаса до двух часов. Приблизительная оценка времени счета показана под статусом процесса. В данном примере расчёт занял около двух минут. Столь малое время объясняется множеством структурных ограничений и особенностями остатков, сужающими область поиска. В частности, бациллозамин, две аминогруппы которого уже заняты ацетильными группами, имеет лишь два замещаемых положения – C1 и C3, а лизин – три, причём два из них (C2 и C6) могут образовывать только амидные связи. В условиях наличия в остатках этой задачи всего двух карбоксильных групп это приводит к комбинаторному уменьшению числа вариантов.

Если во время расчёта в задании обнаружилась ошибка или расчёт стал неактуальным по другой причине, его можно прервать, нажав на кнопку **Stop** (3). Это также может понадобиться, чтобы не превышать разрешённое сервером ко-

личество расчётов на одного пользователя (не более двух процессов одновременно).

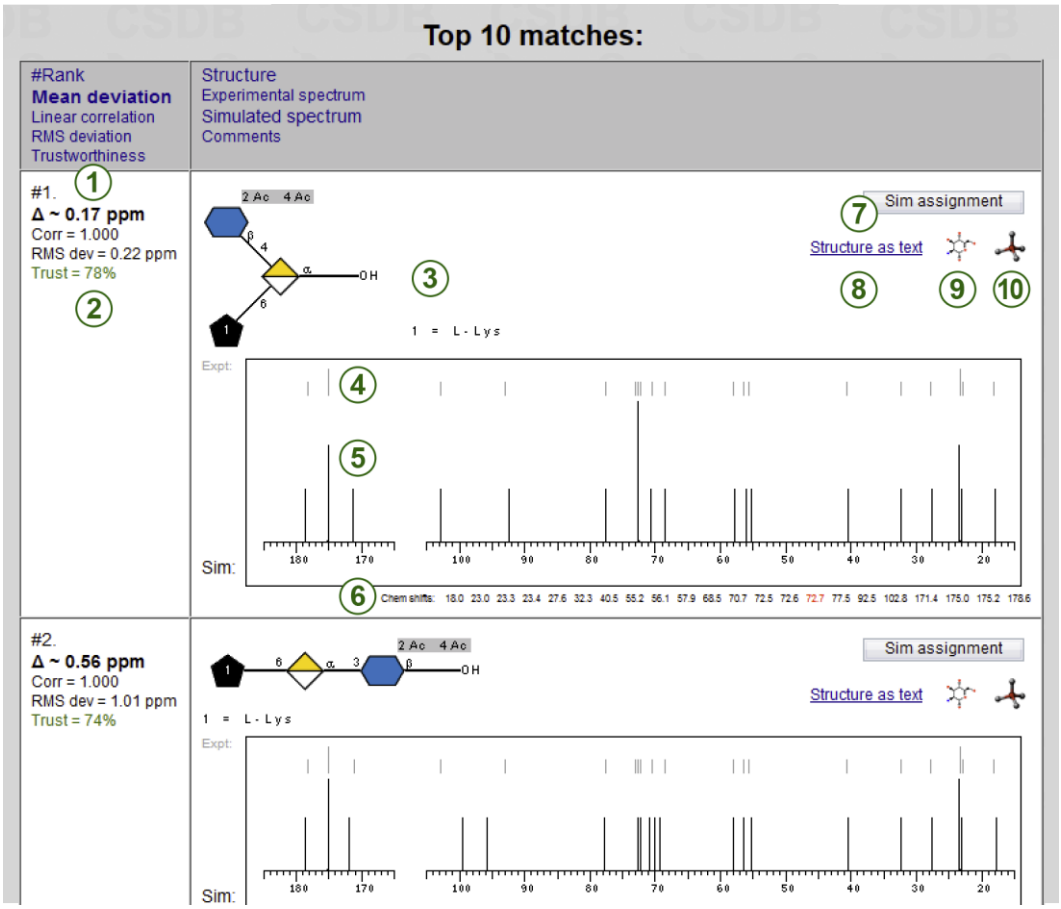



Рис. 79. Результаты ранжирования структурных гипотез (показаны две наилучших гипотезы).

После завершения расчёта, результат отображается в форме таблицы структурных гипотез (Рис. 79). Лучшие гипотезы показаны в начале таблицы. Как и в предыдущем примере, левая колонка содержит номер гипотезы в рейтинге (**#**), метрику соответствия модели эксперименту (**Δ**) (1), коэффициент линейной корреляции, среднеквадратичное отклонение (**ppm**) и оценку достоверности модели (**%**), отображённую в цвете от красного (0%) до зелёного (100%).

Правая колонка отображает структуры в формате SNFG (3) или SweetDB, промоделированные (5, чёрным цветом) и экспериментальные (4, серым цветом) спектры ЯМР $^{13}C$, кнопку перехода к отнесению и моделям остальных спектров (7) и дополнительные инструменты (8-10), подробно рассмотренные в разделе 2.1.8. В списках химических сдвигов (6) отдельных гипотез, некоторые значения подсвечены красным. Это сигналы, которые не могли быть промоделированы

статистически с достаточной точностью в режиме `Accurate`, и вместо этого использованы значения, полученные гибридным методом (см. разд. 3.4.1).

Две наилучших структурных гипотезы, показанные на Рис. 79, имеют заметные отличия в спектрах, особенно в регионе аномерных атомов углерода, что обеспечивает статистически значимую разницу между метрикой соответствия первой и второй гипотезы, не требующую дополнительного подтверждения. Тем не менее, если выбрать полную глубину поиска (**All structures** в селекторе **Search Depth**), результат пополняется гипотезами, занимающими промежуточное положение между двумя показанными на рисунке. В этом случае разница между метриками соответствия соседних гипотез составляет менее 0.2 м.д. и не может служить достаточным основанием для доказательства строения. В этом случае для выбора верной гипотезы потребуются данные эксперимента по метилированию для хотя бы одного остатка.

*4.1.10. Изучение состава гликанов двух видов аспергилл (A. oryzae и A. fumigatus) с особым вниманием к моносахаридам на концах боковых цепей.*

Углеводные эпитопы, локализованные на концах боковых цепей и на невосстанавливающем конце основной цепи гликанов микроорганизмов часто ассоциируют с формированием иммунного ответа в высших организмах. Терминальные части углеводных структур патогенного для человека грибка *Aspergillus fumigatus* – это потенциальные кандидаты для проверки влияния их присутствия на иммунноспецифичность штаммов. Инструмент анализа распределения малых фрагментов по организмам, доступный по ссылке **Fragment abundance** в меню **Extras**, представлен на Рис. 80.

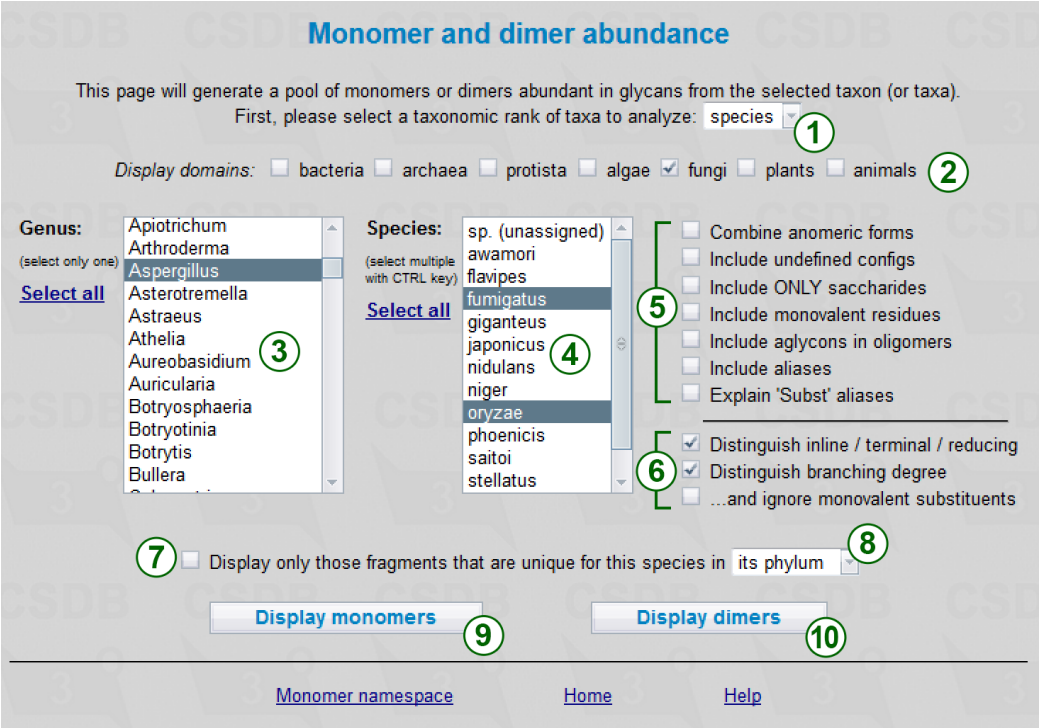



Рис. 80. Запрос на анализ структурных фрагментов в гликомах грибных видов.

Для начала работы выберем таксономический ранг, распределение структур по которому мы будем анализировать. В данном случае это вид (1, `Species`). Селекторы конкретных таксонов выглядят по-разному в зависимости от выбранного ранга но всегда позволяют выбрать один или больше таксонов. Прочие доступные ранги включают царство, тип, класс, род и подвид/штамм. Для упрощения навигации в списке родов (3, **Genus**, первый шаг для выбора видов) отграничим его только представителями царства грибов, поставив единственную отметку `fungi` в блоке **Display domains** (2). В селекторе **Genus** (3), пере-

числяющем все грибные рода, присутствующие в CSDB, выберем род `Aspergillus`. Для навигации по списку можно начать быстро набирать имя рода на клавиатуре. Когда род выбран, список видов (4; **Species**) пересчитывается. Он содержит только те виды выбранного рода, углеводные структуры которых присутствуют в базе данных. Выберем два интересующих нас вида - *A. fumigatus* и *A. oryzae*, используя клавишу Ctrl для множественного выбора.

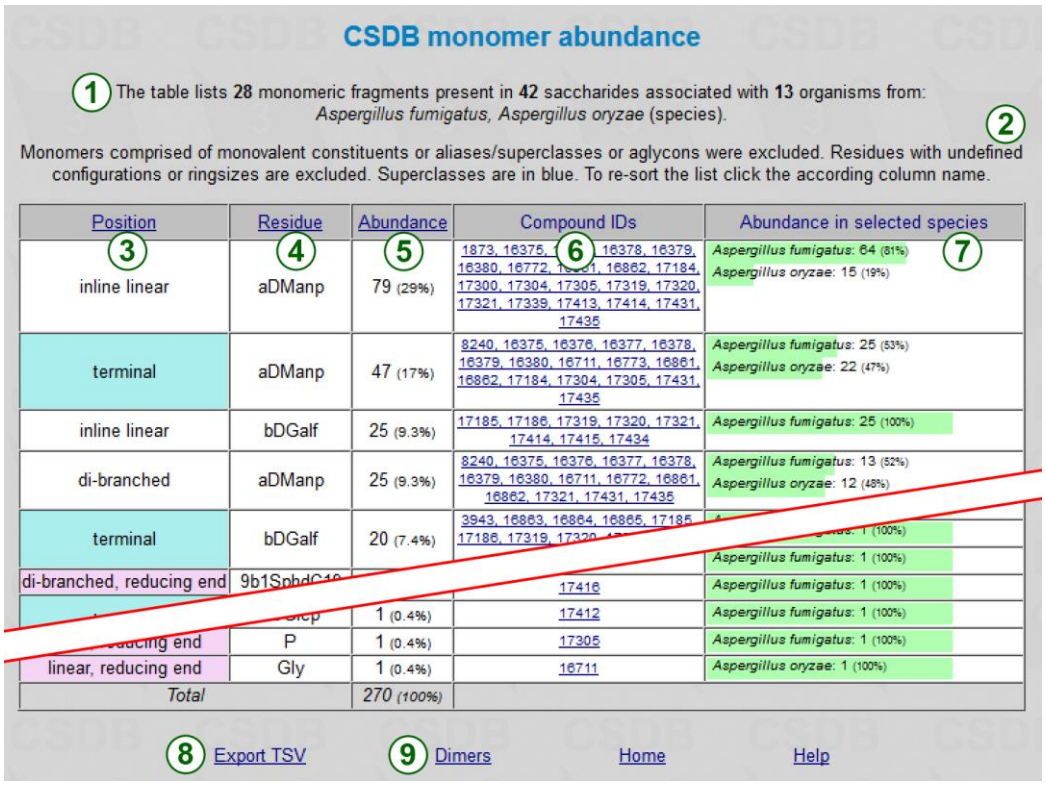
Группа отметок (5) определяет опции поиска на уровне входящих во фрагмент моносахаридов. Отметка **Combine anomeric forms** означает трактовку различных аномерных форм как одного и того же остатка. Отметка **Include undefined configs** включит в рассмотрение не только полностью определённые остатки, но и остатки с неизвестной аномерной или абсолютной конфигурацией или размером цикла. Для упрощения первоначального анализа эти отметки сняты, как и включение агликонов (**Include aglycons**), моновалентных остатков (метанол, уксусная кислота и т.д.) (**Include monovalent**), и остатков-заменителей, предназначенных для описания редких фрагментов, не поддерживаемых нотацией CSDB Linear (**Include aliases**).

Группа отметок (6) задаёт опции, связанные со структурным окружением фрагментов: разделять ли фрагменты в зависимости от их локализации в структурной единице (**Distinguish inline / terminal / reducing**) и от разветвлённости узла, в котором находится фрагмент (**Distinguish branching degree**). При определении числа заместителей для фрагментов в точках разветвления можно учитывать или не учитывать моновалентные модификации моносахаридов, что определяется состоянием отметки **And ignore monovalent substituents.**

Для выявления структурных фрагментов, характеристичных для таксонов в пределах большей биологической группы, предназначена отметка (7) и селектор (8), рассмотренные в следующем примере. Кнопки **Monomers** (9) и **Dimers** (10) запускают анализ для мономерных и димерных фрагментов, соответственно. Таблица распределения мономеров в структурах, синтезируемых организмами выбранных видов аспергилл показана на Рис. 81. Заголовок (1) показывает число найденных фрагментов, соответствующих им структур и организмов, ассоциированных в базе с этими структурами. Заданные ограничения на тип и положение фрагментов, перечислены в блоке (2).

Результаты представлены в виде таблицы со следующими колонками:

- положение фрагмента в структуре (терминальный – на голубом фоне, восстанавливающий – на розовом фоне, линейный, точка разветвления и её тип и т.д.), поскольку было заказано их дифференцировать (3, **Position**);
- структура фрагмента – в данном случае фрагменты являются мономерами, и эта часть включает единственную колонку **Residue** (4) с названиями и конфигурациями остатков;
- относительная распространённость **Abundance** (5) – сколько раз этот фрагмент встречается в структурах, удовлетворяющих запросу;
- ссылки на соединения (6, **Compound IDs**), содержащие данный фрагмент, пользуясь которыми можно перейти к остальным данным;
- распределение фрагмента по выбранным таксонам (в данном случае – по двум видам аспергилл) в виде абсолютных и относительных значений и гистограмм (7, **Abundance in selected taxa**).



**CSDB monomer abundance**

(1) The table lists **28** monomeric fragments present in **42** saccharides associated with **13** organisms from: *Aspergillus fumigatus, Aspergillus oryzae* (species).

(2) Monomers comprised of monovalent constituents or aliases/superclasses or aglycons were excluded. Residues with undefined configurations or ringsizes are excluded. Superclasses are in blue. To re-sort the list click the according column name.

| Position | Residue | Abundance | Compound IDs | Abundance in selected species |
|---|---|---|---|---|
| (3) inline linear | (4) aDManp | (5) 79 (29%) | (6) 1873, 16375, [illegible] 16378, 16379, 16380, 16772, [illegible] 16862, 17184, 17300, 17304, 17305, 17319, 17320, 17321, 17339, 17413, 17414, 17431, 17435 | (7) *Aspergillus fumigatus*: 64 (61%) *Aspergillus oryzae*: 15 (19%) |
| terminal | aDManp | 47 (17%) | 8240, 16375, 16376, 16377, 16378, 16379, 16380, 16711, 16773, 16861, 16862, 17184, 17304, 17305, 17431, 17435 | *Aspergillus fumigatus*: 25 (53%) *Aspergillus oryzae*: 22 (47%) |
| inline linear | bDGalf | 25 (9.3%) | 17185, 17186, 17319, 17320, 17321, 17414, 17415, 17434 | *Aspergillus fumigatus*: 25 (100%) |
| di-branched | aDManp | 25 (9.3%) | 8240, 16375, 16376, 16377, 16378, 16379, 16380, 16711, 16772, 16861, 16862, 17321, 17431, 17435 | *Aspergillus fumigatus*: 13 (52%) *Aspergillus oryzae*: 12 (48%) |
| terminal | bDGalf | 20 (7.4%) | 3943, 16863, 16864, 16865, 17185, 17186, 17319, 17320, [illegible] | [illegible] 1 (100%) |
| di-branched, reducing end | 9b1Sphd[illegible] | [illegible] | [illegible] | *Aspergillus fumigatus*: 1 (100%) |
| [illegible] | [illegible] | 1 (0.4%) | 17416 | *Aspergillus fumigatus*: 1 (100%) |
| [illegible] | [illegible] | 1 (0.4%) | 17412 | *Aspergillus fumigatus*: 1 (100%) |
| [illegible] reducing end | P | 1 (0.4%) | 17305 | *Aspergillus fumigatus*: 1 (100%) |
| linear, reducing end | Gly | 1 (0.4%) | 16711 | *Aspergillus oryzae*: 1 (100%) |
| *Total* | | *270 (100%)* | | |

(8) Export TSV (9) Dimers Home Help

Рис. 81. Распределение мономеров и их положения в структурах гликанов двух изучаемых видов.

После последнего ряда с кумулятивными значениями находится ссылка **Export TSV** (8) для экспорта данных в табличный процессор (напр., Microsoft

Excel) и ссылка **Dimers** (9) для перехода к другому типу фрагментов. Сортировку рядов в таблице можно изменить, нажав на заголовок соответствующей колонки (3, 4 или 5).

В соответствии с полученными результатами база CSDB содержит 42 сахарида, продуцируемых аспергиллами видов *A. fumigatus* и *A. oryzae*, и эти соединения построены из 28 мономерных «строительных блоков». α-D-маннопираноза является среди них самой распространённой. Относительно редко встречающийся в других доменах остаток β-D-галактофуранозы присутствует на невосстанавливающем конце 20 структур *A. fumigatus*, что приходится примерно на половину всех вхождений этого остатка в углеводах проанализированных видов аспергилл. Это выделяет β-галактофуранозу как потенциальный эпитоп иммуноспецифичности грибка *A. fumigatus* и позволяет спланировать последовательность экспериментов по удалению фрагментов структуры с последующей проверкой антигенной активности методами иммуноферментного анализа на антителах, выработанных к указанному грибку.

*4.1.11. Выявление димерных фрагментов (включая сахара и агликоны) гликанов высших растений, уникальных для рода люпинов.*

Эта задача направлена на выявления уникальных особенностей рода люпинов с точки зрения биосинтеза гликанов. Её решение позволяет предсказать специфические люпиновые гликозилтрансферазы для последующего поиска в протеомных базах данных. Для подобного анализа используется тот же инструмент, что и в предыдущем примере, но с другим набором опций (Рис. 82).

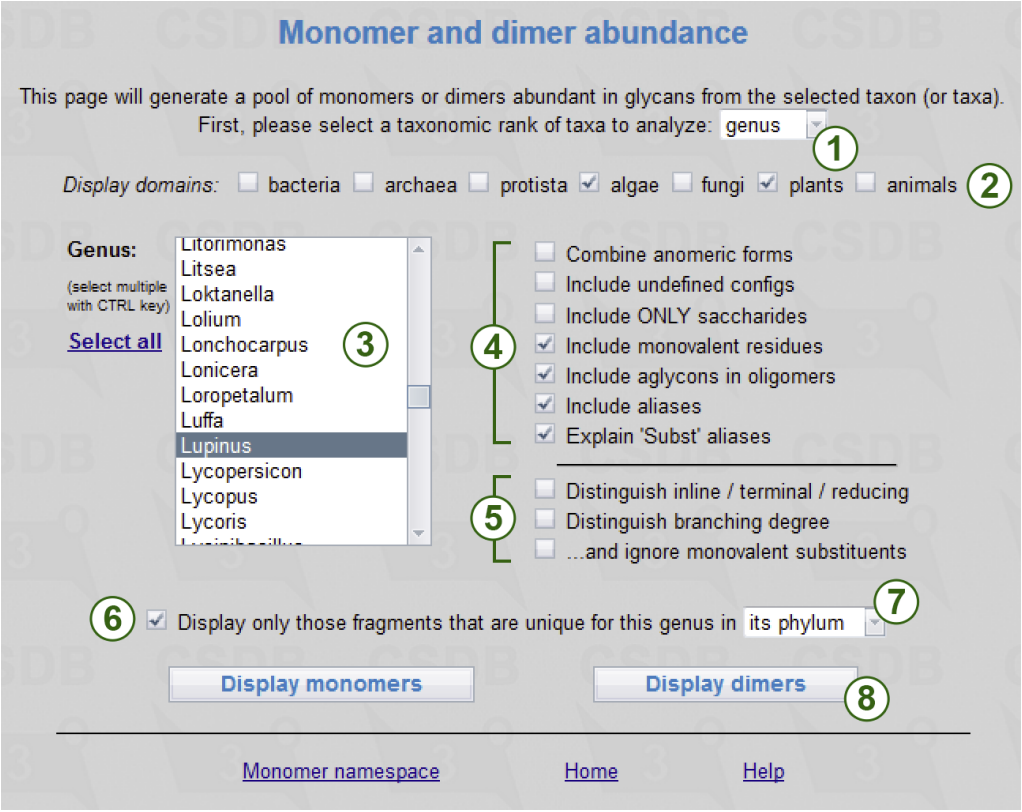
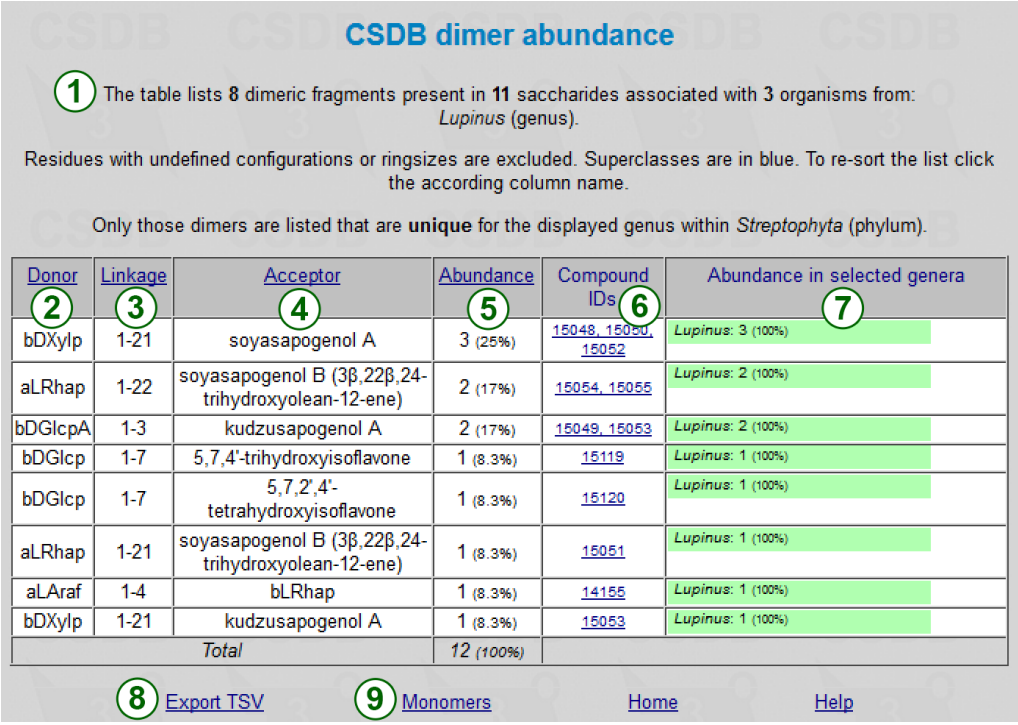



Рис. 82. Запрос на анализ структурных фрагментов в гликомах растительных и грибных родов.

Здесь в качестве таксономического ранга (1) выберем род (`genus`), чтобы охватить все принадлежащие ему виды, включая случаи, когда структура ассоциирована с люпинами неизвестного вида. Домены, для которых будет представлен список родов, ограничены растениями (**algae + plants**, 2). В селекторе родов (3, **Genus**) списке родов выберем `Lupinus` (люпины).

Растительные гликозиды часто содержат единственный моносахарид, присоединённый к остатку агликона. Многие агликоны не поддерживаются нотацией CSDB Linear, тем не менее они могут быть ключевыми факторами специфичности биогликанов. Поэтому включим в анализ не только дисахариды, но

и димеры, построенные из моносахаридной и неуглеводной частей (агликоны и остатки-заменители), поставив соответствующие отметки в блоке (4). В случае, если отметка **Explain 'Subst' aliases** не поставлена, компоненты структуры, для которых не зарезервировано название в контролируемом словаре мономеров будут обработаны как одинаковые псевдо-остатки и показаны как остаток `Subst`. Так как разнообразие агликонов является важной частью разнообразия растительных димеров, в данном примере эта отметка поставлена. Отметка `Include undefined configs` снята, так как исследование разнообразия и распределения димеров – это первый шаг к выявлению гликозилтрансфераз, а гликозилтрансферазы связаны с однозначными конфигурациями доноров и субстратов и определяют конфигурации и связи в продуктах. Таким образом, статистический анализ не будет включать димеры с неизвестными конфигурациями моносахаридов или с неопределёнными положениями замещения. С точки зрения работы углевод-активных ферментов, положение субстрата в конечной структуре, которая может наращиваться далее другими ферментами, не принципиально, поэтому отметки в блоке (5) не поставлены.



**CSDB dimer abundance**

(1) The table lists **8** dimeric fragments present in **11** saccharides associated with **3** organisms from: *Lupinus* (genus).

Residues with undefined configurations or ringsizes are excluded. Superclasses are in blue. To re-sort the list click the according column name.

Only those dimers are listed that are **unique** for the displayed genus within *Streptophyta* (phylum).

| Donor (2) | Linkage (3) | Acceptor (4) | Abundance (5) | Compound IDs (6) | Abundance in selected genera (7) |
|---|---|---|---|---|---|
| bDXylp | 1-21 | soyasapogenol A | 3 (25%) | 15048, 15050, 15052 | *Lupinus*: 3 (100%) |
| aLRhap | 1-22 | soyasapogenol B (3β,22β,24-trihydroxyolean-12-ene) | 2 (17%) | 15054, 15055 | *Lupinus*: 2 (100%) |
| bDGlcpA | 1-3 | kudzusapogenol A | 2 (17%) | 15049, 15053 | *Lupinus*: 2 (100%) |
| bDGlcp | 1-7 | 5,7,4'-trihydroxyisoflavone | 1 (8.3%) | 15119 | *Lupinus*: 1 (100%) |
| bDGlcp | 1-7 | 5,7,2',4'-tetrahydroxyisoflavone | 1 (8.3%) | 15120 | *Lupinus*: 1 (100%) |
| aLRhap | 1-21 | soyasapogenol B (3β,22β,24-trihydroxyolean-12-ene) | 1 (8.3%) | 15051 | *Lupinus*: 1 (100%) |
| aLAraf | 1-4 | bLRhap | 1 (8.3%) | 14155 | *Lupinus*: 1 (100%) |
| bDXylp | 1-21 | kudzusapogenol A | 1 (8.3%) | 15053 | *Lupinus*: 1 (100%) |
| *Total* | | | *12 (100%)* | | |

(8) Export TSV (9) Monomers Home Help

Рис. 83. Распределение родоспецифичных димеров в гликанах рода люпинов.

Для нахождения уникальных фрагментов, т.е. присутствующих только в структурах, выделенных из указанного рода, но не из других растительных ро-

дов, установим отметку (6)и в селекторе (7) выберем таксономический тип (`its phylum`) в качестве более высокого ранга, уникальность в пределах которого нас интересует. Для рода люпинов это тип *Streptophyta* (высшие растения). Кроме типа селектор позволяет выбрать царство (`its kingdom`) или все живые организмы (`all biota`; для поиска структурных компонентов, уникальных во всей природе).

Нажатие на кнопку Dimers (8) запускает анализ и приводит к результатам по распределению димеров (Рис. 83). Кумулятивная информация и условия выбора фрагментов представлены в верхней части страницы (1). В таблице с результатами отсутствует дифференцировка положения фрагмента в структуре, а описание фрагментов представлено тремя колонками: донор гликозидной связи (2, **Donor**), связываемые положения (3, **Linkage**) и акцептор гликозидной связи (4, **Acceptor**). Колонки распространённости (5), ссылок на соединения (6) и распределения по таксонам (7) аналогичны предыдущему примеру. В данном примере для анализа был выбран единственный таксон, поэтому распределение фрагментов по таксонам – всегда 100% в роду люпинов. Ссылка **Export TSV** (8) экспортирует данные для внешней обработки; ссылка **Monomers** (9) позволяет показать распределение мономеров для тех же условий задачи.

Результаты содержат 7 родоспецифичных димеров моносахаридов с неуглеводными компонентами, продуцируемых из всех высших растений только люпинами. Наиболее часто встречающаяся связь (из уникальных) – гликозилирование соясапогенола A в положение 21 остатком β-D-ксилопиранозы, который сам по себе характерен для растительных гликомов. В контексте структур, присутствующих в базе CSDB, уникальным для люпинов является единственный дисахарид: 4-O-α-L-арабинофуранозил-β-D-рамнопираноза. Он синтезируется люпинами в составе соединения 14155, рамногалактуронана с диарабинофуранозными боковыми цепями, выделенного в 1993 году из *Люпина узколистного* (*Lupinus angustifolius)*. Перейти к этому соединению в базе CSDB можно, нажав на ссылку `14155` в колонке **Compound IDs**.

*4.1.12. Получение статистических данных об изученности гликома протеобактерий.*

Для получения данных о полноте базы в пределах одной или нескольких таксономических групп предназначен инструмент статистической оценки покрытия (ссылка **Coverage Stats** в меню **Extras**, Рис. 84).

Здесь можно выбрать таксономический ранг (1) – в данном случае тип (`phylum`), после чего выбрать один или несколько таксонов данного ранга из появившегося селектора **Phylum** (3). Наполненность селектора вариантами ограничивается поставленными отметками в блоке **Display domains** (2). Выберем царство бактерий для упрощения навигации в списке типов и тип *Протеобактерии* (`Proteobacteria`). Отображаемые результаты можно отфильтровать по критерию даты публикации (4, **Publication year**, диапазон лет) и типа структуры (5, **Structure type**). Кнопка **Display coverage** (6) показывает количество записей разного типа для выбранного таксона или таксонов.

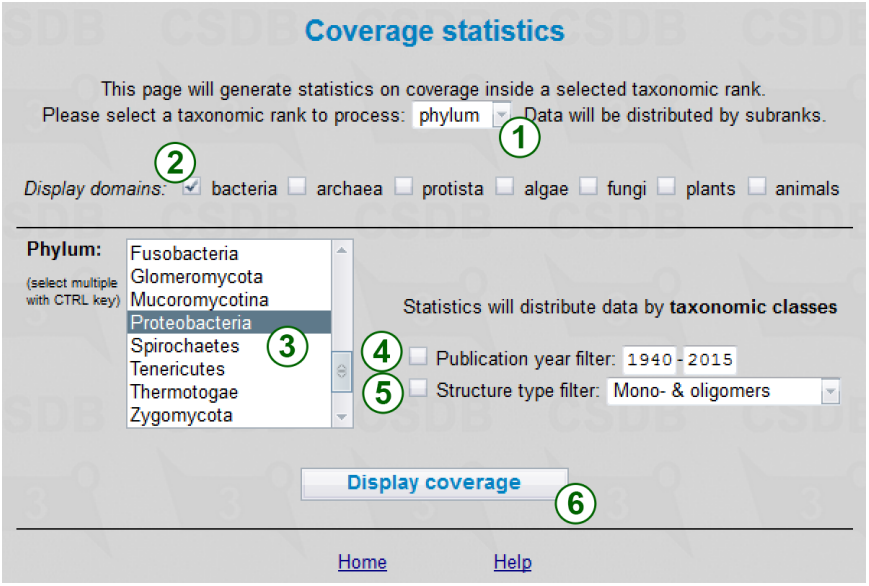



Рис. 84. Запрос на статистический анализ покрытия базы в пределах таксономического типа.

Результирующая таблица (Рис. 85) содержит следующие колонки:

- Выбранные таксоны (1). В данном примере был выбран единственный таксон Proteobacteria, поэтому колонка нерепрезентативна.
- Субтаксоны выбранных таксонов, присутствующие в базе (2). Для таксономического типа это класс. Заголовки первой и второй колонок (в дан-

ном случае – **Phylum** и **Class**) и характер классификации организмов в них зависят от изначально выбранного таксономического ранга.

- Покрытие по структурам (3, **Structures**) – число структур, ассоциированных с субтаксоном из второй колонки, и их доля в общем числе структур, характеризующих таксон.
- Покрытие по публикациям (3, **Publications**) – число публикаций, содержащих структуры из предыдущего пункта и доля публикаций по субтаксону в публикациях по всему таксону.
- Покрытие по организмам (3, **Organisms**) – число таксономически различных организмов в пределах субтаксона и их доля в таксоне.
- Покрытие по спектрам ЯМР (3, **NMR spectra**) – число спектров ЯМР, хранящихся в базе для структур из колонки **Structures**.

Кумулятивные значения, удовлетворяющие указанным таксонам и фильтрам, приведены в последнем ряду. В случае выбора нескольких таксонов, имеющих множество субтаксонов низшего ранга, таблица может быть громоздкой. Для упрощения навигации предусмотрена сортировка по содержимому любой колонки. Для этого следует нажать на заголовок колонки.

Протеобактерии – крупнейший тип прокариот, о чем свидетельствует несколько тысяч объектов каждого сорта, ассоциированных с протеобактериями. Распределение показывает, что наиболее изучен класс гамма-протеобактерий.

**CSDB coverage statistics**

The table lists number of entities distributed by classes and associated with organisms from: *Proteobacteria* (phylum).

To re-sort the list click the according column name. To show instances, click on numbers

| 1 Phylum | 2 Class | 3 Structures | 4 Publications | 5 Organisms | 6 NMR spectra |
|---|---|---|---|---|---|
| Proteobacteria | Gammaproteobacteria | 6764 (79%) | 2832 (80%) | 4200 (81%) | 2607 (81%) |
| Proteobacteria | Betaproteobacteria | 801 (9.3%) | 371 (11%) | 392 (7.6%) | 323 (10.0%) |
| Proteobacteria | Alphaproteobacteria | 753 (8.7%) | 369 (10%) | 347 (6.7%) | 303 (9.4%) |
| Proteobacteria | Epsilonproteobacteria | 586 (6.8%) | 157 (4.4%) | 221 (4.3%) | 181 (5.6%) |
| Proteobacteria | Deltaproteobacteria | 6 (0.1%) | 3 (0.1%) | 3 (0.1%) | 6 (0.2%) |
| Proteobacteria | *unresolved* | 2 (0.0%) | 2 (0.1%) | 2 (0.0%) | 4 (0.1%) |
| 7 | *Merged* | *8606 (100%)* | *3533 (100%)* | *5165 (100%)* | *3232 (100%)* |

Export TSV  Home  New query  Help

Рис. 85. Покрытие базы по протеобактериям и их классам.

### *4.2 Использование знаний, полученных из CSDB, в других исследованиях*

База данных CSDB используется как источник фактических экспериментальных данных и как платформа для создания собственных сервисов. Практически полное покрытие по бактериальным углеводным структурам и многочисленные инструменты позволяют использовать CSDB для решения различных химических и аналитических задач. С её помощью проводится идентификация и установление структур сложных природных углеводов [367-377], в том числе с привлечением конформационных расчётов молекулярно-механическими [378] и квантово-механическими [166] методами и со сравнением геометрических предсказаний с экспериментальными данными ЯМР.

Среди распространённых применений следует отметить поиск характерных эпитопов в углеводных структурах патогенных бактерий, проверку новизны структуры, симуляцию и отнесение спектров ЯМР, исследование влияния структурных параметров на спектральные [379-383]. В биохимических, молекулярно-биологических и медицинских исследованиях с помощью CSDB выявляют характерные мотивы и эпитопы в углеводных структурах микроорганизмов с целью изучения иммунного ответа организма-хозяина и обоснования использования таких углеводов в составе вакцин. Результаты статистического анализа характеристик и разнообразия бактериальных углеводов, представленных в CSDB, привлекаются в качестве обоснования исследований углеводных стимулов, посредством которых бактерии взаимодействуют с иммунной системой [384, 385], для отбора объектов для извлечения знаний из «больших данных» по микробным геномам [386], исследования распространения ферментов по таксонам [387], объяснения геологической концентрации углеводов бактериальным вкладом [388], исследования биосинтеза сахаридов [389], объяснения кросс-реакций на бактериальные гликоэпитопы [390], разработки методов моделирования геометрии гликополимеров [391] и других исследований. Например, в работе Ю.А. Книреля и коллег [381] с помощью базы данных CSDB был проведён поиск эпитопов системы групп крови ABH (αGalNAc(1-3)[αFuc(1-2)]βGal(1-3)βGlcNAc; -2)αFuc(1-2)βGal(1-3)βGalNAc(1-; -2)αGal(1-3)[αFuc(1-2)]βGal(1-3)αGalNAc(1-) в бактериальных углеводах с целью создания массива структур для скрининга

взаимодействия с человеческими галектинами 4, 8 и 9, которые могут участвовать в подавлении бактериальной инфекции.

Репозитории углеводных структур также востребованы в исследованиях биосинтетического аппарата, вовлечённого в их синтез и процессинг. Несмотря на огромное количество предсказанных ферментативных активностей, наличие соответствующих природных структур является необходимым условием доказательства работы этих ферментов *in vivo*. В статье Овчинниковой и коллег, посвящённой новому семейству бактериальных гликозилтрансфераз, переносящих остатки β-Kdo [392], база CSDB была использована для поиска углеводных структур, содержащих данные остатки, что позволило связать структурные данные с последовательностями генов биосинтетического аппарата и предсказать активность β-Kdo-гликозилтрансферазы, которая впоследствии была охарактеризована биохимическими методами. В работе, посвящённой фосфорилазе 3-O-α-D-глюкопиранозил-L-рамнозы из *Clostridium phytofermentans*, Т. Нихира и коллеги использовали CSDB для поиска данного субстрата в составе углеводных структур микроорганизмов [379].

База данных CSDB процитирована в научной литературе около 500 раз. Методологическое использование CSDB в гликохимических исследованиях, как правило, не цитируется, но фиксируется сервером в виде статистики по запросам пользователей. Нагрузка на веб-проект составляет около 600 уникальных посетителей ежемесячно, исключая роботов и пользователей, посетивших только одну страницу. В среднем пользователи проводят на сайте проекта по 5-10 минут, что соответствует 1-2 запросам.

### *4.3 Выявление ошибок в базах и публикациях*

В рамках работ над базой данных был проведён анализ возможностей улучшения качества данных, содержащихся в базах, а также всего накопленного и опубликованного массива информации в химии углеводов, прежде всего первичной структуры биогликанов. Наиболее эффективным средством оказалась разработанная автором программа поиска отклонений химических сдвигов в опубликованных спектрах ЯМР от ожидаемых на основании моделирования (см. раздел 3.4.1). Было обнаружено около 600 несовпадений, превышающих 6 м.д. для сигналов $^{13}$C или 1 м.д. для сигналов $^{1}$H. Детальный анализ этих случаев с учётом оригинальных публикаций и известных зависимостей спектров ЯМР от структуры выявил следующие причины несовпадений:

1. 42%. Ошибки аннотирования (исправлены на основании исходных публикаций).
2. 27%. Аномальный для данной структуры химический сдвиг может быть объяснён нестандартной геометрией молекулы, либо существующие статистические и эмпирические данные не позволяют промоделировать ЯМР-параметры атома в данном химическом окружении с требуемой точностью, т.е. существующих знаний в области корреляции структура-спектр недостаточно для однозначного доказательства наличия ошибки. Данные не корректировались.
3. 12%. Ошибки в структурах, взятых из предыдущих статей, в которых структура установлена неверно, при том, что существуют последующие статьи, в которых структура уточнена на основании новых данных или более тщательного анализа, в том числе при помощи базы CSDB. Исправлены с указанием ссылок на уточняющие публикации, пути миграции ошибок из статьи в статью отслежены и описаны в комментариях к ошибочным структурам.
4. 10%. Ошибки в оригинальных публикациях, связанные с неправильным отнесением спектров и/или неверно установленной структурой, исправление которых невозможно на основании анализа информации, опубликованной в этой и других статьях. Выявлено 57 случаев (Табл. 19), противоречащие структуре сигналы заменены в базе на «неизвестно», в записи

внесены комментарии о несоответствии спектров структуре. В отдельных случаях эти ошибки были исправлены на основании повторной интерпретации спектров членами коллектива CSDB или повторения структурного исследования другими группами.

5. 9%. Ошибки в статьях, связанные с некорректным оформлением, которые могут быть исправлены без повторной интерпретации спектров (например: на спектре сигнал подписан правильно, а в таблице указан неправильный; опечатка в отнесении спектра в случаях, когда правильное значение химического сдвига очевидно; неправильный перенос данных о структуре из оригинальной публикации в последующие статьи). Ошибки исправлены.

Табл. 19. Наиболее явные несоответствия спектров структурам в существующих публикациях, выявленные в автоматическом режиме и доказанные вручную.

| *Природный объект*[*] | *Противоречие*[**] | *Действие*[***] | *Запись CSDB* |
|---|---|---|---|
| *Acetobacter tropicalis* SKU1100 | t)-β-Galf C4 ~72 (обе молекулы, Табл. 3) | | 26126, 26451 |
| *Aneurinibacillus thermoaerophilus* DSM 10155 | -3/4)-β-ManpNAc C3 и C4 ~74 (вне зависимости от замещения по 4) и, возможно, t)-α-MurpNAc6P C6 72 | не изменено | 27325 |
| *Bacillus anthracis* | t)-β-Quip4NAcyl-2OMe C4 72.2 и α-L-Rhap-OH C4 56.4 (Табл. 2) | поменяны местами | 23632 |
| *Bacillus cereus* ATCC 10987 | -3,6)-α-GalpNAc C3 68.7 | | 22688 |
| *Bacteroides vulgatus* IMCJ1204 | -3)-β-Glcp C4 75.4 и, возможно, C3 80.3 | | 500 |
| *Burkholderia cenocepacia* ET-12 J2315 | -4,5)-α-Kdop C4 65.8, -4)-α-L-gro-D-manHepp C2 55.8, t)-α-L-gro-D-manHepp C4 71.2 | | 21453 |
| *Burkholderia multivorans* | -3,4)-α-L-gro-D-manHepp C3 70.5, t)-β-Glcp C3 72.3, C5 68.9 (Табл. S2) | | 22979 |

| | | | |
|---|---|---|---|
| *Burkholderia phytofirmans* PsJN | -3)-β-GalpNAc C4 64.0 | не изменено | 23418 |
| *Burkholderia vietnamiensis* LMG 10929 | t)-α-L-Fucp C3 79.0 (неверное отнесение либо остаток 3-замещен) | | 29444 |
| *Butyrivibrio fibrisolvens* H10b | S-Lac-(2-C1 64.8 (Табл. 4) | исправлено | 120, 121 |
| *Cronobacter malonaticus* 3267 | -2,6)-β-Galp C6 61.0 (Табл. 1) | | 24103 |
| *Cronobacter sakazakii* HPB 3290 | t)-β-GlcpA C5 70.3 (Табл. 3) | | 24098 |
| *Escherichia coli* O103:H2 | -1)-Gro C2 60.8 (Табл. 2) | | 25881 |
| *Escherichia coli* O157:H7 | t)-β-Glcp C4 75.4 и C5 69.2 (Табл. 1) | поменяны местами | 28342 |
| *Escherichia coli* O167 | -2,6)-α-L-4dthrHexp4enA C5 14.3 | | 3231 |
| *Escherichia coli* O41 | -3)-α-Galp C4 79.2 (Табл. 1) | | 27371 |
| *Escherichia coli* O8:K41:H11 | -3)-β-GalpNAc(1- C5 72.07 (Табл. 1) | | 3143 |
| *Escherichia coli* O86:K62:H2 WbnI-mutant | -3)-α-GalpNAc и -3)-β-GalpNAc | все сигналы остатков поменяны местами | 20328 |
| *Escherichia coli* WBB22 | -6)-β-GlcpN4P C6 62.5 | | 6285 |
| *Haemophilus influenzae* NTHi 1247 | t)-β-GalpNAc C5 67.2 (Табл. 6) | | 26993 |
| *Haemophilus parasuis* str. Nagasaki sv. 5; *Haemophilus parasuis* ER-6P sv. 15 | -5)-Kdop4P C6 81.1 | | 29724 |
| *Hafnia alvei* PCM 1194 | -3)-β-GalpN(1- C6 66.5 (Табл. 4) | | 4047 |
| *Hafnia alvei* PCM 1206 | -2)-β-Ribf C4 76.5 | | 4048 |

| | | | |
|---|---|---|---|
| *Lactobacillus delbrueckii* NCFB 2074 | t)-β-Galp C5 71.18,<br>-3,6)-α-Galp C2 76.6 | частично исправлено | 10051 |
| *Lactobacillus helveticus* sp. Rosyjski | t)-β-Glcp C3 70.67 (Табл. 1b) | | 29309 |
| *Lactobacillus rhamnosus* E/N | t)-α-Galp4,6Pyr C5 72.19 и C3 70.35 | | 11063 |
| *Lactobacillus rhamnosus* LOCK 0900 | -3)-α-Manp(1- C4 75.8 и C5 68.3 | поменяны местами | 15050 |
| *Lactococcus lactis* 3107 | -2)-β-Galf(1- C3 83.9 и C4 76.8 (Табл. S2) | поменяны местами | 30193 |
| *Myxococcus xanthus* DK1622 | -6)-α-Glcp и -4)-α-GalpNAc-6OMe (Табл. 1, все атомы) | все сигналы остатков поменяны местами | 21880 |
| *Piscirickettsia salmonis* AL10005 | -6)-α-Glcp C4 79.0 | исправлено по аналогии с OS-2 | 29794 |
| *Proteus mirabilis* O20 | -3,4)-α-GlcpNAc C3 70.2 | | 9274 |
| *Proteus mirabilis* O27 | -6)-β-Glcp C6 61.8 | | 5373 |
| *Proteus mirabilis* O43 | -4)-α-GalpA C2 79.3 | исправлено | 1700 |
| *Proteus penneri* 12, 13, 37, 44 | -2)-α-D-gro-D-manHepp C2 70.8 | | 5641,5642,5645,5646 |
| *Proteus vulgaris* CP2-96 | -4)-α-GlcpA C5 74.6 и др. («альтернативная» β-конфигурация противоречит другим сигналам) | аномерная конфигурация удалена | 4306 |
| *Proteus vulgaris* OX2 | t)-α-L-gro-D-manHepp H7 4.74 (все остатки) | | 5331,5332 |
| *Pseudoalteromonas agarivorans* KMM 232 | -3)-α-L-Rhap C2 78.0 | | 25308 |
| *Pseudoalteromonas carrageenovora* IAM 12662T | t/2)-α-Colp C3 21-22 (оба остатка),<br>-2)-β-Galp C3 68.9 | | 10623 |

| | | | |
|---|---|---|---|
| *Pseudomonas syringae* pv. tomato CFBP2545 | -2)-α-L-Rhap C3 76.9 и -2,3)-α-L-Rhap C3 71.2 | поменяны местами | 4932 |
| *Raoultella terrigena* | -2)-β-Manp4,6Pyr C4 67.9 и C5 74.6 | поменяны местами | 21905 |
| *Serratia marcescens* IFO 3735 | -2,4)-α-GalpA C4 71.3 (Табл. 3) | | 20473 |
| *Shigella flexneri* M90T | t)-β-GlcpN C5 71.8, H1 5.31 (остаток F); вероятно, α-аномер. ошибка распространилась в других публикациях. | | 22705 => 25832 и др. |
| *Sphaerotilus natans* | -4)-β-GlcpA C3 68.7 | | 21441 |
| *Sphingomonas* sp. ATCC 31555 | -3)-α-L-Rhap-OH C1 105.61, t)-α-Galp C4/C5 78.07 и др. | | 27196 |
| *Staphylococcus epidermidis* 5 | два остатка -6)-β-GlcpN C5 65.21, C5 72.31 (и, возможно, C6 71.21) | | 22752 |
| *Streptococcus pneumoniae* 10A mutant JA3, JB3 | либо перепутаны спектры мутантов JA3 и JB3, либо ошибка в структуре (2,5-Rib-ol или 4,5-Rib-ol) | структуры поменяны местами | 27018, 27019 |
| *Streptococcus pneumoniae* 11B, 11C, 11F | -3/4)-β-Galp C4 ~75 и C5 ~66 во всех соединениях (Табл. S1-S6) | поменяны местами | 30999-31003 |
| *Streptococcus pneumoniae* 9V SSISP | -3)-β-ManpNAc C3 69.5, C4 79.6, -3)-α-Galp C2 79.3, C4 73.9 (Табл. S1) | | 27131 |
| *Streptococcus thermophilus* Sfi39 | -3,6)-β-Glcp C4 75.8, C5 70.7 (Табл. 4) | поменяны местами | 415 |
| *Vibrio cholerae* O139 | -3,4)-β-GlcpNAc C3 75.44, C4 76.20 | не изменено | 20387 |
| *Yersinia pseudotuberculosis* O3 | t)-β-Parp C4 77.4 | | 23026 |
| *Cordyceps sinensis* | -3)-α-Glcp C5 76.2 (Табл. 3, остатки B и C) | | 41808 |

| | | | |
|---|---|---|---|
| *Phellinus* sp. P0988 | -2,4)-α-Glcp C5, t)-α-Glc C5, -2)-β-Manp C4 и др. | спектры удалены | 42501 |
| *Schizosaccharomyces pombe* | t)-β-Glcp C2 70.0 (все остатки) | | 41286 |
| *Trichoderma reesei* RUTC 30 | t)-α-Manp C5 61.47, -3)-α-Manp C3 (несущая Glc) 70.22 | | 40018, 40039 |
| *Albizia julibrissin* | t)-β-Glc''' C6 71.6 | | 60605 |
| *Aster bellidiastrum* | t)-β-Glcp C3 71.1 и C4 78.2 | поменяны местами | 60347 |
| *Sechium pittieri*; *Sechium talamanense* | -2)-β-L-Arap-1OMe C3 79.4 (Табл. 6) | исправлено | 60571 |

* Цветом обозначено царство: чёрный – бактерии, синий – грибы, зелёный – растения.

** Указано замещение, остаток, конфигурации (когда не очевидно), атом, химический сдвиг. «t)» означает незамещённый остаток.

*** Отсутствие действия означает, что сигнал отмечен как ошибочный и исключён из рассмотрения в модуле ЯМР-моделирования. «Поменяны местами» означает, что исправление сделано в предположении, что структура верна, но в отнесении спектров сигналы перепутаны местами.

Всего по результатам систематического сопоставления структурных данных из разных записей друг с другом и со спектроскопическими данными выявлены и исправлены ошибки в 343 публикациях (из которых 305 посвящены углеводам бактерий). Это число превышает количество публикаций в пп. 3 и 5 списка типов ошибок (см. выше), так как часть ошибок была обнаружена и исправлена вручную в 2009-2017 гг. до разработки специализированной программы; кроме того, ошибки в публикациях не исчерпываются неправильными структурами. В случае, если ошибочная структура не противоречила правилам химии и биосинтеза сахаридов для сохранения возможности поиска в базу попадали как опубликованные данные со специальной пометкой, так и правильные. Кроме ошибок в публикациях с помощью повторного аннотирования по оригинальным статьям было исправлено несколько тысяч ошибок в записях импортированных из базы Carbbank, что ограничило миграцию этих ошибок в современные проек-

ты, использующие данные Carbbank (включая CSDB). Основные типы исправленных ошибок опубликованы в критическом анализе качества данных Carbbank [42] и суммированы в Табл. 9.

Возможность выявлять противоречия между спектрами и структурой с помощью ЯМР-моделирования в CSDB используется и другими коллективами. Например, чл.-корр. Н.Э. Нифантьев и коллеги нашли ошибки в отнесении галактофурананов [393] и хондроитинсульфатов [394], из-за которых в опубликованных статьях других авторов были неправильно идентифицированы критические сигналы (structure reporting signals, C1 и C6 β-D-Gal*f* [395-397], H6/C6 GlcA и C5 β-D-Gal*p*NAc [398]), что привело к неправильному установлению множества природных структур.

Часть ошибок, идентифицированных на основании анализа химических сдвигов, полученных из CSDB для аналогичных структурных фрагментов, была исправлена авторами оригинальных исследований, что привело к пересмотру первичной структуры биогликана того или иного организма и лишению обоснования переноса результатов исследования синтетических моделей на моделируемые природные объекты.

Так, например, выявленная несовместимость структуры О-антигена *Citrobacter braakii* O6 [399] (ID 861) с химическими сдвигами C1 и C5 4-дезоксиарабиногексозы позволила М. Вангу и коллегам [400] (ID 11231) пересмотреть аномерную конфигурацию этого остатка и сопоставить структуру с генетическими данными, доступными для родственного антигена *Franconibacter pulveris* O1. Структура О-антигена *Shigella dysenteria* 5, исследованная в 1977-1990 гг. и многократно опубликованная [401, 402] (ID 122453, 122454 и др.), была позже исправлена авторами и другими группами, но при этом были внесены другие ошибки. Окончательный правильный вариант (ID 23215) получен на основании рассуждений сотрудников группы проф. Ю.А. Книреля [403] и анализа химических сдвигов коллективом CSDB. Подобные примеры не учтены в вышеприведённой статистике, так как к моменту начала систематического скрининга несоответствий они уже были исправлены в базе CSDB по результатам аннотирования поздних статей.

Исправление ошибок в отнесении спектров или хотя бы исключение из рассмотрения заведомо ошибочных данных позволили увеличить точность статистического предсказания химических сдвигов модулем GODDESS (см. раздел 3.4.1) до значений, недостижимых другими методами.

## 5. Выводы

1. Созданы и объединены в согласованную систему многочисленные компьютерные инструменты гликохимии и гликобиологии. Все разработки верифицированы на модельных системах и использованы для реальных исследований. В результате сформировалась молодая область знания – гликоинформатика, был задан и обеспечен мировой вектор её развития. Наработки популяризированы среди химиков и биологов несколькими обзорами, излагающими авторское видение проблем и решений в этом новом разделе биоорганической химии.
2. На основании аннотирования более 10000 публикаций создана и регулярно обновляется уникальная база данных Carbohydrate Structure Database (CSDB) по углеводам микроорганизмов, грибов и растений, обеспечивающая практически полное покрытие в домене прокариот. CSDB содержит структурную, таксономическую, библиографическую, экспериментально-аналитическую и другую информацию и оснащена многочисленными видами поиска, представления и анализа данных. Её функции, алгоритмы и документация свободно доступны научному сообществу через веб-портал http://csdb.glycoscience.ru и активно используются химиками и биологами в собственных исследованиях.
3. Собраны данные по преимущественным конформациям моносахаридов и на основании анализа методов предсказания молекулярной геометрии создан инструмент быстрого автоматического моделирования структуры биогликанов, тем самым заложен фундамент для статистических и прямых расчётов корреляции «структура – свойство» в химии углеводов.
4. Исследована взаимосвязь «структура – спектр ЯМР» для углеводных поли- и олигомеров и конъюгатов. Определены структурные дескрипторы, оказывающие влияние на химические сдвиги ЯМР $^{1}$H и $^{13}$C. Аннотировано более 9000 спектров ЯМР биогликанов[a], созданы способы объединения эмпирических и статистических ЯМР-моделей и оценки их достоверности.

[a] На 2018-й год.

Впервые разработан метод обобщения атомного окружения в углеводных структурах, что открывает путь к статистическому предсказанию не только химических сдвигов, но и других физико-химических параметров. Метод реализован в алгоритме предсказания спектров ЯМР сложных углеводов со средней точностью 0.06 м.д. для $^{1}$H и 0.69 м.д. для $^{13}$C, что превышает показатели остальных существующих подходов.

5. Разработана новая программа генерирования структурных гипотез и их ранжирования по степени соответствия первичной структуры экспериментальным данным (ЯМР, ГЖХ, метилирование и др.), что существенно облегчает процесс установления первичной структуры природных сахаридов и гликоконъюгатов и в настоящее время является единственным инструментом, справляющимся с полуавтоматическим установлением строения произвольных углеводов по спектрам.
6. Разработан углеводный язык CSDB Linear (с поддержкой не полностью определенных структур), впервые сочетающий машино- и человекочитаемость. Реализован перевод с этого языка на другие углеводные и общехимические языки, тем самым семантическое описание углеводов, используемое в большинстве публикаций, впервые эффективно связано с поатомным описанием, используемым в химических расчётах. Разработаны протоколы визуализации углеводных структур (нотация SNFG, в сотрудничестве с Консорциумом по функциональной гликомике) и их одно- и двумерных спектров ЯМР. SNFG признана в качестве рекомендованного стандарта ведущими углеводными научными журналами.
7. Создана уникальная база данных по подтверждённым активностям гликозилтрансфераз (~1200 установленных и ~600 предсказанных функций ферментов[a]), объединяющая гены, ферменты, субстраты, синтезируемую структуру, таксономию, библиографию, методы и достоверность установления активности. База является полной по двум наиболее изученным таксонам бактерий и растений (*E. coli* и *A. thaliana*).

[a] На 2018-й год.

8. Проведён статистический анализ распространённости структурных особенностей углеводов в различных таксономических группах, выявлены характерные признаки этих групп. Создан инструмент построения альтернативных «деревьев жизни», основанный на схожести и различиях химической структуры биогликанов. Полученные дендрограммы позволили объяснить сходство отдельных таксонов, выходящее за рамки классической филогенетики.
9. Разработаны стандарты и форматы хранения и обработки данных по углеводам (включая углеводную онтологию GlycoRDF), протоколы построения и взаимодействия долговременных углеводных баз данных и программ. Эти правила были признаны большинством мировых коллективов, работающих с информацией об углеводах, в качестве способствующих прогрессу систематизации углеводов. Налажено взаимодействие между основными углеводными базами данных на автоматическом уровне, что позволяет учёным получать знания, неявно содержащиеся в нескольких базах разного типа.
10. По результатам критического анализа качества данных в других базах в них выявлено ~2300 ошибок, в том числе ~340 ошибок в первичной структуре или ЯМР-спектрах углеводов в существующих публикациях. Найденные ошибки исправлены на основании анализа спектров при размещении данных в CSDB.

## 6. Возможные направления дальнейшего развития

Компьютерная часть проекта CSDB находится в непрерывном развитии, чтобы соответствовать новым запросам пользователей и публикуемым данным. Направления, разработка которых не была полностью закончена и опубликована к 2019-му году, не вошли в диссертацию. Они не влияют на завершенность представленных исследований и работоспособность представленных инструментов. Эти направления включают:

*Глобальные направления развития:*

1. Поддержание актуального покрытия по углеводам прокариот путём систематического экспертного аннотирования статей и достижение полного покрытия по углеводам других организмов, в первую очередь растений и патогенных грибов.
2. Адаптация базы к особенностям биогликанов животных (включая человека) с учётом информации, собранной в других проектах, и расширение покрытия на все природные углеводы.
3. Достижение полного покрытия по подтверждённым активностям гликозилтрансфераз и других углевод-активных ферментов.
4. Обеспечение связи всех существующих языков описания углеводов с атомными моделями и дальнейшая интеграция с другими сервисами, использующими эти модели (квантовомеханические расчёты).
5. Генерирование, валидация и хранение конформационных карт ди- и тримерных фрагментов, получение начальных геометрий произвольных гликанов, предсказание стерических свойств и ЯЭО (NOE).
6. Оптимизация ресурсоёмких алгоритмов ЯМР-моделирования и организация их многопоточного выполнения, расширение ЯМР-спектроскопической базы на все опубликованные спектры углеводов.
7. Добавление в базу данных подробной масс-спектрометрической информации и интеграция с программами анализа структур по масс-спектрам.

8. Автоматический сбор и интерпретация неструктурированных данных из других баз (Pubmed, MESH) и Интернет, создание инструментов извлечения знаний, машинного аннотирования и верификации статей.
9. Разработка алгоритмов оценки структурно-химического родства углеводов и обеспечение возможности поиска в базе данных структур по характерным мотивам; предсказание иммунологических кросс-реакций на основании статистического анализа и «нечёткого» сравнения таких мотивов.
10. Валидация опубликованных структур, для которых обнаружены несоответствия между структурами и спектрами, путём повторного анализа экспериментальных данных из оригинальных исследований.
11. Разработка новых подходов к визуализации первичной и пространственной структуры биогликанов.

*Технические и административные задачи:*

1. Систематическое выявление и исправление ошибок в записях базы и в исходных публикациях.
2. Создание формализованного программного интерфейса (API) для использования CSDB роботами других проектов.
3. Выпуск новой версии нотации CSDB Linear с возможностью комбинирования повторяющихся и неповторяющихся фрагментов структуры, улучшенной поддержкой гетерогенных структур, и интеграцией с языком SMILES для описания агликонов.
4. Обеспечение связи CSDB с каталогом ICD11 для систематического выявления связи определённых гликанов с заболеваниями.
5. Адаптация новых версий модуля SugarSketcher для графического ввода и редактирования углеводных структур.
6. Создание анализатора графических (CFG, SNFG) и псевдографических изображений углеводных структур в статьях, и автоматическое генерирование из них описаний молекул в формальной нотации.
7. Поиск структур, содержащих гликоэпитопы, размещённые в базе GlycoEpitope, и обеспечение кросс-ссылок с этой базой.

8. Поиск ссылок Uniprot на гликопротеины, несущие гликаны, аналогичные тем, что присутствуют в записях CSDB, и обеспечение кросс-ссылок с этой базой.
9. Дополнение списка детектируемых несоответствий в данных и реализация новых проверок, в том числе путём статистического анализа всего содержимого базы.
10. Расширение функциональности нотации GlycoCT, используемой для связи с другими проектами, и модификация структурного транслятора для обработки структур с неопределённостями.
11. Создание веб-робота для автоматического опроса сайтов издательств и библиографических агрегаторов, выявляющего подходящие для аннотирования публикации на основе мета-информации и семантического анализа рефератов статей.
12. Создание многоязычного интерфейса (или, как минимум, добавление русского языка).
13. Продолжение непрерывного бета-тестирования, сбора пожеланий пользователей и устранения выявляемыми недочётов в работе веб-сервисов проекта.
14. Лоббирование использования универсальных идентификаторов структур в публикациях.

## 7. Использованные сокращения

Жирным шрифтом приведены аббревиатуры, появившиеся в результате данной работы.

| | |
|---|---|
| БД | база данных |
| ВЭЖХ | высокоэффективная жидкостная хроматография |
| ГЖХ | газожидкостная хроматография |
| ГХ | газовая хроматография |
| КССВ | константа спин-спинового взаимодействия |
| МС | масс-спектрометрия |
| РСА | рентгено-структурный анализ |
| рРНК | рибосомальная рибонуклеиновая кислота |
| СУБД | система управления базами данных |
| ХС | химический сдвиг |
| ЯМР | ядерный магнитный резонанс |
| ЯЭО | ядерный эффект Оверхаузера |
| АВС | Artificial Bee Colony<br>искусственная пчелиная колония |
| **BCSDB** | Bacterial Carbohydrate Structure DataBase<br>база данных бактериальных углеводных структур |

| | |
|---|---|
| **BIOPSEL** | BIOPolymer Structure ELucidation<br>установление структуры биополимеров |
| BIONJ | BIOlogical Neighbor Joining<br>биологическое соединение соседних узлов |
| CASPER | Computer Assisted SPectrum Evaluation of Regular polysaccharides<br>компьютерный анализ спектров регулярных полисахаридов |
| CAZy | carbohydrate-active enzymes,<br>углевод-активные ферменты |
| CCSD | Complex Carbohydrate Structure Database<br>база данных сложных углеводов |
| CFG | Consortium for Functional Glycomics<br>консорциум по функциональной гликомике |
| COSY | COrrelation SpectroscopY<br>корреляционная спектроскопия |
| **CSDB** | Carbohydrate Structure DataBase<br>база данных углеводных структур |
| CSV | Comma-Separated Values<br>значения, разделённые запятыми |
| DEPT | Distorsionless Enhancement by Polarization Transfer<br>улучшение без искажений за счёт переноса поляризации |
| DFT B3LYP | Density Functional Theory: Becke 3-parameter Lee-Yang-Parr<br>теория функционала плотности: Бекке-Ли-Янг-Парр |
| DHTML | Dynamic HyperText Markup Language<br>динамический язык разметки гипертекста |

| | |
|---|---|
| DOI | Digital Object Identifier<br>цифровой идентификатор объекта |
| GLIC | GLycoInformatic Consortium<br>консорциум по гликоинформатике |
| **GODDESS** | Glyco-Optimized Database-Driven Empirical Spectrum Simulation<br>оптимизированное для углеводов эмпирическое предсказание спектров по базе данных |
| **GRASS** | Generation, Ranking and Assignment of Saccharide Structures<br>генерирование, ранжирование и отнесение спектров структур сахаридов |
| GlycoCT | Glycologic Connection Table<br>гликологическая таблица связности |
| GT | GlycosylTransferase<br>гликозилтрансфераза |
| HMBC | Heteronuclear Multiple Bond Correlation<br>гетероядерная многосвязевая корреляция |
| HOSE | Hierarchical Organization of Spherical Environment<br>иерархическая организация сферического окружения |
| HSQC | Heteronuclear Single Quantum Coherence<br>гетероядерная одноквантовая когерентность |
| ICD | International Classification of Diseases<br>международная классификация заболеваний |
| IUPAC | International Union of Pure and Applied Chemistry<br>международное объединение по фундаментальной и прикладной химии |

| | |
|---|---|
| LinUCS | LInear Notation for Unique description of Carbohydrate Sequences<br>линейная нотация для уникального описания углеводных последовательностей |
| MESH | MEdical Subject Headings<br>медицинские предметные термины |
| MMFF94 | Merck Molecular Force Field 1994<br>молекулярно-механическое силовое поле лаборатории Мерк (1994) |
| MSDB | MonoSaccharide DataBase<br>база данных моносахаридов |
| NCBI | National Center for Biotechnology Information<br>национальный центр по биотехнологической информации (США) |
| NIH | National Institute of Health<br>Национальный институт здоровья (США) |
| NLM | National Library of Medicine<br>национальная медицинская библиотека (США) |
| NOE | Nuclear Overhauser Effect<br>ядерный эффект Оверхаузера |
| OWL | Ontology Web Language<br>сетевой язык описания онтологий |
| PDB | Protein Data Bank<br>белковая база данных |
| **PFCSDB** | Plant and Fungal Carbohydrate Structure DataBase<br>база данных растительных и грибных углеводных структур |
| PMID | PubMed IDentifier<br>идентификатор Pubmed |

| | |
|---|---|
| RDF | Resource Description Framework<br>среда описания ресурсов |
| **REStLESS** | ResiduEs as SMILES, LinkagEs as SMARTS<br>остатки как SMILES, связи как SMARTS |
| SMARTS | SMiles ARbitrary Target Specification<br>произвольное указание присоединённых компонентов структуры в SMILES |
| SMILES | Simplified Molecular-Input Line-Entry System<br>упрощённая система для ввода молекул одной строкой |
| **SNFG** | Symbolic Notation For Glycans<br>символическая углеводная нотация |
| SPARQL | SPARQL Protocol and RDF Query Language<br>протокол SPARQL и язык запросов для среды описания ресурсов RDF |
| SQL | Structured Query Language<br>структурированный язык запросов |
| TaxID | Taxonomical IDentifier<br>таксономический идентификатор |
| TOCSY | TOtal Correlation SpectroscopY<br>тотальная корреляционная спектроскопия |
| WSDL | Web Service Description Language<br>язык описания сетевых сервисов |
| WURCS | Web-3 Unique Representation of Carbohydrate Structures<br>уникальное представление углеводных структур для всемирной сети 3 |

| | |
|---|---|
| XML | eXtensible Markup Language<br>расширяемый язык разметки |

Сокращённые названия мономерных остатков используются в тексте в соответствии с правилами IUPAC и нотацией CSDB Linear. Полная таблица этих сокращений (489 остатков) приведена в разделе Monomer namespace на сайте проекта[a].

[a] http://csdb.glycoscience.ru/database/core/residues.php

## 8. Список литературы


1) Robyt J. F. General Occurrence of Carbohydrates // Glycoscience: Chemistry and Chemical Biology I–III / Fraser-Reid B. O. и др. Springer, Berlin, Heidelberg, 2001. – Гл. 1.4, С. 75-114.

2) Sharon N. Carbohydrates // Scientific American. – 1980. – Т. 243, № 5. – С. 90-116.

3) Varki A. Biological roles of oligosaccharides: all of the theories are correct // Glycobiology. – 1993. – Т. 3, № 2. – С. 97-130.

4) Varki A. Biological roles of glycans // Glycobiology. – 2017. – Т. 27, № 1. – С. 3-49.

5) Springer S. A., Gagneux P. Glycan evolution in response to collaboration, conflict, and constraint // Journal of Biological Chemistry. – 2013. – Т. 288, № 10. – С. 6904-6911.

6) Ohtsubo K., Marth J. D. Glycosylation in cellular mechanisms of health and disease // Cell. – 2006. – Т. 126, № 5. – С. 855-867.

7) Dwek M., Markiv A. Glycosylation and Disease // Encyclopedia of Life Sciences. – Chichester: John Wiley & Sons Ltd, 2018.

8) Jaeken J. Glycosylation and its Disorders: General Overview // Reference Module in Biomedical SciencesElsevier, 2016.

9) Freeze H. H., Ng B. G. Glycomics, glycobiology, and glyco-medicine // Genomic and Personalized Medicine / Ginsburg G. S., Willard H. F. Academic Press, 2013. – Гл. 15, С. 173-191.

10) Almeida A., Kolarich D. The promise of protein glycosylation for personalised medicine // Biochimica et Biophysica Acta. – 2016. – Т. 1860, № 8. – С. 1583-1595.

11) Lauc G. Sweet secret of the multicellular life // Biochimica Biophysica Acta. – 2006. – Т. 1760, № 4. – С. 525-526.

12) Griffin M. E., Hsieh-Wilson L. C. Glycan engineering for cell and developmental biology // Cell Chemical Biology. – 2016. – Т. 23, № 1. – С. 108-121.

13) Weir D. M. Carbohydrates as recognition molecules in infection and immunity // FEMS Microbiology and Immunolohy. – 1989. – Т. 1, № 6-7. – С. 331-340.

14) Cobb B. A., Kasper D. L. Coming of age: carbohydrates and immunity // European Journal of Immunology. – 2005. – Т. 35, № 2. – С. 352-356.

15) Lehninger Principles of Biochemistry. 7th Edition. / Nelson D. L., Cox M. M. – New York City: W. H. Freeman and Company, 2017.

16) Bennun S. V., Hizal D. B., Heffner K., Can O., Zhang H., Betenbaugh M. J. Systems glycobiology: integrating glycogenomics, glycoproteomics, glycomics, and other 'omics data sets to characterize cellular glycosylation processes // Journal of Molecular Biology. – 2016. – Т. 428, № 16. – С. 3337-3352.

17) Lu H., Zhang Y., Yang P. Advancements in mass spectrometry-based glycoproteomics and glycomics // National Science Review. – 2016. – Т. 3, № 3. – С. 345-364.

18) Toukach P. V., Egorova K. S. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts // Nucleic Acids Research. – 2016. – Т. 44, № D1. – С. D1229-D1236.

19) Lowe J. B., Marth J. D. A genetic approach to Mammalian glycan function // Annual Reviews in Biochemistry. – 2003. – Т. 72. – С. 643-691.

20) Liu G., Neelamegham S. Integration of systems glycobiology with bioinformatics toolboxes, glycoinformatics resources, and glycoproteomics data // WIREs Systems Biology and Medicine. – 2015. – Т. 7, № 4. – С. 163-181.

21) Glycoinformatics. Methods in Molecular Biology. Под ред. Walker J. M. – New York: Springer, 2015. Methods in Molecular Biology. – 506 с.

22) Prokaryotic Cell Wall Compounds. – 1 изд. – Berlin - Heidelberg: Springer, 2010. – 517 с.

23) Fungal Cell Wall: Structure, Synthesis, and Assembly. 2nd Edition. / Ruiz-Herrera J. – Boca Raton: CRC Press, 2012.

24) Koch A. L. Bacterial wall as target for attack: past, present, and future research // Clinical microbiology reviews. – 2003. – Т. 16, № 4. – С. 673–687.

25) Latge J. P. The cell wall: a carbohydrate armour for the fungal cell // Molecular Microbiology. – 2007. – Т. 66, № 2. – С. 279-290.

26) Herget S., Toukach P. V., Ranzinger R., Hull W. E., Knirel Y. A., von der Lieth C. W. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans // BMC Structural Biology. – 2008. – Т. 8. – С. ID 35.

27) Snapper C. M. Mechanisms underlying in vivo polysaccharide-specific immunoglobulin responses to intact extracellular bacteria // Annals of the New York Academy of Sciences. – 2012. – Т. 1253. – С. 92-101.

28) Zhou J. Y., Oswald D. M., Oliva K. D., Kreisman L. S. C., Cobb B. A. The glycoscience of immunity // Trends in Immunology. – 2018. – Т. 39, № 7. – С. 523-535.

29) Sukhithasri V., Nisha N., Biswas L., Anil Kumar V., Biswas R. Innate immune recognition of microbial cell wall components and microbial strategies to evade such recognitions // Microbiological Research. – 2013. – Т. 168, № 7. – С. 396-406.

30) Jones C. Vaccines based on the cell surface carbohydrates of pathogenic bacteria // Academia Brasileira de Ciências. – 2005. – Т. 77, № 2. – С. 293-324.

31) Egorova K. S., Kondakova A. N., Toukach P. V. Carbohydrate Structure Database: tools for statistical analysis of bacterial, plant and fungal glycomes // Database (Oxford). – 2015. – Т. 2015. – С. ID bav073.

32) Khalid E. B., Ayman E. E., Rahman H., Abdelkarim G., Najda A. Natural products against cancer angiogenesis // Tumour Biology. – 2016. – Т. 37, № 11. – С. 14513-14536.

33) Lutteke T. The use of glycoinformatics in glycochemistry // Beilstein Journal of Organic Chemistry. – 2012. – Т. 8. – С. 915-929.

34) Benson D. A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D. J., Ostell J., Sayers E. W. GenBank // Nucleic Acids Research. – 2013. – Т. 41, № Database issue. – С. D36-D42.

35) Consortium T. U. UniProt: the universal protein knowledgebase // Nucleic Acids Research. – 2017. – Т. 45, № D1. – С. D158-D169.

36) Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature // Database (Oxford). – 2011. – Т. 2011. – С. ID baq036.

37) Bebbington P. Welcome to ICD-10 // Social Psychiatry and Psychiatric Epidemiology. – 1992. – Т. 27, № 6. – С. 255-257.

38) Chen C., Huang H., Wu C. H. Protein bioinformatics databases and resources // Methods in Molecular Biology. – 2017. – Т. 1558. – С. 3-39.

39) Bertozzi C. R., Rabuka D. Structural Basis of Glycan Diversity // Essentials of Glycobiology / Varki A. и др. – New-York: Cold Spring Harbor Laboratory Press, 2009. – Гл. 2, С. 23-36.

40) Egorova K. S., Toukach P. V. Glycoinformatics: bridging isolated islands in the sea of data // Angewandte Chemie International Edition. – 2018. – Т. 57, № 46. – С. 14986-14990.

41) Lombard V., Golaconda Ramulu H., Drula E., Coutinho P. M., Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013 // Nucleic Acids Research. – 2014. – Т. 42, № Database issue. – С. D490-D495.

42) Egorova K. S., Toukach P. V. Critical analysis of CCSD data quality // Journal of Chemical Information and Modeling. – 2012. – Т. 52, № 11. – С. 2812–2814.

43) Willighagen E. L., Brandle M. P. Resource description framework technologies in chemistry // Journal of Cheminformatics. – 2011. – Т. 3, № 1. – С. ID 15.

44) Ranzinger R., Aoki-Kinoshita K. F., Campbell M. P., Kawano S., Lutteke T., Okuda S., Shinmachi D., Shikanai T., Sawaki H., Toukach P., Matsubara M., Yamada I., Narimatsu H. GlycoRDF: an ontology to standardize glycomics data in RDF // Bioinformatics. – 2015. – Т. 31, № 6. – С. 919–925.

45) Katayama T., Wilkinson M. D., Aoki-Kinoshita K. F., Kawashima S., Yamamoto Y., Yamaguchi A., Okamoto S., Kawano S., Kim J. D., Wang Y., Wu H., Kano Y., Ono H., Bono H., Kocbek S., Aerts J., Akune Y., Antezana E., Arakawa K., Aranda B., Baran J., Bolleman J., Bonnal R. J., Buttigieg P. L., Campbell M. P., Chen Y. A., Chiba H., Cock P. J., Cohen K. B., Constantin A., Duck G., Dumontier M., Fujisawa T., Fujiwara T., Goto N., Hoehndorf R., Igarashi Y., Itaya H., Ito M., Iwasaki W., Kalas M., Katoda T., Kim T., Kokubu A., Komiyama Y., Kotera M., Laibe C., Lapp H., Lutteke T., Marshall M. S., Mori T., Mori H., Morita M., Murakami K., Nakao M., Narimatsu H., Nishide H., Nishimura Y., Nystrom-Persson J., Ogishima S., Okamura Y., Okuda S., Oshita K., Packer N. H., Prins P., Ranzinger R., Rocca-Serra P., Sansone S., Sawaki H., Shin S. H., Splendiani A., Strozzi F., Tadaka S., Toukach P., Uchiyama I., Umezaki M., Vos R., Whetzel P. L., Yamada I., Yamasaki C., Yamashita R., York W. S., Zmasek C. M., Kawamoto S., Takagi T. BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains // Journal of Biomedical Semantics. – 2014. – Т. 5, № 1. – С. ID 5.

46) Herget S., Ranzinger R., Maass K., Lieth C. W. GlycoCT-a unifying sequence format for carbohydrates // Carbohydrate Research. – 2008. – Т. 343, № 12. – С. 2162-2171.

47) Matsubara M., Aoki-Kinoshita K. F., Aoki N. P., Yamada I., Narimatsu H. WURCS 2.0 update to encapsulate ambiguous carbohydrate structures // Journal of Chemical Information and Modeling. – 2017. – Т. 57, № 4. – С. 632-637.

48) Tiemeyer M., Aoki K., Paulson J., Cummings R. D., York W. S., Karlsson N. G., Lisacek F., Packer N. H., Campbell M. P., Aoki N. P., Fujita A., Matsubara M., Shinmachi D., Tsuchiya S., Yamada I., Pierce M., Ranzinger R., Narimatsu H., Aoki-Kinoshita K. F. GlyTouCan: an accessible glycan structure repository // Glycobiology. – 2017. – Т. 27, № 10. – С. 915-919.

49) Lisacek F., Mariethoz J., Alocci D., Rudd P. M., Abrahams J. L., Campbell M. P., Packer N. H., Stahle J., Widmalm G., Mullen E., Adamczyk B., Rojas-Macias M. A., Jin C., Karlsson N. G. Databases and associated tools for glycomics and glycoproteomics // High-Throughput Glycomics and Glycoproteomics / Lauc G., Wuhrer M. – New York, NY: Humana Press, 2017. – Гл. 18, С. 235-264.

50) A Practical Guide to Using Glycomics Databases. – 1 изд.: Springer Japan, 2017. – 370 с.

51) Frank M., Schloissnig S. Bioinformatics and molecular modeling in glycobiology // Cellular and molecular life sciences : CMLS. – 2010. – Т. 67, № 16. – С. 2749-2772.

52) Aoki-Kinoshita K. F. Using databases and web resources for glycomics research // Molecular and Cellular Proteomics. – 2013. – Т. 12, № 4. – С. 1036-1045.

53) Toukach P. V. CSDB and other carbohydrate databases // Glycoconjugate Journal. – 2013. – Т. 30. – С. 347-349.

54) Lütteke T., Bohne-Lang A., Loss A., Goetz T., Frank M., von der Lieth C. W. GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research // Glycobiology. – 2006. – Т. 16, № 5. – С. 71R–81R.

55) Maeda M., Fujita N., Suzuki Y., Sawaki H., Shikanai T., Narimatsu H. JCGGDB: Japan Consortium for Glycobiology and Glycotechnology Database // Glycoinformatics / Lütteke T., Frank M. – New York: Humana Press, 2015. – Гл. 12, С. 161–179.

56) Campbell M. P., Peterson R., Mariethoz J., Gasteiger E., Akune Y., Aoki-Kinoshita K. F., Lisacek F., Packer N. H. UniCarbKB: building a knowledge platform for glycoproteomics // Nucleic Acids Research. – 2014. – Т. 42, № Database issue. – С. D215-D221.

57) Rojas-Macias M. A., Ståhle J., Lütteke T., Widmalm G. Development of the ECODAB into a relational database for Escherichia coli O-antigens and other bacterial polysaccharides // Glycobiology. – 2015. – Т. 25, № 3. – С. 341-347.

58) Loss A., Stenutz R., Schwarzer E., von der Lieth C. W. GlyNest and CASPER: two independent approaches to estimate 1H and 13C NMR shifts of glycans available through a common web-interface // Nucleic Acids Research. – 2006. – Т. 34, № Web Server issue. – С. W733-W737.

59) Kirschner K. N., Yongye A. B., Tschampel S. M., Gonzalez-Outeirino J., Daniels C. R., Foley B. L., Woods R. J. GLYCAM06: a generalizable biomolecular force field. Carbohydrates // Journal of Computational Chemistry. – 2008. – Т. 29, № 4. – С. 622-655.

60) Frank M., Lutteke T., von der Lieth C. W. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages // Nucleic Acids Research. – 2007. – Т. 35, № Database issue. – С. 287-290.

61) Doubet S., Bock K., Smith D., Darvill A., Albersheim P. The Complex Carbohydrate Structure Database // Trends in Biochemical Sciences. – 1989. – Т. 14, № 12. – С. 475–477.

62) Doubet S., Albersheim P. CarbBank // Glycobiology. – 1992. – Т. 2, № 6. – С. 505–507.

63) Birch J., Van Calsteren M.-R., Pérez S., Svensson B. The Exopolysaccharide Properties and Structures Database: EPS-DB. Application to Bacterial Exopolysaccharides // Carbohydrate Polymers. – 2018.10.1016/j.carbpol.2018.10.063. – С. ePub ahead of print.

64) Toukach P. V. Bacterial carbohydrate structure database 3: principles and realization // Journal of Chemical Information and Modeling. – 2011. – Т. 51, № 1. – С. 159-170.

65) Cooper C. A., Harrison M. J., Wilkins M. R., Packer N. H. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources // Nucleic Acids Research. – 2001. – Т. 29, № 1. – С. 332–335.

66) Cooper C. A., Joshi H. J., Harrison M. J., Wilkins M. R., Packer N. H. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update // Nucleic Acids Research. – 2003. – Т. 31, № 1. – С. 511-513.

67) Raman R., Venkataraman M., Ramakrishnan S., Lang W., Raguram S., Sasisekharan R. Advancing glycomics: Implementation strategies at the consortium for functional glycomics // Glycobiology. – 2006. – Т. 16, № 5. – С. 82R–90R.

68) Aoki-Kinoshita K. F., Kanehisa M. Glycomic analysis using KEGG GLYCAN // Glycoinformatics / Lütteke T., Frank M. – New York: Humana Press, 2015. – Гл. 7, С. 97–107.

69) Shinmachi D., Yamada I., Aoki N. P., Matsubara M., Aoki-Kinoshita K. F., Narimatsu H. Using GlyTouCan Version 1.0: The First International Glycan Structure Repository // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. Springer Japan, 2017. – Гл. 4, С. 41–73.

70) Toukach P. V., Egorova K. S. Bacterial, Plant, and Fungal Carbohydrate Structure Databases: daily usage // Glycoinformatics / Lütteke T., Frank M. – New York: Humana Press, 2015. – Гл. 5, С. 55-85.

71) Hizal D. B., Wolozny D., Colao J., Jacobson E., Tian Y., Krag S. S., Betenbaugh M. J., Zhang H. Glycoproteomic and glycomic databases // Clinical Proteomics. – 2014. – Т. 11. – С. ID 15.

72) Ranzinger R., Herget S., von der Lieth C. W., Frank M. GlycomeDB - a unified database for carbohydrate structures // Nucleic Acids Research. – 2011. – Т. 39. – С. D373-D376.

73) von der Lieth C. W., Freire A. A., Blank D., Campbell M. P., Ceroni A., Damerell D. R., Dell A., Dwek R. A., Ernst B., Fogh R., Frank M., Geyer H., Geyer R., Harrison M. J., Henrick K., Herget S., Hull W. E., Ionides J., Joshi H. J., Kamerling J. P., Leeflang B. R., Lutteke T., Lundborg M., Maass K., Merry A., Ranzinger R., Rosen J., Royle L., Rudd P. M., Schloissnig S., Stenutz R., Vranken W. F., Widmalm G., Haslam S. M. EUROCarbDB: An open-access platform for glycoinformatics // Glycobiology. – 2011. – Т. 21, № 4. – С. 493–502.

74) Egorova K. S., Toukach P. V. CSDB_GT: a new curated database on glycosyltransferases // Glycobiology. – 2017. – Т. 27, № 4. – С. 285-290.

75) Toukach P. V., Knirel Y. A. New database of bacterial carbohydrate structures // Glycoconjugate Journal. – 2005. – Т. 22. – С. 216-217.

76) Hashimoto K., Goto S., Kawano S., Aoki-Kinoshita K. F., Ueda N., Hamajima M., Kawasaki T., Kanehisa M. KEGG as a glycome informatics resource // Glycobiology. – 2006. – Т. 16, № 5. – С. 63R-70R.

77) Lütteke T. Glycan data retrieval and analysis using GLYCOSCIENCES. de applications // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. – Tokyo, Japan: Springer Japan, 2017. – Гл. 16, С. 335–350.

78) Bohm M., Bohne-Lang A., Frank M., Loss A., Rojas-Macias M. A., Lutteke T. Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update) // Nucleic Acids Research. – 2018.10.1093/nar/gky994 № Database Issue. – C. ePub ahead of print.

79) Campbell M. P., Royle L., Radcliffe C. M., Dwek R. A., Rudd P. M. GlycoBase and autoGU: tools for HPLC-based glycan analysis // Bioinformatics. – 2008. – T. 24, № 9. – C. 1214-1216.

80) Zhao S., Walsh I., Abrahams J. L., Royle L., Nguyen-Khuong T., Spencer D., Fernandes D. L., Packer N. H., Rudd P. M., Campbell M. P. GlycoStore: a database of retention properties for glycan analysis // Bioinformatics. – 2018.10.1093/bioinformatics/bty319.

81) Nakahara T., Hashimoto R., Nakagawa H., Monde K., Miura N., Nishimura S. Glycoconjugate Data Bank: Structures--an annotated glycan structure database and N-glycan primary structure verification service // Nucleic Acids Research. – 2008. – T. 36, № Database issue. – C. D368-D371.

82) Kameyama A., Kikuchi N., Nakaya S., Ito H., Sato T., Shikanai T., Takahashi Y., Takahashi K., Narimatsu H. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library // Analytical Chemistry. – 2005. – T. 77, № 15. – C. 4719-4725.

83) van Kuik J. A., Hard K., Vliegenthart J. F. A 1H NMR database computer program for the analysis of the primary structure of complex carbohydrates // Carbohydrate Research. – 1992. – T. 235. – C. 53-68.

84) Ranzinger R., Herget S., Wetter T., von der Lieth C. W. GlycomeDB - integration of open-access carbohydrate structure databases // BMC Bioinformatics. – 2008. – T. 9. – C. ID 384.

85) Ranzinger R., Frank M., von der Lieth C. W., Herget S. Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences // Glycobiology. – 2009. – T. 19, № 12. – C. 1563-1567.

86) Lundborg M., Modhukur V., Widmalm G. Glycosyltransferase functions of E. coli O-antigens // Glycobiology. – 2010. – T. 20, № 3. – C. 366-368.

87) Cantarel B. L., Coutinho P. M., Rancurel C., Bernard T., Lombard V., Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics // Nucleic Acids Research. – 2009. – T. 37, № Database issue. – C. D233-8.

88) Chang A., Schomburg I., Placzek S., Jeske L., Ulbrich M., Xiao M., Sensen C. W., Schomburg D. BRENDA in 2015: exciting developments in its 25th year of existence // Nucleic Acids Research. – 2015. – Т. 43, № Database issue. – С. D439-D446.

89) Scheer M., Grote A., Chang A., Schomburg I., Munaretto C., Rother M., Sohngen C., Stelzer M., Thiele J., Schomburg D. BRENDA, the enzyme information system in 2011 // Nucleic Acids Research. – 2011. – Т. 39, № Database issue. – С. D670-D676.

90) Togayachi A., Dae K.-Y., Shikanai T., Narimatsu H. A database system for glycogenes (GGDB) // Experimental glycoscience / Taniguchi N. и др. Japan Springer, 2008, С. 423–425.

91) Akiyoshi S., Nomura K. H., Dejima K., Murata D., Matsuda A., Kanaki N., Takaki T., Mihara H., Nagaishi T., Furukawa S., Ando K. G., Yoshina S., Mitani S., Togayachi A., Suzuki Y., Shikanai T., Narimatsu H., Nomura K. RNAi screening of human glycogene orthologs in the nematode *Caenorhabditis elegans* and the construction of the C. elegans glycogene database // Glycobiology. – 2015. – Т. 25, № 1. – С. 8-20.

92) Kawasaki T., Nakao H., Takahashi E., Tominaga T. GlycoEpitope: the integrated database of carbohydrate antigens and antibodies // Trends in Glycoscience and Glycotechnology. – 2006. – Т. 18, № 102. – С. 267-272.

93) Kawasaki T., Nakao H., Tominaga T. GlycoEpitope: a database of carbohydrate epitopes and antibodies // Experimental Glycoscience / Taniguchi N. и др. Springer Tokyo, 2008. – Гл. 104, С. 429-431.

94) Manyam G., Birerdinc A., Baranova A. KPP: KEGG Pathway Painter // BMC System Biology. – 2015. – Т. 9 Suppl 2. – С. S3.

95) Tanabe M., Kanehisa M. Using the KEGG database resource // Current Protocols in Bioinformatics. – 2012. – Т. 38, № 1. – С. 1.12.1-1.12.43.

96) Hirabayashi J., Tateno H., Shikanai T., Aoki-Kinoshita K. F., Narimatsu H. The Lectin Frontier Database (LfDB), and data generation based on frontal affinity chromatography // Molecules. – 2015. – Т. 20, № 1. – С. 951-973.

97) Kaji H., Shikanai T., Sasaki-Sawa A., Wen H., Fujita M., Suzuki Y., Sugahara D., Sawaki H., Yamauchi Y., Shinkawa T., Taoka M., Takahashi N., Isobe T., Narimatsu H. Large-scale identification of N-glycosylated proteins of mouse tissues and construction of a glycoprotein database, GlycoProtDB // Journal of Proteome Research. – 2012. – Т. 11, № 9. – С. 4553-4566.

98) Gupta R., Birch H., Rapacki K., Brunak S., Hansen J. E. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins // Nucleic Acids Research. – 1999. – Т. 27, № 1. – С. 370-372.

99) Shakhsheer B., Anderson M., Khatib K., Tadoori L., Joshi L., Lisacek F., Hirschman L., Mullen E. SugarBind database (SugarBindDB): a resource of pathogen lectins and corresponding glycan targets // Journal of Molecular Recognition. – 2013. – Т. 26, № 9. – С. 426-431.

100) Mariethoz J., Khatib K., Alocci D., Campbell M. P., Karlsson N. G., Packer N. H., Mullen E. H., Lisacek F. SugarBindDB, a resource of glycan-mediated host-pathogen interactions // Nucleic Acids Research. – 2016. – Т. 44, № D1. – С. D1243-D1250.

101) Lutteke T., von der Lieth C. W. MonoSaccharideDB: A reference resource to unify the notation of carbohydrate residues // Glycobiology. – 2005. – Т. 15, № 11. – С. 1209-1210.

102) Lütteke T. Translation and validation of carbohydrate residue names with MonosaccharideDB routines // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. Springer Japan, 2017. – Гл. 3, С. 29-40.

103) Krishnakumar V., Hanlon M. R., Contrino S., Ferlanti E. S., Karamycheva S., Kim M., Rosen B. D., Cheng C. Y., Moreira W., Mock S. A., Stubbs J., Sullivan J. M., Krampis K., Miller J. R., Micklem G., Vaughn M., Town C. D. Araport: the *Arabidopsis* information portal // Nucleic Acids Research. – 2015. – Т. 43, № Database issue. – С. D1003-D1009.

104) Lamesch P., Berardini T. Z., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., Dreher K., Alexander D. L., Garcia-Hernandez M., Karthikeyan A. S., Lee C. H., Nelson W. D., Ploetz L., Singh S., Wensel A., Huala E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools // Nucleic Acids Research. – 2012. – Т. 40, № Database issue. – С. D1202-D1210.

105) Mueller L. A., Zhang P., Rhee S. Y. AraCyc: a biochemical pathway database for *Arabidopsis* // Plant Physiology. – 2003. – Т. 132, № 2. – С. 453–460.

106) Cao P. J., Bartley L. E., Jung K. H., Ronald P. C. Construction of a rice glycosyltransferase phylogenomic database and identification of rice-diverged glycosyltransferases // Molecular Plant. – 2008. – Т. 1, № 5. – С. 858-877.

107) Mariethoz J., Alocci D., Gastaldello A., Horlacher O., Gasteiger E., Rojas-Macias M., Karlsson N. G., Packer N., Lisacek F. Glycomics@ExPASy: Bridging the gap // Molecular and Cellular Proteomics. – 2018.10.1074/mcp.RA118.000799. – C. ePub ahead of print.

108) Ceroni A., Maass K., Geyer H., Geyer R., Dell A., Haslam S. M. GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans // Journal of Proteome Research. – 2008. – Т. 7, № 4. – С. 1650-1659.

109) Maass K., Ranzinger R., Geyer H., von der Lieth C. W., Geyer R. "Glyco-peakfinder" - de novo composition analysis of glycoconjugates // Proteomics. – 2007. – Т. 7, № 24. – С. 4435-4444.

110) Irungu J., Go E. P., Dalpathado D. S., Desaire H. Simplification of mass spectral analysis of acidic glycopeptides using GlycoPep ID // Analytical Chemistry. – 2007. – Т. 79, № 8. – С. 3065-3074.

111) Pompach P., Chandler K. B., Lan R., Edwards N., Goldman R. Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search // Journal of Proteome Research. – 2012. – Т. 11, № 3. – С. 1728-1740.

112) Ozohanics O., Krenyacz J., Ludanyi K., Pollreisz F., Vekey K., Drahos L. GlycoMiner: a new software tool to elucidate glycopeptide composition // Rapid Communications in Mass Spectrometry. – 2008. – Т. 22, № 20. – С. 3245-3254.

113) Cooper C. A., Gasteiger E., Packer N. H. GlycoMod - a software tool for determining glycosylation compositions from mass spectrometric data // Proteomics. – 2001. – Т. 1, № 2. – С. 340-349.

114) Takahashi N., Kato K. GALXY(Glycoanalysis by the Three Axes of MS and Chromatography): a Web Application that Assists Structural Analyses of N-Glycans // Trends in Glycoscience and Glycotechnology. – 2003. – Т. 15, № 84. – С. 235-251.

115) Vranken W. F., Boucher W., Stevens T. J., Fogh R. H., Pajon A., Llinas M., Ulrich E. L., Markley J. L., Ionides J., Laue E. D. The CCPN data model for NMR spectroscopy: development of a software pipeline // Proteins. – 2005. – Т. 59, № 4. – С. 687-696.

116) Toukach F. V., Shashkov A. S. Computer-assisted structural analysis of regular glycopolymers on the basis of 13C NMR data // Carbohydrate Research. – 2001. – Т. 335, № 2. – С. 101-114.

117) Kapaev R. R., Egorova K. S., Toukach P. V. Carbohydrate structure generalization scheme for database-driven simulation of experimental observables, such as NMR chemical shifts // Journal of Chemical Information and Modeling. – 2014. – Т. 54, № 9. – С. 2594-2611.

118) Kapaev R. R., Toukach P. V. Improved carbohydrate structure generalization scheme for (1)H and (13)C NMR Simulations // Analytical Chemistry. – 2015. – Т. 87, № 14. – С. 7006-7010.

119) Kapaev R. R., Toukach P. V. Simulation of 2D NMR spectra of carbohydrates using GODESS software // Journal of Chemical Information and Modeling. – 2016. – Т. 56, № 6. – С. 1100-1104.

120) Kapaev R. R., Toukach P. V. GRASS: semi-automated NMR-based structure elucidation of saccharides // Bioinformatics. – 2018. – Т. 34, № 6. – С. 957-963.

121) Tanaka K., Aoki-Kinoshita K. F., Kotera M., Sawaki H., Tsuchiya S., Fujita N., Shikanai T., Kato M., Kawano S., Yamada I., Narimatsu H. WURCS: the Web3 unique representation of carbohydrate structures // Journal of Chemical Information and Modeling. – 2014. – Т. 54, № 6. – С. 1558-1566.

122) McNaught A. D. IUPAC and IUBMB. Joint Commission on Biochemical Nomenclature. Nomenclature of carbohydrates // Carbohydrate Research. – 1997. – Т. 297, № 1. – С. 1-92.

123) McNaught A. D. Nomenclature of carbohydrates (recommendations 1996) // Advances in carbohydrate chemistry and biochemistry. – 1997. – Т. 52. – С. 43-177.

124) Dalby A., Nourse J. G., Hounshell W. D., Gushurst A. K. I., Grier D. L., Leland B. A., Laufer J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited // Journal of Chemical Information and Computer Sciences. – 1992. – Т. 32, № 3. – С. 244-255.

125) Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // Journal of Chemical Information and Computer Sciences. – 1988. – Т. 28, № 1. – С. 31-36.

126) Heller S., McNaught A., Stein S., Tchekhovskoi D., Pletnev I. InChI - the worldwide chemical structure identifier standard // BMC Journal of Cheminformatics. – 2013. – Т. 5, № 1. – С. ID 7.

127) Sahoo S. S., Thomas C., Sheth A., Henson C., York W. S. GLYDE-an expressive XML standard for the representation of glycan structure // Carbohydrate Research. – 2005. – Т. 340, № 18. – С. 2802-2807.

128) Kikuchi N., Kameyama A., Nakaya S., Ito H., Sato T., Shikanai T., Takahashi Y., Narimatsu H. The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures // Bioinformatics. – 2005. – Т. 21, № 8. – С. 1717-1718.

129) Banin E., Neuberger Y., Altshuler Y., Halevi A., Inbar O., Nir D., Dukler A. A novel linear code nomenclature for complex carbohydrates // Trends in Glycoscience and Glycotechnology. – 2002. – Т. 14, № 77. – С. 127-137.

130) Bohne-Lang A., Lang E., Förster T., von der Lieth C.-W. LINUCS: LInear Notation for Unique description of Carbohydrate Sequences // Carbohydrate Research. – 2001. – Т. 336, № 1. – С. 1-11.

131) Ranzinger R., Kochut K. J., Miller J. A., Eavenson M., Lutteke T., York W. S. GLYDE-II: The GLYcan data exchange format // Perspectives in Science. – 2017. – Т. 11. – С. 24-30.

132) Harvey D. J., Merry A. H., Royle L., Campbell M. P., Dwek R. A., Rudd P. M. Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds // Proteomics. – 2009. – Т. 9, № 15. – С. 3796-801.

133) Varki A., Cummings R. D., Esko J. D., Freeze H. H., Stanley P., Marth J. D., Bertozzi C. R., Hart G. W., Etzler M. E. Symbol nomenclature for glycan representation // Proteomics. – 2009. – Т. 9, № 24. – С. 5398-5399.

134) Varki A., Cummings R. D., Aebi M., Packer N. H., Seeberger P. H., Esko J. D., Stanley P., Hart G., Darvill A., Kinoshita T., Prestegard J. J., Schnaar R. L., Freeze H. H., Marth J. D., Bertozzi C. R., Etzler M. E., Frank M., Vliegenthart J. F., Lutteke T., Perez S., Bolton E., Rudd P., Paulson J., Kanehisa M., Toukach P., Aoki-Kinoshita K. F., Dell A., Narimatsu H., York W., Taniguchi N., Kornfeld S. Symbol nomenclature for graphical representations of glycans // Glycobiology. – 2015. – Т. 25, № 12. – С. 1323-1324.

135) Herget S., von der Lieth C. W. Digital Representations of Oligo- and Polysaccharides // Bioinformatics for Glycobiology and Glycomics / von der Lieth C. W. и др., 2009. – Гл. 3, С. 49-68.

136) Lutteke T. Handling and conversion of carbohydrate sequence formats and monosaccharide notation // Glycoinformatics / Lütteke T., Frank M. – New York: Humana Press, 2015. – Гл. 4, С. 43-54.

137) York W. S., Kochut K. J., Miller J. A. Integration of Glycomics Knowledge and Data // Handbook of Glycomics / Cummings R. D., Pierce J. M. – San Diego: Academic Press, 2009. – Гл. 8, С. 177-195.

138) Damerell D., Ceroni A., Maass K., Ranzinger R., Dell A., Haslam S. M. The GlycanBuilder and GlycoWorkbench glycoinformatics tools: updates and new developments // Biological Chemistry. – 2012. – Т. 393, № 11. – С. 1357-1362.

139) Tsuchiya S., Aoki N. P., Shinmachi D., Matsubara M., Yamada I., Aoki-Kinoshita K. F., Narimatsu H. Implementation of GlycanBuilder to draw a wide variety of ambiguous glycans // Carbohydrate Research. – 2017. – Т. 445. – С. 104-116.

140) Cheng K., Zhou Y., Neelamegham S. DrawGlycan-SNFG: a robust tool to render glycans and glycopeptides with fragmentation information // Glycobiology. – 2017. – Т. 27, № 3. – С. 200-205.

141) Thieker D. F., Hadden J. A., Schulten K., Woods R. J. 3D implementation of the symbol nomenclature for graphical representation of glycans // Glycobiology. – 2016. – Т. 26, № 8. – С. 786-787.

142) Loss A., Bunsmann P., Bohne A., Loss A., Schwarzer E., Lang E., von der Lieth C. W. SWEET-DB: an attempt to create annotated data collections for carbohydrates // Nucleic Acids Research. – 2002. – Т. 30, № 1. – С. 405-408.

143) Essentials of glycobiology. – 2 изд.: Cold Spring Harbor Laboratory Press, 1999. – 784 с.

144) Kornfeld S., Li E., Tabas I. The synthesis of complex-type oligosaccharides. II. Characterization of the processing intermediates in the synthesis of the complex oligosaccharide units of the vesicular stomatitis virus G protein // Journal of Biological Chemistry. – 1978. – Т. 253, № 21. – С. 7771-7778.

145) Perez S., Aoki-Kinoshita K. F. Development of carbohydrate nomenclature and representation // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. – Japan: Springer, 2017. – Гл. 2, С. 7-25.

146) Chernyshov I. Y., Toukach P. V. REStLESS: automated translation of glycan sequences from residue-based notation to SMILES and atomic coordinates // Bioinformatics. – 2018. – Т. 34, № 15. – С. 2679-2681.

147) Perez S., Tubiana T., Imberty A., Baaden M. Three-dimensional representations of complex carbohydrates and polysaccharides--SweetUnityMol: a video game-based computer graphic software // Glycobiology. – 2015. – Т. 25, № 5. – С. 483-91.

148) Sehnal D., Grant O. C. Rapidly display glycan symbols in 3D structures: 3D-SNFG in LiteMol // Journal of Proteome Research. – 2018.10.1021/acs.jproteome.8b00473. – С. ePub ahead of print.

149) Humphrey W., Dalke A., Schulten K. VMD: Visual molecular dynamics // Journal of Molecular Graphics. – 1996. – Т. 14, № 1. – С. 33-38.

150) Burley S. K., Berman H. M., Kleywegt G. J., Markley J. L., Nakamura H., Velankar S. Protein Data Bank (PDB): the single global macromolecular structure archive // Protein Crystallography / Wlodawer A. и др. – New York, NY: Humana Press, 2017. – Гл. 26, С. 627-641.

151) Groom C. R., Bruno I. J., Lightfoot M. P., Ward S. C. The Cambridge Structural Database // cta crystallographica Section B, Structural science, crystal engineering and materials. – 2016. – Т. 72, № Pt 2. – С. 171-179.

152) Bhat T. N., Bourne P., Feng Z., Gilliland G., Jain S., Ravichandran V., Schneider B., Schneider K., Thanki N., Weissig H., Westbrook J., Berman H. M. The PDB data uniformity project // Nucleic Acids Research. – 2001. – Т. 29, № 1. – С. 214-218.

153) Lütteke T., von der Lieth C.-W. The protein data bank (PDB) as a versatile resource for glycobiology and glycomics // Biocatalysis and biotransformation. – 2009. – Т. 24, № 1-2. – С. 147-155.

154) Lutteke T., Frank M., von der Lieth C. W. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB // Nucleic Acids Research. – 2005. – Т. 33, № Database issue. – С. D242-D246.

155) Jo S., Im W. Glycan fragment database: a database of PDB-based glycan 3D structures // Nucleic Acids Research. – 2013. – Т. 41, № Database issue. – С. D470-D474.

156) Kerzmann A., Neumann D., Kohlbacher O. SLICK - scoring and energy functions for protein-carbohydrate interactions // Journal of Chemical Information and Modeling. – 2006. – Т. 46, № 4. – С. 1635-1642.

157) Kerzmann A., Fuhrmann J., Kohlbacher O., Neumann D. BALLDock/SLICK: a new method for protein-carbohydrate docking // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 8. – С. 1616-1625.

158) Case D. A., Cheatham T. E., 3rd, Darden T., Gohlke H., Luo R., Merz K. M., Jr., Onufriev A., Simmerling C., Wang B., Woods R. J. The Amber biomolecular simulation programs // Journal of Computational Chemistry. – 2005. – Т. 26, № 16. – С. 1668-1688.

159) Jo S., Song K. C., Desaire H., MacKerell A. D., Jr., Im W. Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins // Journal of Computational Chemistry. – 2011. – Т. 32, № 14. – С. 3135-3141.

160) Brooks B. R., Brooks C. L., 3rd, Mackerell A. D., Jr., Nilsson L., Petrella R. J., Roux B., Won Y., Archontis G., Bartels C., Boresch S., Caflisch A., Caves L., Cui Q., Dinner A. R., Feig M., Fischer S., Gao J., Hodoscek M., Im W., Kuczera K., Lazaridis T., Ma J., Ovchinnikov V., Paci E., Pastor R. W., Post C. B., Pu J. Z., Schaefer M., Tidor B., Venable R. M., Woodcock H. L., Wu X., Yang W., York D. M., Karplus M. CHARMM: the biomolecular simulation program // Journal of Computational Chemistry. – 2009. – Т. 30, № 10. – С. 1545-1614.

161) Bohne-Lang A., von der Lieth C. W. GlyProt: in silico glycosylation of proteins // Nucleic Acids Research. – 2005. – Т. 33, № Web Server issue. – С. W214-W219.

162) Lütteke T., Frank M., von der Lieth C.-W. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures // Carbohydrate Research. – 2004. – Т. 339, № 5. – С. 1015-1020.

163) Lütteke T., von der Lieth C. W. pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files // BMC Bioinformatics. – 2004. – Т. 5. – С. ID 69.

164) Bohne A., Lang E., von der Lieth C. W. SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides // Bioinformatics. – 1999. – Т. 15, № 9. – С. 767-768.

165) Kuttel M. M., Stahle J., Widmalm G. CarbBuilder: Software for building molecular models of complex oligo- and polysaccharide structures // Journal of Computational Chemistry. – 2016. – Т. 37, № 22. – С. 2098-105.

166) Toukach F. V., Ananikov V. P. Recent advances in computational predictions of NMR parameters for the structure elucidation of carbohydrates: methods and limitations // Chemical Society Reviews. – 2013. – Т. 42, № 21. – С. 8376-8415.

167) Adcock S. A., McCammon J. A. Molecular dynamics: survey of methods for simulating the activity of proteins // Chemical Reviews. – 2006. – Т. 106, № 5. – С. 1589-1615.

168) Re S., Nishima W., Miyashita N., Sugita Y. Conformational flexibility of *N*-glycans in solution studied by REMD simulations // Biophysical Reviews. – 2012. – Т. 4, № 3. – С. 179-187.

169) Wavefunction, Inc. *Spartan software*. URL: http://www.wavefun.com/products/spartan.html.

170) MOSCITO. 4th edition. / Paschek, Geiger A. – Dortmund, Germany: Department of Physical Chemistry University of Dortmund, 2002.

171) Möllhoff M., Sternberg U. Molecular mechanics with fluctuating atomic charges – a new force field with a semi-empirical charge calculation // Journal of Molecular Modeling. – 2001. – Т. 7. – С. 90-102.

172) Gaussian Inc. *Gaussian*. URL: http://www.gaussian.com.

173) Schmidt M. W., Baldridge K. K., Boatz J. A., Elbert S. T., Gordon M. S., Jensen J. H., Koseki S., Matsunaga N., Nguyen K. A., Su S., Windus T. L., Dupuis M., Montgomery J. A. General atomic and molecular electronic structure system // Journal of Computational Chemistry. – 1993. – Т. 14, № 11. – С. 1347-1363.

174) Rahal-Sekkal M., Sekkal N., Kleb D. C., Bleckmann P. Structures and energies of D-galactose and galabiose conformers as calculated by ab initio and semiempirical methods // Journal of Computational Chemistry. – 2003. – Т. 24, № 7. – С. 806-818.

175) Stewart J. J. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements // Journal of Molecular Modeling. – 2007. – Т. 13, № 12. – С. 1173-1213.

176) Introduction to computational chemistry. / Jensen F. – 2 изд.: John Wiley & Sons Ltd., 2007. – 620 с.

177) Zhao Y., Truhlar D. G. Density functionals with broad applicability in chemistry // Accounts of Chemical Research. – 2008. – Т. 41, № 2. – С. 157-167.

178) Becke A. D. Density-functional exchange-energy approximation with correct asymptotic behavior // Physical Review. A. – 1988. – Т. 38, № 6. – С. 3098-3100.

179) Perdew J. P., Wang Y. Accurate and simple analytic representation of the electron-gas correlation energy // Physical Review. B. – 1992. – Т. 45, № 23. – С. 13244-13249.

180) Svensson M., Humbel S., Froese R. D. J., Matsubara T., Sieber S., Morokuma K. ONIOM: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions // Journal of Physical Chemistry. – 1996. – T. 100. – C. 19357-19363.

181) Lodola A., Woods C. J., Mulholland A. J. Applications and advances of QM/MM methods in computational enzymology // Annual Reports in Computational Chemistry. – 2008. – T. 4. – C. 155-169.

182) Imberty A., Pérez S. Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations // Chemical Reviews. – 2000. – T. 100, № 12. – C. 4567-4588.

183) Miertuš S., Tomasi J. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes // Chemical Physics. – 1982. – T. 65, № 2. – C. 239-245.

184) Cossi M., Rega N., Scalmani G., Barone V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model // Journal of Computational Chemistry. – 2003. – T. 24, № 6. – C. 669-681.

185) Marenich A. V., Cramer C. J., Truhlar D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions // Journal of Physical Chemistry. B. – 2009. – T. 113, № 18. – C. 6378-6396.

186) Klamt A., Schüürmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient // Journal of the Chemical Society. Perkin transactions 2. – 1993.10.1039/p29930000799 № 5. – C. 799-805.

187) Frank M., Collins P. M., Peak I. R., Grice I. D., Wilson J. C. An unusual carbohydrate conformation is evident in Moraxella catarrhalis oligosaccharides // Molecules. – 2015. – T. 20, № 8. – C. 14234-14253.

188) Lipkind G. M., Shashkov A. S., Knirel Y. A., Vinogradov E. V., Kochetkov N. K. A computer-assisted structural analysis of regular polysaccharides on the basis of $^{13}$C-n.m.r. data // Carbohydrate Research. – 1988. – T. 175, № 1. – C. 59-75.

189) Jansson P.-E., Kenne L., Widmalm G. Computer-assisted structural analysis of polysaccharides with an extended version of casper using 1H- and 13C-n.m.r. data // Carbohydrate Research. – 1989. – Т. 188. – С. 169-191.

190) Smurnyy Y. D., Blinov K. A., Churanova T. S., Elyashberg M. E., Williams A. J. Toward more reliable $^{13}C$ and $^{1}H$ chemical shift prediction: a systematic comparison of neural-network and least-squares regression based approaches // Journal of Chemical Information and Modeling. – 2008. – Т. 48, № 1. – С. 128-134.

191) Bremser W. Hose - a novel substructure code // Analytica chimica acta. – 1978. – Т. 103, № 4. – С. 355-365.

192) Sasaki R. R., Lefebvre B. A. On the importance of structure stereochemical markers in 13C NMR predictions – Burlington, VT, 2006. –.

193) Trepalin S. V., Yarkov A. V., Dolmatova L. M., Zefirov N. S., Finch S. A. E. Windat - an NMR database compilation tool, user-interface, and spectrum libraries for personal computers // Journal of Chemical Information and Computer Sciences. – 1995. – Т. 35, № 3. – С. 405-411.

194) Steinbeck C., Krause S., Kuhn S. NMRShiftDB-constructing a free chemical information system with open-source components // Journal of Chemical Information and Computer Sciences. – 2003. – Т. 43, № 6. – С. 1733-1739.

195) Neural Networks in Analytical Chemistry. Artificial neural networks: methods and applications. / Jalali-Heravi M.; Под ред. Livingstone D. J., 2008. Artificial neural networks: methods and applications.

196) Radomski J. P., van Halbeek H., Meyer B. Neural network–based recognition of oligosaccharide 1H-NMR spectra // Nature Structural Biology. – 1994. – Т. 1. – С. 217-218.

197) Aires-de-Sousa J., Hemmer M. C., Gasteiger J. Prediction of $^{1}H$ NMR chemical shifts using neural networks // Analytical Chemistry. – 2002. – Т. 74, № 1. – С. 80-90.

198) Gerbst A. G., Grachev A. A., Ustuzhanina N. E., Nifantiev N. E., Vyboichtchik A. A., Shashkov A. S., Usov A. I. Application of artificial neural networks for analysis of $^{13}C$ NMR spectra of fucoidans // Journal of Carbohydrate Chemistry. – 2010. – Т. 29, № 2. – С. 92-102.

199) McIntyre M. K., Small G. W. Carbon-13 nuclear magnetic resonance spectrum simulation methodology for the structure elucidation of carbohydrates // Analytical Chemistry. – 1987. – Т. 59, № 14. – С. 1805-1811.

200) Clouser D. L., Jurs P. C. Simulation of the $^{13}C$ nuclear magnetic resonance spectra of ribonucleosides using multiple linear regression analysis and neural networks // Journal of Chemical Information and Computer Sciences. – 1996. – Т. 36, № 2. – С. 168-172.

201) Mitchell B. E., Jurs P. C. Computer assisted simulation of $^{13}C$ nuclear magnetic spectra of monosaccharides // Journal of Chemical Information and Computer Sciences. – 1996. – Т. 36, № 1. – С. 58-64.

202) Abraham R. J., Byrne J. J., Griffiths L., Koniotou R. $^{1}H$ chemical shifts in NMR: Part 22-Prediction of the $^{1}H$ chemical shifts of alcohols, diols and inositols in solution, a conformational and solvation investigation // Magnetic Resonanace in Chemistry. – 2005. – Т. 43, № 8. – С. 611-624.

203) Lundborg M., Widmalm G. Structural analysis of glycans by NMR chemical shift prediction // Analytical Chemistry. – 2011. – Т. 83, № 5. – С. 1514−1517.

204) Jansson P. E., Kenne L., Widmalm G. CASPER: a computer program used for structural analysis of carbohydrates // Journal of Chemical Information and Computer Sciences. – 1991. – Т. 31, № 4. – С. 508–516.

205) Jansson P. E., Stenutz R., Widmalm G. Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER // Carbohydrate Research. – 2006. – Т. 341, № 8. – С. 1003-1010.

206) Sternberg U. Theory of the influence of the second co-ordination sphere on the chemical shift // Molecular Physics. – 1988. – Т. 63, № 2. – С. 249-267.

207) Sternberg U., Prieß W. New semi-empirical approach for the calculation of $^{13}C$ chemical-shift tensors // Journal of Magnetic Resonanace. – 1997. – Т. 125, № 1. – С. 8-19.

208) Ditchfield R. Self-consistent perturbation theory of diamagnetism // Molecular Physics. – 1974. – Т. 27, № 4. – С. 789-807.

209) Schindler M., Kutzelnigg W. Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules // Journal of Chemical Physics. – 1982. – Т. 76, № 4. – С. 1919-1933.

210) Malkin V. G., Malkina O. L., Casida M. E., Salahub D. R. Nuclear magnetic resonance shielding tensors calculated with a aum-over-states density functional perturbation theory // Journal of American Chemical Society. – 1994. – T. 116, № 13. – C. 5898-5908.

211) Hansen A. E., Bouman T. D. Localized orbital/local origin method for calculation and analysis of NMR shieldings. Applications to $^{13}C$ shielding tensors // Journal of Chemical Physics. – 1985. – T. 82, № 11. – C. 5035-5047.

212) Pfrommer B. G., Demmel J., Simon H. Unconstrained energy functionals for electronic structure calculations // Journal of Computational Physics. – 1999. – T. 150, № 1. – C. 287-298.

213) Friedrich K., Seifert G., Großmann G. Nuclear magnetic shielding in molecules. The application of GIAO's in LCAO-Xα-calculation // Zeitschrift für Physik D. Atoms, Molecules and Clusters. – 1990. – T. 17, № 1. – C. 45-46.

214) Cheeseman J. R., Trucks G. W., Keith T. A., Frisch M. J. A comparison of models for calculating nuclear magnetic resonance shielding tensors // Journal of Chemical Physics. – 1996. – T. 104, № 14. – C. 5497-5509.

215) Lii J. H., Ma B. Y., Allinger N. L. Importance of selecting proper basis set in quantum mechanical studies of potential energy surfaces of carbohydrates // Journal of Computational Chemistry. – 1999. – T. 20, № 15. – C. 1593-1603.

216) Pankratyev E. Y., Tulyabaev A. R., Khalilov L. M. How reliable are GIAO calculations of 1H and 13C NMR chemical shifts? A statistical analysis and empirical corrections at DFT (PBE/3z) level // Journal of Computational Chemistry. – 2011. – T. 32, № 9. – C. 1993-1997.

217) Laikov D. N., Ustynyuk Y. A. PRIRODA_04: a quantum_chemical program suite. New possibilities in the study of molecular systems with the application of parallel computing // Russ Chem Bull Int Ed. – 2005. – T. 54, № 3. – C. 820-826.

218) Tafazzoli M., Ghiasi M. Structure and conformation of α-, β- and γ-cyclodextrin in solution: Theoretical approaches and experimental validation // Carbohydrate Polymers. – 2009. – T. 78, № 1. – C. 10-15.

219) Sergeyev I., Moyna G. Determination of the three-dimensional structure of oligosaccharides in the solid state from experimental 13C NMR data and ab initio chemical shift surfaces // Carbohydrate Research. – 2005. – T. 340, № 6. – C. 1165-1174.

220) Roslund M. U., Tahtinen P., Niemitz M., Sjoholm R. Complete assignments of the (1)H and (13)C chemical shifts and J(H,H) coupling constants in NMR spectra of D-glucopyranose and all D-glucopyranosyl-D-glucopyranosides // Carbohydrate Research. – 2008. – Т. 343, № 1. – С. 101-112.

221) Esrafili M. D., Elmi F., Hadipour N. L. Density functional theory investigation of hydrogen bonding effects on the oxygen, nitrogen and hydrogen electric field gradient and chemical shielding tensors of anhydrous chitosan crystalline structure // Journal of Physical Chemistry. A. – 2007. – Т. 111, № 5. – С. 963-970.

222) Bagno A., Rastrelli F., Saielli G. Prediction of the $^{1}H$ and $^{13}C$ NMR spectra of alpha-D-glucose in water by DFT methods and MD simulations // Journal of Organic Chemistry. – 2007. – Т. 72, № 19. – С. 7373-7381.

223) Kasat R. B., Wang N. H., Franses E. I. Effects of backbone and side chain on the molecular environments of chiral cavities in polysaccharide-based biopolymers // Biomacromolecules. – 2007. – Т. 8, № 5. – С. 1676-1685.

224) Suzuki S., Horii F., Kurosu H. Theoretical investigations of $^{13}C$ chemical shifts in glucose, cellobiose, and native cellulose by quantum chemistry calculations // Journal of Molecular Structure. – 2009. – Т. 921, № 1-3. – С. 219-226.

225) Yates J. R., Pham T. N., Pickard C. J., Mauri F., Amado A. M., Gil A. M., Brown S. P. An investigation of weak CH...O hydrogen bonds in maltose anomers by a combination of calculation and experimental solid-state NMR spectroscopy // Journal of American Chemical Society. – 2005. – Т. 127, № 29. – С. 10216-10220.

226) Sefzik T. H., Turco D., Iuliucci R. J., Facelli J. C. Modeling NMR chemical shift: A survey of density functional theory approaches for calculating tensor properties // Journal of Physical Chemistry. A. – 2005. – Т. 109, № 6. – С. 1180-1187.

227) Helgaker T., Jaszuński M., Pecul M. The quantum-chemical calculation of NMR indirect spin-spin coupling constants // Prog Nucl Magn Reson Spectrosc. – 2008. – Т. 53. – С. 249-268.

228) Ramsey N. F. Electron coupled interactions between nuclearspins in molecules // Phys Rev. – 1953. – Т. 91, № 2. – С. 303-307.

229) Stenutz R., Weintraub A., Widmalm G. The structures of Escherichia coli O-polysaccharide antigens // FEMS Microbiology Reviews. – 2006. – Т. 30, № 3. – С. 382-403.

230) Vollmer W., Blanot D., de Pedro M. A. Peptidoglycan structure and architecture // FEMS Microbiology Reviews. – 2008. – Т. 32, № 2. – С. 149-167.

231) Bishop J. R., Gagneux P. Evolution of carbohydrate antigens--microbial forces shaping host glycomes? // Glycobiology. – 2007. – Т. 17, № 5. – С. 23R-34R.

232) Gagneux P., Varki A. Evolutionary considerations in relating oligosaccharide diversity to biological function // Glycobiology. – 1999. – Т. 9, № 8. – С. 747-755.

233) Boulnoisl G. J., Jann K. Bacterial polysaccharide capsule synthesis, export and evolution of structural diversity // Molecular Microbiology. – 1989. – Т. 3, № 12. – С. 1819-1823.

234) Gagneux P., Aebi M., Varki A. Evolution of Glycan Diversity // Essentials of Glycobiology / Varki A. и др. – New-York: Cold Spring Harbor Laboratory Press, 2017. – Гл. 20, С. 253-264.

235) Seeberger P. H. The logic of automated glycan assembly // Accounts of Chemical Research. – 2015. – Т. 48, № 5. – С. 1450-1463.

236) Hahm H. S., Schlegel M. K., Hurevich M., Eller S., Schuhmacher F., Hofmann J., Pagel K., Seeberger P. H. Automated glycan assembly using the Glyconeer 2.1 synthesizer // Proceedings of the National Academy of Sciences of the U.S.A. – 2017. – Т. 114, № 17. – С. E3385-E3389.

237) Pardo-Vargas A., Delbianco M., Seeberger P. H. Automated glycan assembly as an enabling technology // Current Opinions in Chemical Biology. – 2018. – Т. 46. – С. 48-55.

238) Woese C. R. There must be a prokaryote somewhere: microbiology's search for itself // Microbiological Reviews. – 1994. – Т. 58. – С. 1-9.

239) Woese C. R., Kandler O., Wheelis M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya // Proceedings of the National Academy of Sciences of the U.S.A. – 1990. – Т. 87, № 12. – С. 4576-4579.

240) Fitz-Gibbon S. T., House C. H. Whole genome-based phylogenetic analysis of free-living microorganisms // Nucleic Acids Research. – 1999. – Т. 27, № 21. – С. 4218-4222.

241) Bansal A. K., Meyer T. E. Evolutionary analysis by whole-genome comparisons // Journal of Bacteriology. – 2002. – Т. 184, № 8. – С. 2260-2272.

242) Osawa S., Jukes T. H., Watanabe K., Muto A. Recent evidence for evolution of the genetic code // Microbiological Reviews. – 1992. – Т. 56, № 1. – С. 229-264.

243) Sankoff D., Leduc G., Antoine N., Paquin B., Lang B. F., Cedergren R. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome // Proceedings of the National Academy of Sciences of the U.S.A. – 1992. – Т. 89, № 14. – С. 6575-6579.

244) Wolf Y. I., Brenner S. E., Bash P. A., Koonin E. V. Distribution of protein folds in the three superkingdoms of life // Genome Research. – 1999. – Т. 9, № 1. – С. 17-26.

245) Lin J., Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels // Genome Research. – 2000. – Т. 10, № 6. – С. 808-818.

246) Tekaia F., Yeramian E., Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis // Gene. – 2002. – Т. 297, № 1-2. – С. 51-60.

247) Aguilar D., Aviles F. X., Querol E., Sternberg M. J. Analysis of phenetic trees based on metabolic capabilites across the three domains of life // Journal of Molecular Biology. – 2004. – Т. 340, № 3. – С. 491-512.

248) Jeong H., Tombor B., Albert R., Oltvai Z. N., Barabasi A. L. The large-scale organization of metabolic networks // Nature. – 2000. – Т. 407, № 6804. – С. 651-654.

249) Ravasz E., Somera A. L., Mongru D. A., Oltvai Z. N., Barabasi A. L. Hierarchical organization of modularity in metabolic networks // Science. – 2002. – Т. 297, № 5586. – С. 1551-1555.

250) Rigden D. J., Fernandez-Suarez X. M., Galperin M. Y. The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection // Nucleic Acids Research. – 2016. – Т. 44. – С. D1-D6.

251) Artemenko N. V., McDonald A. G., Davey G. P., Rudd P. M. Databases and tools in glycobiology // Therapeutic Proteins / Voynov V., Caravella J. – Totowa, NJ: Humana Press, 2012. – Гл. 21, С. 325-350.

252) Zhulin I. B. Databases for Microbiologists // Journal of Bacteriology. – 2015. – Т. 197, № 15. – С. 2458-2467.

253) Yamada K., Kakehi K. Recent advances in the analysis of carbohydrates for biomedical use // Journal of Pharmaceutical and Biomedical Analysis. – 2011. – Т. 55, № 4. – С. 702-727.

254) Egorova K. S., Kalinchuk N. A., Knirel Y. A., Toukach P. V. Carbohydrate Structure Database (CSDB): new features // Russian Chemical Bulletin. – 2015. – Т. 64, № 5. – С. 1205-1210.

255) Toukach P., Joshi H. J., Ranzinger R., Knirel Y., von der Lieth C. W. Sharing of worldwide distributed carbohydrate-related digital resources: online connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de // Nucleic Acids Research. – 2007. – Т. 35, № Database issue. – С. D280-D286.

256) Egorova K. S., Toukach P. V. Expansion of coverage of Carbohydrate Structure Database (CSDB) // Carbohydrate Research. – 2014. – Т. 389. – С. 112-114.

257) Toukach P. V., Egorova K. S. Bacterial, Plant, and Fungal Carbohydrate Structure Database (CSDB) // Glycoscience: Biology and Medicine / Endo T. и др. – Japan: Springer, 2015. – Гл. 29, С. 241-250.

258) Sevinc A. Web of science: a unique method of cited reference searching // Journal of the National Medical Association. – 2004. – Т. 96, № 7. – С. 980-983.

259) Wiggins E. V. The NLM current catalog // Bulletin of the Medical Library Association. – 1969. – Т. 57, № 1. – С. 36-40.

260) Federhen S. The NCBI Taxonomy database // Nucleic Acids Research. – 2012. – Т. 40, № Database issue. – С. D136-D143.

261) Kim S., Thiessen P. A., Bolton E. E., Chen J., Fu G., Gindulyte A., Han L., He J., He S., Shoemaker B. A., Wang J., Yu B., Zhang J., Bryant S. H. PubChem Substance and Compound databases // Nucleic Acids Research. – 2016. – Т. 44, № D1. – С. D1202-D1213.

262) Weininger D. SMILES-A Language for Molecules and Reactions // Handbook of Chemoinformatics: From Data to Knowledge / Gasteiger J. WILEY-VCH Verlag GmbH & Co., 2003. – Гл. II.3, С. 80-102.

263) Hanson R. M. Jmol SMILES and Jmol SMARTS: specifications and applications // Journal of cheminformatics. – 2016. – Т. 8. – С. ID 50.

264) Halgren T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 // Journal of Computational Chemistry. – 1996. – Т. 17, № 5-6. – С. 490-519.

265) Karaboga D., Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm // Journal of Global Optimization. – 2007. – Т. 39, № 3. – С. 459-471.

266) An introduction to error analysis: the study of uncertainties in physical measurements. / Taylor J. R. – 2 изд. – Sausalito: Calif University Science Books 1997.

267) Buckingham A. D., Schaefer T., Schneider W. G. Solvent Effects in Nuclear Magnetic Resonance Spectra // The Journal of Chemical Physics. – 1960. – Т. 32, № 4. – С. 1227-1233.

268) Tynkkynen T., Tiainen M., Soininen P., Laatikainen R. From proton nuclear magnetic resonance spectra to pH. Assessment of 1H NMR pH indicator compound set for deuterium oxide solutions // Analytica Chimica Acta. – 2009. – Т. 648, № 1. – С. 105-112.

269) Raiford D. S., Fisk C. L., Becker E. D. Calibration of methanol and ethylene glycol nuclear magnetic resonance thermometers // Analytical Chemistry. – 2002. – Т. 51, № 12. – С. 2050-2051.

270) Markley J. L., Bax A., Arata Y., Hilbers C. W., Kaptein R., Sykes B. D., Wright P. E., Wüthrich K. Recommendations for the presentation of NMR structures of proteins and nucleic acids – IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data vases of protein and nucleic acid structures determined by NMR spectroscopy // Journal of Biomolecular NMR. – 1998. – Т. 12, № 1. – С. 1-23.

271) Morcombe C. R., Zilm K. W. Chemical shift referencing in MAS solid state NMR // Journal of Magnetic Resonance. – 2003. – Т. 162, № 2. – С. 479-486.

272) Katzenellenbogen E., Kocharova N. A., Toukach P. V., Gorska S., Korzeniowska-Kowal A., Bogulska M., Gamian A., Knirel Y. A. Structure of an abequose-containing O-polysaccharide from Citrobacter freundii O22 strain PCM 1555 // Carbohydrate Research. – 2009. – Т. 344, № 13. – С. 1724-1728.

273) Sidorczyk Z., Zych K., Toukach F. V., Arbatsky N. P., Shashkov A. S., Knirel Y. A. Structure of the O-polysaccharide and classification ofProteus mirabilisstrain G1 inProteusserogroup O3 // European Journal of Biochemistry. – 2002. – Т. 269, № 5. – С. 1406-1412.

274) Kuhn H., Meier-Dieter U., Hubert M. ECA, the enterobacterial common antigen // FEMS Microbiology Letters. – 1988. – Т. 54, № 3. – С. 195-222.

275) Davies A. N., Lampen P. JCAMP-DX for NMR // Applied Spectroscopy. – 1993. – T. 47, № 8. – C. 1093-1099.

276) Claridge T. MNova: NMR data processing, analysis, and prediction software // Journal of Chemical Information and Modeling. – 2009. – T. 49, № 4. – C. 1136-1137.

277) Ulrich E. L., Akutsu H., Doreleijers J. F., Harano Y., Ioannidis Y. E., Lin J., Livny M., Mading S., Maziuk D., Miller Z., Nakatani E., Schulte C. F., Tolmie D. E., Kent Wenger R., Yao H., Markley J. L. BioMagResBank // Nucleic Acids Research. – 2008. – T. 36, № D. – C. D402-408.

278) Bush C. A. High resolution NMR in the determination of structure in complex carbohydrates // Bulletin of Magnetic Resonance. – 1988. – T. 10, № 3/4. – C. 73-95.

279) Coxon B. Developments in the Karplus equation as they relate to the NMR coupling constants of carbohydrates // Advances in Carbohydrate Chemistry and Biochemistry. – 2009. – T. 62. – C. 17-82.

280) Duus J. Ø., Gotfredsen C. H., Bock K. Carbohydrate Structural Determination by NMR Spectroscopy: Modern Methods and Limitations // Chemical Reviews. – 2000. – T. 100, № 12. – C. 4589-4614.

281) Ritchie R. G. S., Cyr N., Korsch B., Koch H. J., Perlin A. S. Carbon-13 chemical shifts of furanosides and cyclopentanols. Configurational and conformational influences // Canadian Journal of Chemistry. – 1975. – T. 53, № 10. – C. 1424-1433.

282) Bock K., Pedersen C. Carbon-13 nuclear magnetic resonance spectroscopy of monosaccharides // Advances in Carbohydrate Chemistry and Biochemistry / Tipson R. S., Horton D. Academic Press, 1983, C. 27-66.

283) Waeghe T. J., Darvill A. G., McNeil M., Albersheim P. Determination, by methylation analysis, of the glycosyl-linkage compositions of microgram quantities of complex carbohydrates // Carbohydrate Research. – 1983. – T. 123, № 2. – C. 281-304.

284) Leontein K., Lindberg B., Lönngren J. Assignment of absolute configuration of sugars by g.l.c. of their acetylated glycosides formed from chiral alcohols // Carbohydrate Research. – 1978. – T. 62, № 2. – C. 359-362.

285) Claridge T. D. W. One-Dimensional Techniques // High-Resolution NMR Techniques in Organic Chemistry / Claridge T. D. W. – Boston: Elsevier, 2016. – Гл. 4, C. 133-169.

286) Stenutz R. Automatic Spectrum Interpretation Based on Increment Rules: CASPER // Bioinformatics for Glycobiology and Glycomics: An Introduction / von der Lieth C.-W. и др. John Wiley & Sons, Ltd, 2009. – Гл. 16, С. 311-320.

287) Weintraub A. Immunology of bacterial polysaccharide antigens // Carbohydrate Research. – 2003. – Т. 338, № 23. – С. 2539-2547.

288) Hedlund M., Tangvoranuntakul P., Takematsu H., Long J. M., Housley G. D., Kozutsumi Y., Suzuki A., Wynshaw-Boris A., Ryan A. F., Gallo R. L., Varki N., Varki A. N-glycolylneuraminic acid deficiency in mice: implications for human biology and evolution // Molecular and Cellular Biology. – 2007. – Т. 27, № 12. – С. 4340-4346.

289) Hoare A., Bittner M., Carter J., Alvarez S., Zaldivar M., Bravo D., Valvano M. A., Contreras I. The outer core lipopolysaccharide of Salmonella enterica serovar Typhi is required for bacterial entry into epithelial cells // Infection and Immunity. – 2006. – Т. 74, № 3. – С. 1555-1564.

290) He X. M., Liu H. W. Formation of unusual sugars: mechanistic studies and biosynthetic applications // Annual Reviews in Biochemistry. – 2002. – Т. 71. – С. 701-754.

291) Kawagishi S., Araki Y., Ito E. Bacillus cereus autolytic endoglucosaminidase active on cell wall peptidoglycan with N-unsubstituted glucosamine residues // Journal of Bacteriology. – 1980. – Т. 141, № 1. – С. 137-143.

292) Fusco P. C., Farley E. K., Huang C. H., Moore S., Michon F. Protective meningococcal capsular polysaccharide epitopes and the role of O acetylation // Clinical Vaccine Immunology. – 2007. – Т. 14, № 5. – С. 577-584.

293) Strecker G., Herlant-Peers M. C., Fournet B., Montreul J. Structure of seven oligosaccharides excreted in the urine of a patient with Sandhoff's disease (GM2 gangliosidosis-variant O) // European Journal of Biochemistry. – 1977. – Т. 81, № 1. – С. 165-171.

294) Hamming R. W. Error detecting and error correcting codes // Bell Syst. Tech. J. – 1950. – Т. 29, № 2. – С. 147—160.

295) Cardona G., Rosselló F., Valiente G. Extended Newick: it is time for a standard representation of phylogenetic networks // BMC Bioinformatics. – 2008. – Т. 9, № 1. – С. ID 532.

296) Maddison D. R., Swofford D. L., Maddison W. P., Cannatella D. Nexus: an extensible file format for systematic information // Systematic Biology. – 1997. – Т. 46, № 4. – С. 590-621.

297) Pace N. R., Olsen G. J., Woese C. R. Ribosomal RNA phylogeny and the primary lines of evolutionary descent // Cell. – 1986. – Т. 45, № 3. – С. 325-326.

298) Doolittle W. F. Phylogenetic Classification and the Universal Tree // Science. – 1999. – Т. 284, № 5423. – С. 2124-2128.

299) Andam C. P., Williams D., Gogarten J. P. Natural taxonomy in light of horizontal gene transfer // Biology & Philosophy. – 2010. – Т. 25, № 4. – С. 589-602.

300) Aoki-Kinoshita K. F., Sawaki H., An H. J., Campbell M., Cao Q., Cummings R., Hsu D. K., Kato M., Kawasaki T., Khoo K. H., Kim J., Kolarich D., Li X., Liu M., Matsubara M., Okuda S., Packer N. H., Ranzinger R., Shen H., Shikanai T., Shinmachi D., Toukach P., Yamada I., Yamaguchi Y., Yang P., Ying W., Yoo J. S., Zhang Y., Zhang Y., Narimatsu H. The Fifth ACGG-DB Meeting Report: Towards an International Glycan Structure Repository // Glycobiology. – 2013. – Т. 23, № 12. – С. 1422-1424.

301) Paskin N. Digital Object Identifiers for scientific data // Data Science Journal. – 2005. – Т. 4. – С. 12-20.

302) editorial T. L. I. D. ICD-11: in praise of good data // The Lancet Infectious Diseases. – 2018. – Т. 18, № 8. – С. 813.

303) Baumann N. How to use the medical subject headings (MeSH) // International Journal of Clinical Practice. – 2016. – Т. 70, № 2. – С. 171-174.

304) Fielding R. Architectural styles and the design of network-based software architectures: PhD dissertation; University of California. – Irvine, 2000.

305) Alocci D., Suchánková P., Costa R., Hory N., Mariethoz J., Svobodová Vařeková R., Toukach P., Lisacek F. SugarSketcher: quick and intuitive online glycan drawing // MDPI Molecules. – 2018. – Т. 23, № 12. – С. ID 3206.

306) Antezana E., Kuiper M., Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies // Briefings in Bioinformatics. – 2009. – Т. 10, № 4. – С. 392-407.

307) Wu H., Yamaguchi A. Semantic Web technologies for the big data in life sciences // BioScience Trends. – 2014. – Т. 8, № 4. – С. 192-201.

308) Aoki-Kinoshita K. F., Bolleman J., Campbell M. P., Kawano S., Kim J. D., Lutteke T., Matsubara M., Okuda S., Ranzinger R., Sawaki H., Shikanai T., Shinmachi D., Suzuki Y., Toukach P., Yamada I., Packer N. H., Narimatsu H. Introducing glycomics data into the Semantic Web // Journal of Biomedical Semantics. – 2013. – Т. 4, № 1. – С. ID 39.

309) Wollbrett J., Larmande P., de Lamotte F., Ruiz M. Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases // BMC bioinformatics. – 2013. – Т. 14. – С. ID 126.

310) O'Boyle N. M., Banck M., James C. A., Morley C., Vandermeersch T., Hutchison G. R. Open Babel: An open chemical toolbox // Journal of cheminformatics. – 2011. – Т. 3. – С. ID 33.

311) Herraez A. Biomolecules in the computer: Jmol to the rescue // Biochemistry and Molecular Biology Education. – 2006. – Т. 34, № 4. – С. 255-261.

312) Rackers J. A., Wang Z., Lu C., Laury M. L., Lagardere L., Schnieders M. J., Piquemal J. P., Ren P., Ponder J. W. Tinker 8: Software Tools for Molecular Design // Journal of Chemical Theory and Computation. – 2018.10.1021/acs.jctc.8b00529.

313) Allinger N. L., Yuh Y. H., Lii J. H. Molecular Mechanics. The MM3 Force Field for Hydrocarbons // J Am Chem Soc. – 1989. – Т. 1, № 23. – С. 8551-8566.

314) Frisch M. J., Trucks G. W., Schlegel H. B., Scuseria G. E., Robb M. A., Cheeseman J. R., Montgomery J. A., Jr., Vreven T., Kudin K. N., Burant J. C., Millam J. M., Iyengar S. S., Tomasi J., Barone V., Mennucci B., Cossi M., Scalmani G., Rega N., Petersson G. A., Nakatsuji H., Hada M., Ehara M., Toyota K., Fukuda R., Hasegawa J., Ishida M., Nakajima T., Honda Y., Kitao O., Nakai H., Klene M., Li X., Knox J. E., Hratchian H. P., Cross J. B., Adamo C., Jaramillo J., Gomperts R., Stratmann R. E., Yazyev O., Austin A. J., Cammi R., Pomelli C., Ochterski J. W., Ayala P. Y., Morokuma K., Voth G. A., Salvador P., Dannenberg J. J., Zakrzewski V. G., Dapprich S., Daniels A. D., Strain M. C., Farkas O., Malick D. K., Rabuck A. D., Raghavachari K., Foresman J. B., Ortiz J. V., Cui Q., Baboul A. G., Clifford S., Cioslowski J., Stefanov B. B., Liu G., Liashenko A., Piskorz P., Komaromi I., Martin R. L., Fox D. J., Keith T., Al-Laham M. A., Peng C. Y., Nanayakkara A., Challacombe M., Gill P. M. W., Johnson B., Chen W., Wong M. W., Gonzalez C., Pople J. A. // Book / Editor. – Wallingford CT.: Gaussian Inc., 2004.

315) Ihaka R., Gentleman R. R: a language for data analysis and graphics // Journal of Computational and Graphical Statistics. – 1996. – Т. 5, № 3. – С. 299-314.

316) Paradis E., Claude J., Strimmer K. APE: Analyses of phylogenetics and evolution in R language // Bioinformatics. – 2004. – Т. 20, № 2. – С. 289-290.

317) Murtagh F. Complexities of hierarchic clustering algorithms: the state of the art // Computational Statistics Quarterly. – 1984. – Т. 1. – С. 101—113.

318) Murtagh F., Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? // J. Classif. – 2014. – Т. 31, № 3. – С. 274—295.

319) Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data // Mol. Biol. Evol. – 1997. – Т. 14, № 7. – С. 685—695.

320) Desper R., Gascuel O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle // J. Comput. Biol. – 2002. – Т. 9, № 5. – С. 687—705.

321) Nye T. M., Lio P., Gilks W. R. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees // Bioinformatics. – 2006. – Т. 22, № 1. – С. 117—119.

322) Letunic I., Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy // Nucleic Acids Research. – 2011. – Т. 39, № Web Server issue. – С. W475-W478.

323) Aoki-Kinoshita K. F., Sawaki H., An H. J., Cho J. W., Hsu D., Kato M., Kawano S., Kawasaki T., Khoo K. H., Kim J., Kim J. D., Li X., Lutteke T., Okuda S., Packer N. H., Paulson J. C., Raman R., Ranzinger R., Shen H., Shikanai T., Yamada I., Yang P., Yamaguchi Y., Ying W., Yoo J. S., Zhang Y., Narimatsu H. The Third ACGG-DB Meeting Report: Towards an international collaborative infrastructure for glycobioinformatics // Glycobiology. – 2013. – Т. 23, № 2. – С. 144-146.

324) Aoki-Kinoshita K. F. Glycoinformatics: overview // Glycoscience: Biology and Medicine / Endo T. и др. – Japan: Springer, 2015. – Гл. 22, С. 185-192.

325) Campbell M. P., Lisacek F., Wilkins M. R., Rudd P. M., Kolarich D., Hayes C. A., Karlsson N. G., Packer N. H. Linking glycomics repositories with data capture // Cracking the sugar code by navigating the glycospace / Под ред. Hicks M. G., Kettner C. – Potsdam, Germany: Logos Verlag Berlin GmbH, 2011. – С. 191-206.

326) Walsh I., Zhao S., Campbell M., Taron C. H., Rudd P. M. Quantitative profiling of glycans and glycopeptides: an informatics' perspective // Current Opinions in Structural Biology. – 2016. – Т. 40. – С. 70-80.

327) Andersen M. R., Nam J. H., Sharfstein S. T. Protein glycosylation: analysis, characterization, and engineering // Upstream Industrial Biotechnology / Flickinger M. C. John Wiley & Sons, Limited, 2013. – Гл. 23, С. 489-542.

328) Kawano S. Glycobiology meets the semantic web // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. – Japan: Springer, 2017. – Гл. 17, С. 351-370.

329) Campbell M. P. A review of software applications and databases for the interpretation of glycopeptide data // Trends in Glycoscience and Glycotechnology. – 2017. – Т. 29, № 168. – С. E51-E62.

330) Ranzinger R., Maaß K., Lütteke T. Bioinformatics databases and applications available for glycobiology and glycomics // Functional and Structural Proteomics of Glycoproteins / Owens R., Nettleship J. – New York: Springer, Dordrecht, 2011, С. 59-90.

331) Ranzinger R., Herget S., Lütteke T., Frank M. Carbohydrate Structure Databases // Handbook of Glycomics / Cummings R. D., Pierce J. M. Elsevier Inc., 2010. – Гл. 10, С. 211-233.

332) Bennun S. V., Hizal D. B., Ranzinger R., Betenbaugh M. J. Towards integrative glycoinformatics for glycan based biomarker cancer research and discovery // Journal of Glycobiology. – 2013. – Т. S1, № 1.

333) Hayes C. A., Carta G., Karlsson N. G., Duffy F., Rudd P. M. Informatics and analytical tools for glycan analysis and the development of biotherapeutics // Carbohydrate Chemistry: State of the Art and Challenges for Drug Development / Cipolla L. Imperial College Press, 2015. – Гл. 7, С. 173-192.

334) Campbell M. P., Peterson R. A., Gasteiger E., Mariethoz J., Lisacek F., Packer N. H. Navigating the Glycome Space and Connecting the Glycoproteome // Protein Bioinformatics / Wu C. и др. – New York, NY: Humana Press, 2017, С. 139-158.

335) Aoki-Kinoshita K. F., Aoki N. P., Fujita A., Fujita N., Kawasaki T., Matsubara M., Okuda S., Shikanai T., Shinmachi D., Solovieva E., Suzuki Y., Tsuchiya S., Yamada I., Narimatsu H. Latest developments in semantic web technologies applied to the glycosciences // Perspectives in Science. – 2017. – Т. 11. – С. 18-23.

336) Lutteke T. Web resources for the glycoscientist // Chembiochem. – 2008. – Т. 9, № 13. – С. 2155-2160.

337) Łowicki D., Czarny A., Mlynarski J. NMR of carbohydrates // Nuclear Magnetic Resonance / Wojcik J., Kamienska-Trela K. Royal Society of Chemistry, 2013.

338) Aoki-Kinoshita K. F. Using glycome databases for drug discovery // Expert Opinions in Drug Discovery. – 2008. – Т. 3, № 8. – С. 877-890.

339) Sarkar A., Pérez S. Glycoinformatics and Glycosciences // Encyclopedia of Information Science and Technology / Khosrow-Pour D. B. A., M. – Hershey, PA: IGI Global, 2015. – Гл. 40, С. 414-425.

340) Ceroni A., Joshi H. J., Maaß K., Ranzinger R., von der Lieth C.-W. Informatics tools for glycomics: assisted interpretation and annotation of mass spectra // Glycoscience / Fraser-Reid B. O. и др. – Berlin, Heidelberg: Springer, 2008, С. 2219-2240.

341) Kailemia M. J., Xu G., Wong M., Li Q., Goonatilleke E., Leon F., Lebrilla C. B. Recent advances in the mass spectrometry methods for glycomics and cancer // Analytical Chemistry. – 2018. – Т. 90, № 1. – С. 208-224.

342) Верещагин А. Н. Классические и междисциплинарные подходы в дизайне органических и гибридных молекулярных систем // Известия Академии Наук, Серия химическая. – 2017. – Т. 10. – С. 1765-1796.

343) Herget S., Ranzinger R., Thomson R., Frank M., von der Lieth C.-W. Introduction to Carbohydrate Structure and Diversity // Bioinformatics for Glycobiology and Glycomics: An Introduction / von der Lieth C.-W. и др. John Wiley & Sons, Ltd, 2009. – Гл. 2, С. 21-47.

344) Werz D., Koester D., Holkenbrink A. Recent advances in the synthesis of carbohydrate mimetics // Synthesis. – 2010. – Т. 2010, № 19. – С. 3217-3242.

345) Benito J., Ortega-Caballero F. Recent Developments on Synthetic Tools Towards Structural and Functional Glycodiversity // Current Medicinal Chemistry. – 2013. – Т. 20, № 32. – С. 3986-4029.

346) Lütteke T. Problems and pitfalls of residue notation in glycoinformatics // Glycoscience: Biology and Medicine / Endo T. и др. – Japan: Springer, 2015. – Гл. 30, С. 251-258.

347) Campbell M. P., Ranzinger R., Lutteke T., Mariethoz J., Hayes C. A., Zhang J., Akune Y., Aoki-Kinoshita K. F., Damerell D., Carta G., York W. S., Haslam S. M., Narimatsu H., Rudd P. M., Karlsson N. G., Packer N. H., Lisacek F. Toolboxes for a standardised and systematic study of glycans // BMC Bioinformatics. – 2014. – Т. 15 Suppl 1. – С. ID S9.

348) Mariño K., Bones J., Kattla J. J., Rudd P. M. A systematic approach to protein glycosylation analysis: a path through the maze // Nature Chemical Biology. – 2010. – Т. 6, № 10. – С. 713-723.

349) Tansel B. Morphology, composition and aggregation mechanisms of soft bioflocs in marine snow and activated sludge: a comparative review // Journal of Environmental Management. – 2018. – Т. 205. – С. 231-243.

350) Harvey D. J. Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: an update for 2007-2008 // Mass Spectrometry Reviews. – 2012. – Т. 31, № 2. – С. 183-311.

351) Pereira F. 1D 13C-NMR data as molecular descriptors in spectra-structure relationship analysis of oligosaccharides // Molecules. – 2012. – Т. 17, № 4. – С. 3818-3833.

352) Horne G. Iminosugars: Therapeutic Applications and Synthetic Considerations // Carbohydrates as Drugs / Seeberger P., Rademacher C. – Berlin, Heidelberg: Springer, Cham, 2014. – Гл. 2, С. 23-51.

353) Ovchinnikova O. G., Rozalski A., Liu B., Knirel Y. A. O-antigens of bacteria of the genus providencia: structure, serology, genetics, and biosynthesis // Biochemistry (Moscow). – 2013. – Т. 78, № 7. – С. 798-817.

354) Koharudin L. M. I., Gronenborn A. M. Nuclear Magnetic Resonance Studies of Carbohydrate-Protein Interactions // Structural Glycobiology / Yuriev E., Ramsland P. A. CRC Press, 2013. – Гл. 2, С. 29-45.

355) Grachev A. A., Gerbst A. G., Ustyuzhanina N. E., Krylov V. B., Shashkov A. S., Usov A. I., Nifantiev N. E. Modeling of polysaccharides with oligosaccharides: how large should the model be? // Mendeleev Communications. – 2007. – Т. 17, № 2. – С. 57-62.

356) Vauthier C., Bertholon I., Labarre D. Integrated Development of Glycobiologics: From Discovery to Applications in the Design of Nanoparticular Drug Delivery Systems // Handbook of Pharmaceutical Biotechnology / Gad S. C. John Wiley & Sons, Inc., 2006. – Гл. 1.4, С. 125-160.

357) Ståhle J. Structure elucidations of bacterial polysaccharides using NMR spectroscopy and bioinformatics: PhD dissertation; Stockholm University. – Stockholm: Department of Organic Chemistry, Stockholm University, 2017.

358) Orlov N. V. Rational design of complex molecular structures starting from readily available precursors // Russian Chemical Bulletin. – 2017. – Т. 65, № 6. – С. 1418-1440.

359) Ranzinger R., York W. S. Glyco-bioinformatics today (August 2011) – solutions and problems // Cracking the sugar code by navigating the glycospace / Под ред. Hicks M. G., Kettner C. – Potsdam, Germany: Logos Verlag Berlin GmbH, 2011. – С. 125-154.

360) Hoehndorf R., Queralt-Rosinach N., Kuhn T. Data Science and symbolic AI: Synergies, challenges and opportunities // Data Science. – 2017.10.3233/ds-170004. – С. 1-12.

361) Kocbek S., Kim J. D. Exploring biomedical ontology mappings with graph theory methods // PeerJ. – 2017. – Т. 5. – С. e2990.

362) Aoki-Kinoshita K. F. Semantic web technologies applied to glycoscience data to integrate with life science databases // Discovering the Subtleties of Sugars / Под ред. Hicks M. G., Kettner C. – Potsdam, Germany: Logos Verlag Berlin GmbH, 2013. – С. 39-46.

363) Fukushima A., Kanaya S., Nishida K. Integrated network analysis and effective tools in plant systems biology // Frontiers in Plant Science. – 2014. – Т. 5. – С. ID 598.

364) Hofmann-Apitius M., Ball G., Gebel S., Bagewadi S., de Bono B., Schneider R., Page M., Kodamullil A. T., Younesi E., Ebeling C., Tegner J., Canard L. Bioinformatics Mining and Modeling Methods for the Identification of Disease Mechanisms in Neurodegenerative Disorders // International Journal of Molecular Sciences. – 2015. – Т. 16, № 12. – С. 29179-29206.

365) Harvey D. J. Analysis of Protein Glycosylation by Mass Spectrometry // Analysis of Protein Post-Translational Modifications by Mass Spectrometry / Griffiths J. R., Unwin R. D. John Wiley & Sons, Inc., 2016. – Гл. 3, С. 89-159.

366) Egorova K. S., Toukach P. V. Carbohydrate Structure Database (CSDB): examples of usage // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. – Japan: Springer, 2017. – Гл. 5, С. 75-113.

367) Filatov A. V., Wang M., Wang W., Perepelov A. V., Shashkov A. S., Wang L., Knirel Y. A. Structure and genetics of the O-antigen of Enterobacter cloacae C6285 containing di-N-acetyllegionaminic acid // Carbohydrate Research. – 2014. – Т. 392. – С. 21-24.

368) Perepelov A. V., Filatov A. V., Wang M., Shashkov A. S., Wang L., Knirel Y. A. Structure and gene cluster of the O-antigen of Enterobacter cloacae G3421 // Carbohydrate Research. – 2016. – Т. 427. – С. 55-59.

369) Senchenkova S. N., Guo X., Filatov A. V., Perepelov A. V., Liu B., Shashkov A. S., Knirel Y. A. Structure elucidation and gene cluster characterization of the O-antigen of Escherichia coli O80 // Carbohydrate Research. – 2016. – Т. 432. – С. 83-87.

370) Shashkov A. S., Potekhina N. V., Kachala V. V., Senchenkova S. N., Dorofeeva L. V., Evtushenko L. I. A novel galactofuranan from the cell wall of Arthrobacter sp. VKM capital A, Cyrillicsmall es, Cyrillic-2576 // Carbohydrate Research. – 2012. – Т. 352. – С. 215-218.

371) Valueva O. A., Zdorovenko E. L., Kachala V. V., Varbanets L. D., Arbatsky N. P., Shubchynskyy V. V., Shashkov A. S., Knirel Y. A. Structure of the O-polysaccharide of Pragia fontium 27480 containing 2,3-diacetamido-2,3-dideoxy-d-mannuronic acid // Carbohydrate Research. – 2011. – Т. 346, № 1. – С. 146-149.

372) Zdorovenko E. L., Valueva O. A., Kachala V. V., Shashkov A. S., Knirel Y. A., Komaniecka I., Choma A. Structure of the O-polysaccharide of Azorhizobium caulinodans HAMBI 216; identification of 3-C-methyl-D-rhamnose as a component of bacterial polysaccharides // Carbohydrate Research. – 2012. – Т. 358. – С. 106-109.

373) Liu B., Knirel Y. A., Feng L., Perepelov A. V., Senchenkova S. N., Reeves P. R., Wang L. Structural diversity in Salmonella O antigens and its genetic basis // FEMS Microbiology Reviews. – 2014. – Т. 38, № 1. – С. 56-89.

374) Senchenkova S. N., Guo X., Naumenko O. I., Shashkov A. S., Perepelov A. V., Liu B., Knirel Y. A. Structure and genetics of the O-antigens of Escherichia coli O182-O187 // Carbohydrate Research. – 2016. – Т. 435. – С. 58-67.

375) Senchenkova S. N., Shashkov A. S., Popova A. V., Shneider M. M., Arbatsky N. P., Miroshnikov K. A., Volozhantsev N. V., Knirel Y. A. Structure elucidation of the capsular polysaccharide of Acinetobacter baumannii AB5075 having the KL25 capsule biosynthesis locus // Carbohydrate Research. – 2015. – Т. 408. – С. 8-11.

376) Senchenkova S. N., Shashkov A. S., Shneider M. M., Arbatsky N. P., Popova A. V., Miroshnikov K. A., Volozhantsev N. V., Knirel Y. A. Structure of the capsular polysaccharide of Acinetobacter baumannii ACICU containing di-N-acetylpseudaminic acid // Carbohydrate Research. – 2014. – Т. 391. – С. 89-92.

377) Senchenkova S. N., Zhang Y., Perepelov A. V., Guo X., Shashkov A. S., Weintraub A., Liu B., Widmalm G., Knirel Y. A. Structure and gene cluster of the O-antigen of Escherichia coli O165 containing 5-N-acetyl-7-N-[(R)-3-hydroxybutanoyl]pseudaminic acid // Glycobiology. – 2016. – Т. 26, № 4. – С. 335-342.

378) Toukach F. V., Kocharova N. A., Maszewska A., Shashkov A. S., Knirel Y. A., Rozalski A. Structure of the O-polysaccharide of Providencia alcalifaciens O8 containing (2S,4R)-2,4-dihydroxypentanoic acid, a new non-sugar component of bacterial glycans // Carbohydrate Research. – 2008. – Т. 343, № 15. – С. 2706-2711.

379) Nihira T., Nakai H., Kitaoka M. 3-O-alpha-D-glucopyranosyl-L-rhamnose phosphorylase from Clostridium phytofermentans // Carbohydrate Research. – 2012. – Т. 350. – С. 94-97.

380) Katzenellenbogen E., Kocharova N. A., Toukach P. V., Gorska S., Bogulska M., Gamian A., Knirel Y. A. Structures of a unique O-polysaccharide of Edwardsiella tarda PCM 1153 containing an amide of galacturonic acid with 2-aminopropane-1,3-diol and an abequose-containing O-polysaccharide shared by E. tarda PCM 1145, PCM 1151 and PCM 1158 // Carbohydrate Research. – 2012. – Т. 355. – С. 56-62.

381) Knirel Y. A., Gabius H. J., Blixt O., Rapoport E. M., Khasbiullina N. R., Shilova N. V., Bovin N. V. Human tandem-repeat-type galectins bind bacterial non-betaGal polysaccharides // Glycoconj J. – 2014. – Т. 31, № 1. – С. 7-12.

382) Potekhina N. V., Shashkov A. S., Senchenkova S. N., Dorofeeva L. V., Evtushenko L. I. Structure of hexasaccharide 1-phosphate polymer from Arthrobacter uratoxydans VKM Ac-1979(T) cell wall // Biochemistry (Moscow). – 2012. – Т. 77, № 11. – С. 1294-302.

383) Vázquez-Vázquez J. L., Ortega-de la Rosa N. D., Huerta-Ocho S., Gimeno M., Gutiérrez-Rojas M. Novel exopolysaccharide produced by Acinetobacter bouvetii uam25: production, characterization and pahs bioemulsifying capability // Revista Mexicana de Ingeniería Química. – 2017. – Т. 16, № 3. – С. 721-733.

384) Malott R. J., Keller B. O., Gaudet R. G., McCaw S. E., Lai C. C., Dobson-Belaire W. N., Hobbs J. L., St Michael F., Cox A. D., Moraes T. F., Gray-Owen S. D. Neisseria gonorrhoeae-derived heptose elicits an innate immune response and drives HIV-1 expression // Proceedings of the National Academy of Sciences of the U.S.A. – 2013. – Т. 110, № 25. – С. 10234-10239.

385) Gaudet R. G., Gray-Owen S. D. Heptose sounds the alarm: innate sensing of a bacterial sugar stimulates immunity // PLoS Pathology. – 2016. – Т. 12, № 9. – С. e1005807.

386) Kersten R. D., Ziemert N., Gonzalez D. J., Duggan B. M., Nizet V., Dorrestein P. C., Moore B. S. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules // Proceedings of the National Academy of Sciences of the U.S.A. – 2013. – Т. 110, № 47. – С. E4407-E4416.

387) Bishnoi R., Khatri I., Subramanian S., Ramya T. N. Prevalence of the F-type lectin domain // Glycobiology. – 2015. – Т. 25, № 8. – С. 888-901.

388) Quijada M., Riboulleau A., Guerardel Y., Monnet C., Tribovillard N. Neutral aldoses derived from sequential acid hydrolysis of sediments as indicators of diagenesis over 120,000years // Organic Geochemistry. – 2015. – Т. 81. – С. 53-63.

389) Kosma P. Recent advances in Kdo-glycoside formation // Carbohydrate ChemistryThe Royal Society of Chemistry, 2017, С. 116-164.

390) Dugovich B. S., Peel M. J., Palmer A. L., Zielke R. A., Sikora A. E., Beechler B. R., Jolles A. E., Epps C. W., Dolan B. P. Detection of bacterial-reactive natural IgM antibodies in desert bighorn sheep populations // PLoS One. – 2017. – Т. 12, № 6. – С. e0180415.

391) Lubecka E. A., Liwo A. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. II. Backbone-local potentials of coarse-grained O1-->4-bonded polyglucose chains // Journal of Chemical Physics. – 2017. – Т. 147, № 11. – С. 115101.

392) Ovchinnikova O. G., Mallette E., Koizumi A., Lowary T. L., Kimber M. S., Whitfield C. Bacterial beta-Kdo glycosyltransferases represent a new glycosyltransferase family (GT99) // Proceedings of the National Academy of Sciences of the U.S.A. – 2016. – Т. 113, № 22. – С. E3120-E3129.

393) Krylov V. B., Argunov D. A., Solovev A. S., Petruk M. I., Gerbst A. G., Dmitrenok A. S., Shashkov A. S., Latge J. P., Nifantiev N. E. Synthesis of oligosaccharides related to galactomannans from Aspergillus fumigatus and their NMR spectral data // Organic and biomolecular chemistry. – 2018. – Т. 16, № 7. – С. 1188-1199.

394) Ustyuzhanina N. E., Bilan M. I., Dmitrenok A. S., Shashkov A. S., Nifantiev N. E., Usov A. I. Two structurally similar fucosylated chondroitin sulfates from the holothurian species Stichopus chloronotus and Stichopus horrens // Carbohydrate Polymers. – 2018. – Т. 189. – С. 10-14.

395) Completo G. C., Lowary T. L. Synthesis of galactofuranose-containing acceptor substrates for mycobacterial galactofuranosyltransferases // Journal of Organic Chemistry. – 2008. – Т. 73, № 12. – С. 4513-4525.

396) Pathak A. K., Pathak V., Seitz L., Maddry J. A., Gurcha S. S., Besra G. S., Suling W. J., Reynolds R. C. Studies on (b,1-->5) and (b,1-->6) linked octyl Gal(f) disaccharides as

substrates for mycobacterial galactosyltransferase activity // Bioorganic & medicinal chemistry. – 2001. – Т. 9, № 12. – С. 3129-3143.

397) Tilve M. J., Cori C. R., Gallo-Rodriguez C. Regioselective 5-O-opening of conformationally locked 3,5-O-Di-tert-butylsilylene-d-galactofuranosides. Synthesis of (1-->5)-b-D-galactofuranosyl derivatives // Journal of Organic Chemistry. – 2016. – Т. 81, № 20. – С. 9585-9594.

398) Chen S. G., Xue C. H., Yin L. A., Tang Q. J., Yu G. L., Chai W. G. Comparison of structures and anticoagulant activities of fucosylated chondroitin sulfates from different sea cucumbers // Carbohydrate Polymers. – 2011. – Т. 83, № 2. – С. 688-696.

399) Katzenellenbogen E., Kocharova N. A., Zatonsky G. V., Witkowska D., Bogulska M., Shashkov A. S., Gamian A., Knirel Y. A. Structural and serological studies on a new 4-deoxy-d-arabino-hexose-containing O-specific polysaccharide from the lipopolysaccharide of Citrobacter braakii PCM 1531 (serogroup O6) // European Journal of Biochemistry. – 2003. – Т. 270, № 13. – С. 2732-2738.

400) Wang M., Arbatsky N. P., Xu L., Shashkov A. S., Wang L., Knirel Y. A. O antigen of FranconibacterpulverisG3872 (O1) is a 4-deoxy-d-arabino-hexose-containing polysaccharide synthesized by the ABC-transporter-dependent pathway // Microbiology. – 2016. – Т. 162, № 7. – С. 1103-1113.

401) Dmitriev B. A., Backinowsky L. V., Knirel Y. A., Kochetkov N. K. Somatic antigens of Shigella. The strucuture of the specific polysaccharide chain of Shigella dysenteriae type 5 lipopolysaccharide // European Journal of Biochemistry. – 1977. – Т. 78, № 2. – С. 381-387.

402) Dmitriev B. A., Knirel Y. A., Kochetkov N. K., Jann B., Jann K. Cell-wall lipopolysaccharide of the 'Shigella-like' Escherichia coli 058. Structure of the polysaccharide chain // European Journal of Biochemistry. – 1977. – Т. 79, № 1. – С. 111-115.

403) Perepelov A. V., Senchenkova S. N., Shashkov A. S., Knirel' Iu A., Liu B., Feng L., Wang L. [Antigenic polysaccharides of bacteria: 41. Structures of the O-specific polysaccharides of Shigella dysenteriae types 4 and 5 revised by NMR spectroscopy] // Биоорганическая Химия. – 2008. – Т. 34, № 4. – С. 513-521.

## 9. Публикации и апробация работы

По материалам диссертации опубликовано 3 монографии и 28 статей в научных журналах, рекомендованных ВАК, из них 19 - в журналах первого квартиля (Q1). Одна статья [ссылка 4 в разд. 9.2] была включена в кандидатскую диссертацию автора, но оставлена в данной работе, так как на ней базировались дальнейшие разработки. По остальным 30 публикациям диссертации не защищались. Основное содержание работы изложено в публикациях, перечисленных в подразделах 9.1-9.3[a]:

### *9.1. Главы в книгах*

1) Toukach P. V., Egorova K. S. Bacterial, Plant, and Fungal Carbohydrate Structure Database (CSDB) // Glycoscience: Biology and Medicine / Endo T. и др. – Japan: Springer, 2014. – Гл. 29, С. 241-250.
2) Toukach P. V., Egorova K. S. Bacterial, Plant, and Fungal Carbohydrate Structure Databases: daily usage // Glycoinformatics / Lütteke T., Frank M. – New York: Springer, 2015. – Гл. 5, С. 55-85.
3) Egorova K. S., Toukach P. V. Carbohydrate Structure Database (CSDB): examples of usage // A Practical Guide to Using Glycomics Databases / Aoki-Kinoshita K. F. – Japan: Springer, 2017. – Гл. 5, С. 75-113.

### *9.2 Статьи в реферируемых рецензируемых журналах*

4) Toukach F. V., Shashkov A. S. Computer-assisted structural analysis of regular glycopolymers on the basis of 13C NMR data // Carbohydrate Research. – 2001. – Т. 335, № 2. – С. 101-114.
5) Toukach Ph. V., Joshi H. J., Ranzinger R., Knirel Y., von der Lieth C. W. Sharing of worldwide distributed carbohydrate-related digital resources: online

[a] Публикации, на которые приведены ссылки в этом разделе, имеют сквозную нумерацию, не связанную с общим списком литературы.

connection of the Bacterial Carbohydrate Structure DataBase and GLYCOSCIENCES.de // Nucleic Acids Research. – 2007. – T. 35, № Database issue. – C. D280-D286.

6) Herget S., Toukach Ph. V., Ranzinger R., Hull W. E., Knirel Y. A., von der Lieth C. W. Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans // BMC Structural Biology. – 2008. – T. 8. – C. ID 35.
7) Toukach Ph. V. Bacterial carbohydrate structure database 3: principles and realization // Journal of Chemical Information and Modeling. – 2011. – T. 51, № 1. – C. 159–170.
8) Egorova K. S., Toukach Ph. V. Critical analysis of CCSD data quality // Journal of Chemical Information and Modeling. – 2012. – T. 52, № 11. – C. 2812–2814.
9) Toukach F. V., Ananikov V. P. Recent advances in computational predictions of NMR parameters for the structure elucidation of carbohydrates: methods and limitations. // Chemical Society Reviews. – 2013. T. 42, № 21. – C. 8376–8415.
10) Aoki-Kinoshita K. F., Bolleman J., Campbell M. P., Kawano S., Kim J. D., Lutteke T., Matsubara M., Okuda S., Ranzinger R., Sawaki H., Shikanai T., Shinmachi D., Suzuki Y., Toukach Ph. V., Yamada I., Packer N. H., Narimatsu H. Introducing glycomics data into the Semantic Web // Journal of Biomedical Semantics. – 2013. – T. 4, № 1. – C. ID 39.
11) Aoki-Kinoshita K. F., Sawaki H., An H. J., Campbell M., Cao Q., Cummings R., Hsu D. K., Kato M., Kawasaki T., Khoo K. H., Kim J., Kolarich D., Li X., Liu M., Matsubara M., Okuda S., Packer N. H., Ranzinger R., Shen H., Shikanai T., Shinmachi D., Toukach Ph.V., Yamada I., Yamaguchi Y., Yang P., Ying W., Yoo J. S., Zhang Y., Zhang Y., Narimatsu H. The Fifth ACGG-DB Meeting Report: Towards an International Glycan Structure Repository // Glycobiology. – 2013. – T. 23, № 12. – C. 1422-1424.
12) Egorova K. S., Toukach Ph. V. Expansion of coverage of Carbohydrate Structure Database (CSDB) // Carbohydrate Research. – 2014. – T. 389. – C. 112-114.
13) Kapaev R. R., Egorova K. S., Toukach Ph. V. Carbohydrate structure generalization scheme for database-driven simulation of experimental

observables, such as NMR chemical shifts // Journal of Chemical Information and Modeling. – 2014. – Т. 54, № 9. – С. 2594-2611.

14) Katayama T., Wilkinson M. D., Aoki-Kinoshita K. F., Kawashima S., Yamamoto Y., Yamaguchi A., Okamoto S., Kawano S., Kim J. D., Wang Y., Wu H., Kano Y., Ono H., Bono H., Kocbek S., Aerts J., Akune Y., Antezana E., Arakawa K., Aranda B., Baran J., Bolleman J., Bonnal R. J., Buttigieg P. L., Campbell M. P., Chen Y. A., Chiba H., Cock P. J., Cohen K. B., Constantin A., Duck G., Dumontier M., Fujisawa T., Fujiwara T., Goto N., Hoehndorf R., Igarashi Y., Itaya H., Ito M., Iwasaki W., Kalas M., Katoda T., Kim T., Kokubu A., Komiyama Y., Kotera M., Laibe C., Lapp H., Lutteke T., Marshall M. S., Mori T., Mori H., Morita M., Murakami K., Nakao M., Narimatsu H., Nishide H., Nishimura Y., Nystrom-Persson J., Ogishima S., Okamura Y., Okuda S., Oshita K., Packer N. H., Prins P., Ranzinger R., Rocca-Serra P., Sansone S., Sawaki H., Shin S. H., Splendiani A., Strozzi F., Tadaka S., Toukach Ph. V., Uchiyama I., Umezaki M., Vos R., Whetzel P. L., Yamada I., Yamasaki C., Yamashita R., York W. S., Zmasek C. M., Kawamoto S., Takagi T. BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains // Journal of Biomedical Semantics. – 2014. – Т. 5, № 1. – С. ID 5.

15) Egorova K. S., Kalinchuk N. A., Knirel Y. A., Toukach Ph. V. Carbohydrate Structure Database (CSDB): new features // Russian Chemical Bulletin. – 2015. – Т. 64, № 5. – С. 1205-1210.

16) Egorova K. S., Kondakova A. N., Toukach Ph. V. Carbohydrate Structure Database: tools for statistical analysis of bacterial, plant and fungal glycomes // Database (Oxford). – 2015. – Т. 2015. – С. ID bav073.

17) Kapaev R. R., Toukach Ph. V. Improved Carbohydrate Structure Generalization Scheme for (1)H and (13)C NMR Simulations // Analytical Chemistry. – 2015. – Т. 87, № 14. – С. 7006-7010.

18) Ranzinger R., Aoki-Kinoshita K. F., Campbell M. P., Kawano S., Lutteke T., Okuda S., Shinmachi D., Shikanai T., Sawaki H., Toukach Ph. V., Matsubara M., Yamada I., Narimatsu H. GlycoRDF: an ontology to standardize glycomics data in RDF // Bioinformatics. – 2015. – Т. 31, № 6. – С. 919–925.

19) Varki A., Cummings R. D., Aebi M., Packer N. H., Seeberger P. H., Esko J. D., Stanley P., Hart G., Darvill A., Kinoshita T., Prestegard J. J., Schnaar R. L., Freeze H. H., Marth J. D., Bertozzi C. R., Etzler M. E., Frank M., Vliegenthart J. F., Lutteke T., Perez S., Bolton E., Rudd P., Paulson J., Kanehisa M., Toukach Ph. V., Aoki-Kinoshita K. F., Dell A., Narimatsu H., York W., Taniguchi N., Kornfeld S. Symbol Nomenclature for Graphical Representations of Glycans // Glycobiology. – 2015. – Т. 25, № 12. – С. 1323-1324.

20) Kapaev R. R., Toukach Ph. V. Simulation of 2D NMR Spectra of Carbohydrates Using GODESS Software // Journal of Chemical Information and Modeling. – 2016. – Т. 56, № 6. – С. 1100-1104.

21) Toukach Ph. V., Egorova K. S. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts // Nucleic Acids Research. – 2016. – Т. 44, № D1. – С. D1229–D1236.

22) Egorova K. S., Toukach Ph. V. CSDB_GT: a new curated database on glycosyltransferases // Glycobiology. – 2017. – Т. 27, № 4. – С. 285-290.

23) Kapaev R. R., Toukach Ph. V. GRASS: semi-automated NMR-based structure elucidation of saccharides // Bioinformatics. – 2018. – Т. 34, № 6. – С. 957-963.

24) Chernyshov I. Y., Toukach Ph. V. REStLESS: automated translation of glycan sequences from residue-based notation to SMILES and atomic coordinates // Bioinformatics. – 2018. – Т. 34, № 15. – С. 2679-2681.

25) Egorova K. S., Toukach Ph. V. Glycoinformatics: bridging isolated islands in the sea of data // Angewandte Chemie International Edition. – 2018. – Т. 57, № 46. – С. 14986-14990.

26) Alocci D., Suchánková P., Costa R., Hory N., Mariethoz J., Svobodová Vařeková R., Toukach P., Lisacek F. SugarSketcher: quick and intuitive online glycan drawing // MDPI Molecules. – 2018. – Т. 23, № 12. – С. ID 3206.

### *9.3 Тезисы докладов на конференциях*

Результаты представлены автором в виде приглашённых, пленарных, устных и стендовых докладов на 24 международных (33 доклада) и 13 российских (18 докладов) научных конференциях. Опубликованы тезисы 41 доклада, преимущественно на международных конференциях, из них четыре – в виде статей в реферируемых научных журналах:

27) Toukach Ph. V., Knirel Y. A. New database of bacterial carbohydrate structures // Glycoconjugate Journal. – 2005. – Т. 22. – С. 216-217.
28) Toukach Ph. V. Bacterial carbohydrate structures database version 3 // Glycoconjugate Journal. – 2009. – Т. 26. – С. 856.
29) Toukach Ph. V. CSDB and other carbohydrate databases // Glycoconjugate Journal. – 2013. – Т. 30. – С. 347-349.
30) Toukach Ph. V., Egorova K. S. Carbohydrate Structure Database merged form bacterial, plant and fungal parts // Glycoconjugate Journal. – 2015. – Т. 32. – С. 241–242.

Остальные тезисы опубликованы в сборниках материалов конференций (International Carbohydrate Symposium, European Carbohydrate Symposium, International Glycoconjugate Symposium, Baltic Meeting on Microbial Carbohydrates и других аналогичных мероприятий). Полный список докладов представлен на сайте автора[a].

[a] http://toukach.ru/rus/publist.htm#nmrsymp

### *9.4 Коллаборация*

Автор является членом двух международных углеводных консорциумов (GLIC и GlyAG) и в ходе работ по теме диссертации принимал участие в работе научных групп других организаций:

- Консорциум по Гликоинформатике (GLIC)[a];
- Консультативный совет по гликоинформатике (GlyAG) при Национальном Центре по Биотехнологической Информации (NCBI)[b];
- Отдел спектроскопии, Немецкий Центр Исследования Рака[c] (Гейдельберг, Германия);
- Портал Glycosciences.DE, Гисенский Университет Юстуса-Либиха[d] (Гисен, Германия);
- Отдел Биоинформатики, Инженерный факультет Университета Сока[e] (Токио, Япония);
- Центр исследования сложных углеводов, Университет Джорджии[f] (Атенс, США);
- Отдел Химии и Биомолекулярных Наук, Университет Маккуори[g] (Сидней, Австралия);
- Швейцарский Институт Биоинформатики[h] (Женева, Швейцария).

[a] https://glic.glycoinfo.org/

[b] https://www.ncbi.nlm.nih.gov/glycans/glyag.html

[c] https://www.dkfz.de

[d] http://www.glycosciences.de/

[e] http://bioinfo.t.soka.ac.jp/about.cgi

[f] https://www.ccrc.uga.edu/

[g] https://www.mq.edu.au/research/phd-and-research-degrees/areas-of-research/chemical-and-biomolecular-sciences

[h] https://www.sib.swiss/

Результаты работы по теме диссертации были использованы в исследованиях других коллективов (около 500 цитирований), в том числе опубликованных в соавторстве с автором диссертации в 16 статьях:

31) Kocharova N. A., Ovchinnikova O. G., Bushmarinov I. S., Toukach F. V., Torzewska A., Shashkov A. S., Knirel Y. A., Rozalski A. The structure of the O-polysaccharide from the lipopolysaccharide of *Providencia stuartii* O57 containing an amide of D-galacturonic acid with L-alanine // Carbohydrate Research. – 2005. – Т. 340, № 4. – С. 775-780.
32) Kocharova N. A., Ovchinnikova O. G., Toukach F. V., Torzewska A., Shashkov A. S., Knirel Y. A., Rozalski A. The O-polysaccharide from the lipopolysaccharide of *Providencia stuartii O44* contains L-quinovose, a 6-deoxy sugar rarely occurring in bacterial polysaccharides // Carbohydrate Research. – 2005. – Т. 340, № 7. – С. 1419-1423.
33) Bushmarinov I. S., Ovchinnikova O. G., Kocharova N. A., Toukach F. V., Torzewska A., Shashkov A. S., Knirel Y. A., Rozalski A. Structure of the O-polysaccharide from the lipopolysaccharide of *Providencia alcalifaciens O29* // Carbohydrate Research. – 2006. – Т. 341, № 9. – С. 1181-1185.
34) Torzewska A., Grabowski S., Kondakova A. N., Toukach F. V., Senchenkova S. N., Shashkov A. S., Arbatsky N. P., Knirel Y. A., Rozalski A., Kaca W. Structures and serology of the O-antigens of *Proteus* strains classified into serogroup O17 and former serogroup O35 // Archivum Immunologiae et Therapia Experimentalis. – 2006. – Т. 54, № 4. – С. 277-282.
35) Bushmarinov I. S., Ovchinnikova O. G., Kocharova N. A., Toukach F. V., Torzewska A., Shashkov A. S., Knirel Y. A., Rozalski A. Structure of the O-polysaccharide and serological cross-reactivity of the lipopolysaccharide of *Providencia alcalifaciens* O32 containing N-acetylisomuramic acid // Carbohydrate Research. – 2007. – Т. 342, № 2. – С. 268-273.
36) Ovchinnikova O. G., Bushmarinov I. S., Kocharova N. A., Toukach F. V., Wykrota M., Shashkov A. S., Knirel Y. A., Rozalski A. New structure for the O-polysaccharide of *Providencia alcalifaciens* O27 and revised structure for the O-polysaccharide of *Providencia stuartii* O43 // Carbohydrate Research. – 2007. – Т. 342, № 8. – С. 1116-1121.

37) Katzenellenbogen E., Toukach P. V., Kocharova N. A., Korzeniowska-Kowal A., Gamian A., Shashkov A. S., Knirel Y. A. Structure of a phosphoethanolamine-containing O-polysaccharide of *Citrobacter freundii* strain PCM 1443 from serogroup O39 and its relatedness to the *Klebsiella* pneumoniae O1 polysaccharide // FEMS Immunology and Medical Microbiology. – 2008. – Т. 53, № 1. – С. 60-64.

38) Shashkov A. S., Kocharova N. A., Toukach F. V., Kachala V. V., Knirel Y. A. 2,4-dihydroxypentanoic acids: new non-sugar components of bacterial polysaccharides // Natural Product Communications. – 2008. – Т. 3, № 10. – С. 1625-1630.

39) Toukach F. V., Kocharova N. A., Maszewska A., Shashkov A. S., Knirel Y. A., Rozalski A. Structure of the O-polysaccharide of *Providencia alcalifaciens* O8 containing (2S,4R)-2,4-dihydroxypentanoic acid, a new non-sugar component of bacterial glycans // Carbohydrate Research. – 2008. – Т. 343, № 15. – С. 2706-2711.

40) Katzenellenbogen E., Kocharova N. A., Toukach P. V., Gorska S., Korzeniowska-Kowal A., Bogulska M., Gamian A., Knirel Y. A. Structure of an abequose-containing O-polysaccharide from *Citrobacter freundii* O22 strain PCM 1555 // Carbohydrate Research. – 2009. – Т. 344, № 13. – С. 1724-1728.

41) Katzenellenbogen E., Kocharova N. A., Toukach P. V., Gorska S., Bogulska M., Gamian A., Knirel Y. A. Structures of a unique O-polysaccharide of *Edwardsiella tarda* PCM 1153 containing an amide of galacturonic acid with 2-aminopropane-1,3-diol and an abequose-containing O-polysaccharide shared by *E. tarda* PCM 1145, PCM 1151 and PCM 1158 // Carbohydrate Research. – 2012. – Т. 355. – С. 56-62.

42) Shakhmatov E. G., Toukach P. V., Kuznetsov S. P., Makarova E. N. Structural characteristics of water-soluble polysaccharides from *Heracleum sosnowskyi* Manden // Carbohydrate Polymers. – 2014. – Т. 102. – С. 521-528.

43) Shakhmatov E. G., Toukach P. V., Michailowa C., Makarova E. N. Structural studies of arabinan-rich pectic polysaccharides from *Abies sibirica* L. Biological activity of pectins of *A. sibirica* // Carbohydrate Polymers. – 2014. – Т. 113. – С. 515-524.

44) Sizova O. V., Shashkov A. S., Dmitrenok A. S., Toukach P. V., Knirel Y. A., Shaikhutdinova R. Z., Ivanov S. A., Dentovskaya S. V. Structure and gene cluster of the O-polysaccharide of *Yersinia rohdei* H274-36/78 // International Journal of Biological Macromolecules. – 2019. – Т. 122. – С. 555-561.

45) Sigida E. N., Fedonenko Yu. P., Shashkov A. S., Toukach Ph. V., Shelud'ko A. V., Zdorovenko E. L., Knirel Yu. A., Konnova S. A. Structural studies of the two O-specific polysaccharide(s) and biological activity of the lipopolysaccharide from *Azospirillum brasilense* SR8 toward plants // International Journal of Biological Macromolecules. – 2019. – Т. 126. – С. 246-253.

46) Sizova O. V., Shashkov A. S., Toukach P. V., Knirel Y. A., Shaikhutdinova R. Z., Ivanov S. A., Dentovskaya S. V. Structure the O-polysaccharide of *Yersinia pekkanenii* C-134 // Carbohydrate Research. – 2019. – в печати

Хотя публикации в соавторстве с коллаборатораторами демонстрируют востребованность созданной платформы, результаты по теме диссертации не являются в них основными, поэтому они не представлены к защите.

### *9.5 Результаты в сети Интернет*

По результатам работы создан и поддерживается бесплатный для пользователей веб-портал Carbohydrate Structure Database[a], объединяющий разработки за период 2005-2018. Он включён во множество каталогов информационных ресурсов для учёных и в обзоры по методологии химии и биологии углеводов, опубликованные в научной литературе другими авторами. Ежемесячно фиксируется около 300 уникальных запросов пользователей.

[a] http://csdb.glycoscience.ru

## 10.Финансирование и благодарности




В период с 2004 по 2018 гг. работа имела целевое финансирование в виде 12 грантов российских и международных научных фондов (в хронологическом порядке):

1. «Банк данных по углеводным структурам II» - дополнение к гранту Международного Научно-Технического Центра № 1197, 2004, 1 год, рук. Ю.А. Книрель.
2. «Разработка, создание и поддержка Интернет-доступной базы данных 'Природные углеводы'» - Российский Фонд Фундаментальных Исследований, грант № 05-07-90099, 2005, 3 года, рук. Ю.А. Книрель.
3. «Разработка базы данных по бактериальным углеводам» - Комиссия по грантам при президенте РФ, грант МК-1700.2005.4, 2005, 1 год, рук. Ф.В. Тоукач.
4. «Объединение цифровых ресурсов в химии углеводов» - Немецкий Центр Исследования Рака, 2006, 3 месяца, рук. с российской стороны Ф.В. Тоукач, рук. с немецкой стороны К-В. фон дер Лиет.
5. «Сравнительный анализ гликомов бактерий и млекопитающих» - Немецкий Центр Исследования Рака, 2007, 3 месяца, рук. с российской стороны Ф.В. Тоукач, рук. с немецкой стороны С. Хергет.
6. «Информатизация гликобиологии» - Фонд Содействия Отечественной Науке, 2008, 2 года, рук. Ф.В. Тоукач.
7. «Эмпирическое моделирование спектров ЯМР углеводов и база данных экспериментальных данных ЯМР» - Немецкий Центр Исследования Рака, 2008, 3 месяца, рук. с российской стороны Ф.В. Тоукач, рук. с немецкой стороны М. Франк.
8. «Интеграция с CSDB базой GlycomeDB на основе новой углеводной нотации GlycoCT» - Немецкий Центр Исследования Рака, 2009, 3 месяца, рук. с российской стороны Ф.В. Тоукач, рук. с немецкой стороны С. Хергет.

9. «Разработка базы данных по грибным и растительным углеводам выявление с её помощью взаимосвязей 'структура–таксономия'» - Российский Фонд Фундаментальных Исследований, грант № 12-04-00324, 2012, 3 года, рук. Ф.В. Тоукач.
10. «Поиск корреляций 'структура – спектр ЯМР' природных углеводов. Автоматическое предсказание спектров по структуре и структуры по спектру» - Российский Фонд Фундаментальных Исследований, грант № 15-04-01065, 2015, 3 года, рук. Ф.В. Тоукач.
11. «Моделирование структурных характеристик биогликанов» - Российский Фонд Фундаментальных Исследований, грант № 18-04-00094, 2018, 3 года, рук. Ф.В. Тоукач.
12. «Привнесение информационных технологий в гликобиологию» - Российский Научный Фонд, грант № 18-14-00098, 2018, 3 года, рук. Ф.В. Тоукач.

Автор выражает благодарность проф. Ю.А. Книрелю, проф. А.С. Шашкову, проф. А.И. Усову, проф. Т. Люттеке, проф. К. Аоки-Киношита, проф. Ф. Лизачек, проф. К.-В. фон дер Лиету (†), к.б.н. К.С. Егоровой, д-ру Р. Ранцингеру, д-ру С. Хергету, д-ру М. Кэмпбеллу, к.х.н. Н.А. Калинчук, Р.Р. Капаеву, И.Ю. Чернышову за плодотворное сотрудничество.

*версия документа: 1.3*