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

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
доктора химических наук

Москва — 2019

Работа выполнена в Лаборатории металлокомплексных и наноразмерных катализаторов (№30) и Лаборатории химии углеводов (№21) Федерального государственного бюджетного учреждения науки Института органической химии им. Н.Д. Зелинского РАН (ИОХ РАН).

НАУЧНЫЙ КОНСУЛЬТАНТ:	д.х.н. проф. Книрель Юрий Александрович, ИОХ РАН
ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:	д.х.н. академик **Стоник Валентин Аронович**, научный руководитель Института, ФГБУН Тихоокеанский институт биоорганической химии им. Г.Б. Елякова Дальневосточного отделения Российской академии наук.
	д.б.н. проф. **Гельфанд Михаил Сергеевич**, зам. директора, ФГБУН Институт проблем передачи информации им. А.А. Харкевича Российской академии наук.
	д.х.н. **Мирошников Константин Анатольевич**, заведующий Лаборатории молекулярной биоинженерии, ФГБУН Институт биоорганической химии им. М.М. Шемякина и Ю.А. Овчинникова Российской академии наук.
ВЕДУЩАЯ ОРГАНИЗАЦИЯ:	Казанский институт биофизики и биохимии, Казанский научный центр Российской академии наук

Защита состоится ”22” мая 2019 г. в 11:00 на заседании диссертационного совета Д 002.222.01 при Федеральном государственном бюджетном учреждении науки Институте органической химии им. Н.Д. Зелинского Российской академии наук по адресу: 119991, Москва, Ленинский проспект, д. 47.

С диссертацией можно ознакомиться в библиотеке ИОХ РАН и на официальном сайте ИОХ РАН (http://zioc.ru). Автореферат размещён на официальном сайте Высшей аттестационной комиссии при Министерстве образования и науки Российской Федерации по адресу http://vak.ed.gov.ru

Автореферат разослан ”28” марта 2019 года

Учёный секретарь
диссертационного совета Д 002.222.01

доктор химических наук А.Д. Дильман

Общая характеристика работы

Актуальность темы. Углеводы – важные носители биологической информации, наряду с белками и нуклеиновыми кислотами. Углеводы выполняют структурные, энергетические, регуляторные и защитные функции в клетках, 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 отставания информационного обеспечения гликомики от геномики и протеомики, которое 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 информации об углеводах, отсутствием полных хранилищ данных и информационной изолированностью существующих проектов. Также отсутствуют полные репозитории экспериментальной информации о ферментах, вовлечённых в биосинтез углеводов, которая востребована при разработке ферментативных синтетических протоколов получения 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, и значительно облегчает исследования в гликохимии и гликобиологии. 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 информации, привязанных к платформе базы данных, открывает доступ к неявно присутствующим в базе знаниям. Эти инструменты позволяют неподготовленным в плане информатики учёным получать информацию, доступ к которой ранее требовал направленных компьютерных изысканий. В химии углеводов наблюдается несоответствие огромного объёма накопленных структурных данных ограниченным возможностям их обобщения и прогнозирования свойств и структур. В частности, основной аналитический метод гликохимии (спектроскопия ЯМР) плохо обеспечен средствами интерпретации экспериментальных данных, что делает структурные исследования весьма трудоёмкими. 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 полисахаридам, гликозидам и гликоконъюгатам с уже описанной первичной структурой. 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 компьютерной 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 является одним из наиболее перспективных с точки зрения интеграции с существующими гликоресурсами.

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, углевод-активных ферментов, участвующих в их биосинтезе, углеводную онтологию, форматы данных, инструменты предсказания свойств биогликанов (спектры, молекулярная геометрия и т.д.), инструменты ввода, визуализации и статистической обработки данных об углеводах и связи с другими значимыми углеводными проектами. Вышеизложенные соображения делают этот междисциплинарный проект актуальным для современной науки об углеводах.

Цели работы. Целью работы являлась оптимизация и автоматизация структурно-функциональных исследований углеводов, привнесение в гликохимиию и гликобиологию 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 биогликанов. 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 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 использовать модели в ис-

следованиях структуры природных соединений. Создание методологии оценки достоверности моделей и их валидация на большой выборке природных структур.

4. Разработка алгоритма сравнения спектров и инструмента предсказания первичной структуры сахаридов по легко получаемым экспериментальным данным, таким как одномерные спектры ЯМР, данные ГЖХ, данные эксперимента по метилированию.
5. Разработка языка описания структуры биогликанов, пригодного 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. Создание программ-переводчиков на существующие углеводные (GlycoCT, WURCS и др.) и общехимические (IUPAC, SMILES и др.) языки и с них.
6. Разработка интуитивно понятного способа визуализации углеводных структур в программах и в статьях, учитывающего все структурные особенности, характерные для биогликанов, и обратно-совместимого с существующими публикациями.
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, и сравнительный анализ гликоразнообразия в различных таксонах. Построение альтернативных «деревьев жизни», основанных на сходствах и различиях гликомов, их сравнение с классическими филогенетическими деревьями.
9. Разработка идеологии и правил обработки информации об углеводах, которые позволят сократить отставание гликомики от других наук о жизни. Эта теоретическая основа гликоинформатики должна учитывать структурную и биологическую специфику углеводов, а также исторически сложившиеся стандарты и ошибки существующих проектов. Объединение мировых проектов гликоинформатики в единую информационную среду (в сотрудничестве с другими группами), включая прозрачную для пользователей интеграцию баз данных, создание углеводной онтологии, стандартизацию используемых в гликоинформатике моделей данных, индексов и идентификаторов.

Научная новизна и практическая значимость. В работе решена 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 данными между этими базами был значительно затруднён, что ограничивало эффективность работы учёных. В результате представленной работы массив накопленных данных об углеводах получил 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, библиогра-

фической и другой информацией, с данными по ферментативному аппарату, участвующему в их биосинтезе. 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.

Впервые представлены инструменты точного предсказания характеристик биогликанов, пригодные для использования в потоковом режиме на больших множествах структур. Они включают алгоритмы моделирования вторичной структуры, алгоритмы моделирования спектров ЯМР, генерирования и оценки достоверности структурных гипотез. Эти инструменты активно используются в исследованиях гликополимеров микроорганизмов и механизмов их взаимодействия с другими клеточными структурами.

В рамках проекта проведена стандартизация представления информации об углевод-содержащих молекулах и создана специальная углеводная онтология, что 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 повысить эффективность как фундаментальной, так и прикладной работы учёных в широкой области знания – науке об углеводах. Разработки использованы в исследованиях структуры и функций биогликанов, проводимых другими группами биохимиков, молекулярных биологов, иммунологов и фармацевтов. Эти исследования включают установление строения новых бактериальных антигенов, 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, выяснение активности углевод-активных ферментов, хемотаксономическую классификацию патогенных микроорганизмов.

За 13 лет своего развития платформа CSDB заняла ведущие позиции в мировой науке об углеводах и имеет перспективы стать единственной всеобъемлющей базой по природным углеводам (таких баз не существует с 1996 года, когда из-за удорожания обслуживания была прекращена поддержка CarbBank).

Личный вклад автора. Автор непосредственно участвовал в определении направления исследований, разработке компьютерных алгоритмов, сервисов, онтологии и форматов данных, программировании, статистической обработке, верификации, интерпретации и обобщении результатов, сборе данных из компьютерных источников и литературы, написании статей. Бóльшая часть работы сделана автором единолично. Все выводы основаны на данных, полученных автором лично или при его ключевом участии. Под руководством автора по теме корреляции «структура-спектр» углеводов защищено три дипломные работы.

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) в химии и биологии. Из них одна статья была включена в кандидатскую диссертацию автора, но оставлена в данной работе в качестве базы для дальнейших разработок. По остальным 30 публикациям диссертации не защищались. Результаты работы были использова-

ны в исследованиях других коллективов (около 500 цитирований), в том числе опубликованных в соавторстве с автором диссертации ещё в 16 статьях. Наработки были представлены в виде приглашённых, устных и стендовых докладов на 24 международных (33 доклада) и 13 российских (18 докладов) научных конференциях. В процессе работы создан и поддерживается веб-портал Carbohydrate Structure Database, ежегодно фиксирующий более 3000 уникальных запросов. Финансирование работы включало 12 целевых грантов российских и международных научных фондов.

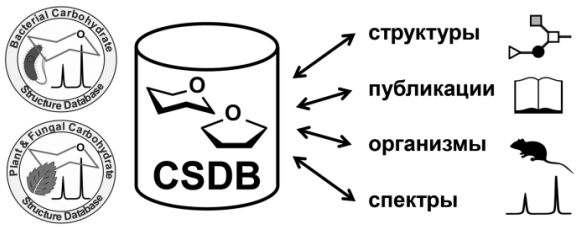
Структура и объем работы. Диссертация изложена на 302 страницах, включает 19 таблиц и 85 иллюстраций. Она состоит из введения, литературного обзора, обсуждения результатов, описания технической реализации проекта, анализа использования результатов в рамках научной области, выводов, списка сокращений, списка литературы (404 источника), информации об апробации и финансировании. Литературный обзор посвящён применению существующих методов информатики в исследованиях углеводов и содержит анализ исторически значимых и конкурирующих проектов.

Основные положения, выносимые на защиту:

- создание универсальной долговременной базы данных природных углеводов,
- высокоточное моделирование спектров ЯМР углеводов,
- полуавтоматическое установление углеводных структур по экспериментальным данным,
- моделирование молекулярной геометрии углеводов,
- кластеризация живых организмов на основании их гликанов,
- стандартизация и формализация знаний об углеводах, включая углеводную онтологию.

Основное содержание работы

1. База данных природных углеводов как платформа гликоинформатики



Новая курируемая база данных Carbohydrate Structure Database (CSBD) была спроектирована, разработана, заполнена данными, снабжена интерфейсом и представлена в сети Интернет (http://csdb.glycoscience.ru). Предназначением базы является предоставление опубликованных данных по природным гликанам, гликополимерам, гликоконъюгатам и другим сахаридам. В ходе своего развития CSBD превратилась в платформу гликоинформатики, предоставляющую не только данные, но и инструменты их выборки, верификации, визуализации, анализа и предсказания.

1.1 Данные CSDB

1.1.1 Типы данных CSDB. Информация, содержащаяся в CSDB, суммирована в Табл. 1. **Жирным** шрифтом показаны обязательные данные, присутствующие во всех записях. *Курсивом* показаны редкие данные, присутствующие менее чем в четверти записей.

Табл. 1. Данные CSDB

Группа	*Данные*	*Индексация*	*Примечания*
Структура	**Первичная структура**	транслируется в SweetDB (2D IUPAC), SNFG, GlycoCT и др.	хранится в виде таблицы связности остатков, для ввода-вывода кодируется на языке CSDB Linear; может содержать неоднозначности; выводится в виде семантических или структурных формул и 3D-моделей
	Ошибки в структуре		при наличии ошибок также приводится неправильная опубликованная структура
	Тип молекулы	контролируемый словарь	моно- / олигомер, химическое / биологическое / циклическое повторяющееся звено, фрагмент, мотив и др.
	Число повторов		для полимеров
	Молекулярный вес		
	Мономерный состав и *брутто-формула*	неявно задан в структуре	состав рассчитывается из структуры; брутто-формула – для олигомеров
	Агликон и положение его присоединения		при возможности включён в первичную структуру, если это невозможно - закодирован в виде SMILES, IUPAC или описания
	Класс, клеточная роль	индексированы	«гликосфинголипид», «О-антиген» и т.п.
	Тривиальное название	индексированы	
	Геометрия молекулы		теоретическая модель
	Биосинтетические и генетические данные, названия ферментов		для *E. coli* и *A. thaliana* – полный набор данных группы «гликозилтрансферазы»; для остальных - только наличие данных
	Тип лаб. 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, семейство, ссылка на GenBank / Uniprot
	Синтезируемая связь	индексированы	в контексте полной структуры
	Донор и акцептор	индексированы	представлены как виртуальные структуры в CSDB
	Степень достоверности	индексированы	*in silico*, *in vitro*, три градации *in vivo*
	Методы подтверждения активности		
Библиография	**Авторы**	индексированы	
	Название работы		
	Название журнала, сборника или книги	индексированы	для журналов – дополнительный внешний индекс NLM ID
	Издательство, редакторы	редакторы индексированы	для книг
	Выходные данные		год, том, номер, страницы
	Ссылки на библиографические базы	внешние индексы	DOI, NCBI PubMed ID, веб-адрес
	e-mail авторов		
	аффилиации авторов	индексированы	
	ключевые слова	индексированы	
	реферат (abstract)		
Запись (уникальная комбинация структуры и статьи, в которой она описана)	**Оригинальность**		установлена ли структура в этой статье
	Применённые методы	индексированы	
	Локализация структуры в статье		номер рисунка, схемы и т.д.
	Комментарии		все, что не закодировано в типизированных полях, включая информацию об ошибках в публикациях

Таксономия и природный контекст	**Царство, тип**	индексированы	
	Род, вид, штамм / серогруппа	индексированы (три индекса)	включая неполные сочетания, гибриды и мутанты
	Переименования и переклассификации таксонов		по отношению к опубликованному названию
	Орган, ткань, стадия развития	индексированы	
	Болезнь организма-хозяина	индексированы	ассоциированная с таксоном микроорганизма или со структурой
	Организм-хозяин	индексированы	для микроорганизмов
	Ссылка на NCBI Taxonomy	внешний индекс NCBI TaxID	
ЯМР	Спектр ^{1}H		с отнесением сигналов
	Спектр ^{13}C		с отнесением сигналов
	Температура и *pH*		
	Растворитель	индексированы	включая стандарт калибровки шкалы
Универсальные данные	**Взаимосвязи между остальными данными и идентификаторами**		
	Свойства мономеров	индексированы; контролируемый словарь; MSDB	название, возможные конфигурации остатка, стереоконфигурации атомов, число протонов и тип заместителя в каждой позиции, записи WURCS и SMILES
	Суперклассы мономеров	индексированы	разновидность неопределённости структуры на уровне остатков, напр. HEX (гексоза)
	Топологии соединения остатков	индексированы	до 12 остатков
	Эмпирические эффекты гликозилирования в спектрах ЯМР ^{13}C	индексированы	215 эффектов
	Спектры ЯМР ^{13}C модельных структур	индексированы	2679 спектров (313 остатков в моно- и олигомерных фрагментах)
	Разрешённые связи	индексированы	для всех комбинаций типов атомов
Вспомогательные данные	**Идентификаторы записей**		запись = уникальная комбинация структуры и статьи, в которой она описана
	Идентификаторы соединений		соединение = уникальная комбинация первичной структуры, её типа, агликона, полимеризации и т.д.
	Идентификаторы публикаций		
	Идентификаторы спектров		
	Идентификаторы организмов		организм = уникальная комбинация таксона (царство, тип, род, вид), штамма и/или серогруппы
	Идентификаторы гликозилтрансфераз		
	Идентификаторы генов и ферментов		
	Ссылки на родственные записи в CSDB		например, структуры, отличающиеся только ацетилированием или рамкой полимеризации, или идентичные структуры, опубликованные в другом контексте.
	Ссылки на записи в других базах		GlyTouCan, GlycomeDB, CCSD (CarbBank), Chemical abstracts, ProtDB, Genbank и др.
	Данные для отслеживания процесса аннотирования		аннотатор, контролёр, дата, ссылка на запись в лабораторной базе, ссылка на файл со статьёй, ошибки в других базах и статьях, исправленные при импорте данных.

1.1.2 Покрытие CSDB и источники данных. Таксономический охват CSDB включает микроорганизмы, растения и грибы. Покрытие по прокариотам близко к полному, что обеспечивает научную ценность даже отрицательных ответов на поисковые запросы. Данные по прокариотам попадают в базу в среднем через год после их опубликования. Покрытие по грибам реализовано до 2010 года 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, они представлены в других углеводных базах. CSDB систематически пополняется по результатам аннотирования 500-1000 публикаций в год.

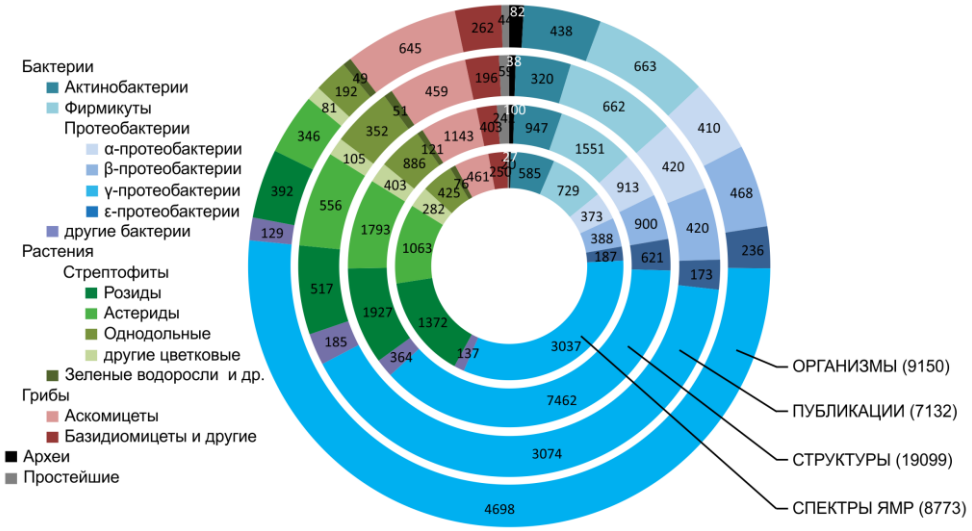


Рис. 1. Количество записей CSDB в основных таксономических группах на 2017-й год.

Объем аннотированных данных CSDB составляет 19483 структуры, 9150 таксонов, 7285 публикаций, 9427 отнесённых спектров ЯМР, 1755 активностей гликозилтрансфераз. Покрытие в пределах таксономических групп, меньших, чем царства, отражает изученность таксонов в публикациях (Рис. 1).

CSDB является первичной базой, т.е. содержит данные, полученные 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 структуры, которая: содержит хотя бы один углеводный остаток (кроме нуклеиновых кислот); определена достаточно полно, чтобы иметь возможность судить о мономерном составе и не менее чем о 50% связей и конфигураций остатков; соотносится со структурой из микроорганизма, растения или гриба. 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 модификация природной структуры. Ежегодный импорт данных включает поиск в глобальных библиографических системах по ключевым словам, первичный отбор на основании рефератов, анализ текстов и вторичный отбор, аннотирование статей в виде текстового дампа, автоматическое выявление ошибок и несоответствий, проверку и исправление аннотаций другим аннотатором, загрузку дампа в базу.
2. Другие базы данных. Структуры, опубликованные до 1996 года (~40% всех структур в CSDB), были отобраны по критериям таксономической принадлежности из базы CarbBank с полным покрытием до 1996 года. 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 и NCBI Taxonomy. Данные по свойствам моносахаридов сопоставляются с немецкой базой MonosaccharideDB и обновляются вместе с ней.

3. Загрузка пользователями собственных опубликованных данных. При этом проводится автоматическая и ручная проверка качества данных.
4. Обобщённые и опосредованные данные (напр., молекулярная геометрия, характеристичные химические сдвиги и эффекты гликозилирования в спектрах), полученные анализом и прогнозированием свойств по собственным моделям. Эти данные попадают в часть базы, изолированную от опубликованных данных, недоступны для поиска и используются для моделирования и предсказания свойств.

1.1.3. Контроль ошибок. CSDB - курируемая база данных, т.е. данные проверяются экспертами для повышения надёжности базы как источника данных. В рамках доменов прокариот, растений и грибов CSDB является единственной первичной курируемой базой. Качество данных традиционно было краеугольным камнем для биохимических баз. В CSDB оно обеспечивается автоматическим выявлением около ста видов ошибок и «подозрительных» сочетаний данных. Проверка и верификация аннотаций также включает ручное выявление ошибок, не обнаруживаемых автоматически. В порядке распространённости к ошибкам относятся: ошибки, пришедшие из других баз; некорректные аннотации; ошибки в публикациях; ошибки в программах обработки данных. Ошибки каталогизированы в трёх группах:

1. «Исправляемые» - автоматически выявляются и исправляются с выдачей предупреждения. Всего 10 типов ошибок (4 структурных, 3 таксономических, 1 спектральный, 2 общих), например, «избыточное указание стереохимии атомов» или «то же химическое повторяющееся звено полимера присутствует в других записях с иным положением рамки полимеризации».
2. «Выявляемые» - выявляются автоматически, но для исправления нужно повторное аннотирование экспертом. Всего 65 типов ошибок (15 структурных, 9 библиографических, 14 таксономических, 10 спектральных, 17 общих), например, «замещённая позиция приходится на точку замыкания цикла» или «химический сдвиг не соответствует аномерной конфигурации». Каждое действие, затрагивающее данные, отслеживается на предмет корректности запроса и соответствия результатов его смыслу. Если запрос или результат некорректен, это свидетельствует либо о наличии новых типов ошибок в данных, проверка которых ещё не реализована, либо об ошибках в программах управления CSDB или нехватке аппаратных ресурсов. В всех подобных случаях проводится ручная проверка данных и кода и выяснение причин.
3. «Экспертные» - можно выявить только путём сопоставления записи с оригинальной публикацией. Всего 36 типов ошибок (6 структурных, 8 библиографических, 5 таксономических, 3 спектральных, 14 общих), например, «химически возможная, но неправильная последовательность остатков» или «не все приведённые в статье методы исследования структуры присутствуют в записи».

Практически во всех современных углеводных базах, включая CSDB, информация, опубликованная до 1996 года, импортирована из CarbBank. В этом контексте выявление и исправление ошибок CarbBank является чрезвычайно важной задачей. В процессе импорта Car-

bBank в CSDB были автоматически проверены все записи и вручную – около 50% записей. Выяснилось, что ошибки содержатся в 40% записей CarbBank, в том числе в 10% структур бактериальных углеводов. Был проведён подробный анализ качества данных и методов обнаружения и исправления ошибок в углеводных базах.

1.2 Поиск данных

Поиск данных позволяет перейти от известных данных к связанным с ними неизвестным. Общая схема таких переходов в CSDB показана на Рис. 2. Поисковые запросы различных типов можно проводить по всему покрытию базы либо комбинировать при помощи логических операторов И (искать в результатах предыдущего запроса), ИЛИ (объединить с результатами предыдущего запроса), НЕ (получить все результаты, кроме удовлетворяющих запросу) и И НЕ (вычесть результаты, удовлетворяющие запросу, из результатов предыдущего запроса). Пример комбинирования разнородных критериев в показан на Рис. 3.

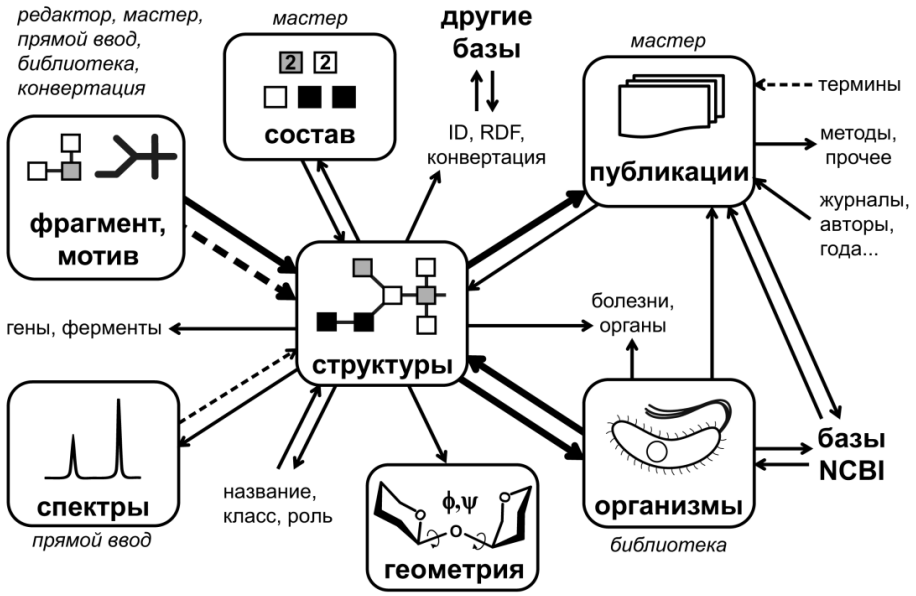


Рис. 2. Типы данных CSDB и возможные переходы между ними (толстые стрелки – наиболее распространённые переходы, тонкие – прочие переходы, пунктир – переходы с нечёткой логикой). Способы ввода приведены курсивом.

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 для идентификации объектов в других базах данных. Эти индексы включают ссылки на другие структурные (GlyTouCan, GlycomeDB, CarbBank, MonosaccharideDB), таксономические (NCBI Taxonomy), объектные (MeSH), библиографические (система DOI, NCBI Pubmed, каталог NLM) и протеомно-генетические (NCBI Genbank, NCBI Enzyme, Uniprot) базы.

Веб-интерфейс предоставляет пользователям следующие возможности поиска:

1. *Поиск структур и их фрагментов (Substructure).* Запрос включает описание структурного фрагмента на языке CSDB Linear (см. ниже). Его можно ввести: в графическом формате SNFG с помощью визуального редактора; с помощью мастера, позволяющего «собрать» структуру посредством операций в браузере; путём прямого ввода на языке CSDB Linear

либо редактирования структуры, сгенерированной мастером; путём редактирования предыдущего запроса; путём выбора из библиотеки 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 фильтрация результатов по типу молекулы, по классам (функциям) соединений, по наличию данных ЯМР, по таксономии на уровне царств. Также поддерживается текстовый поиск в тривиальных названиях соединений и агликонов. Результатом поиска являются соответствующие запросу структуры, а также сопутствующая структурная информация и список публикаций, в которых они описаны, с привязкой к таксономии.

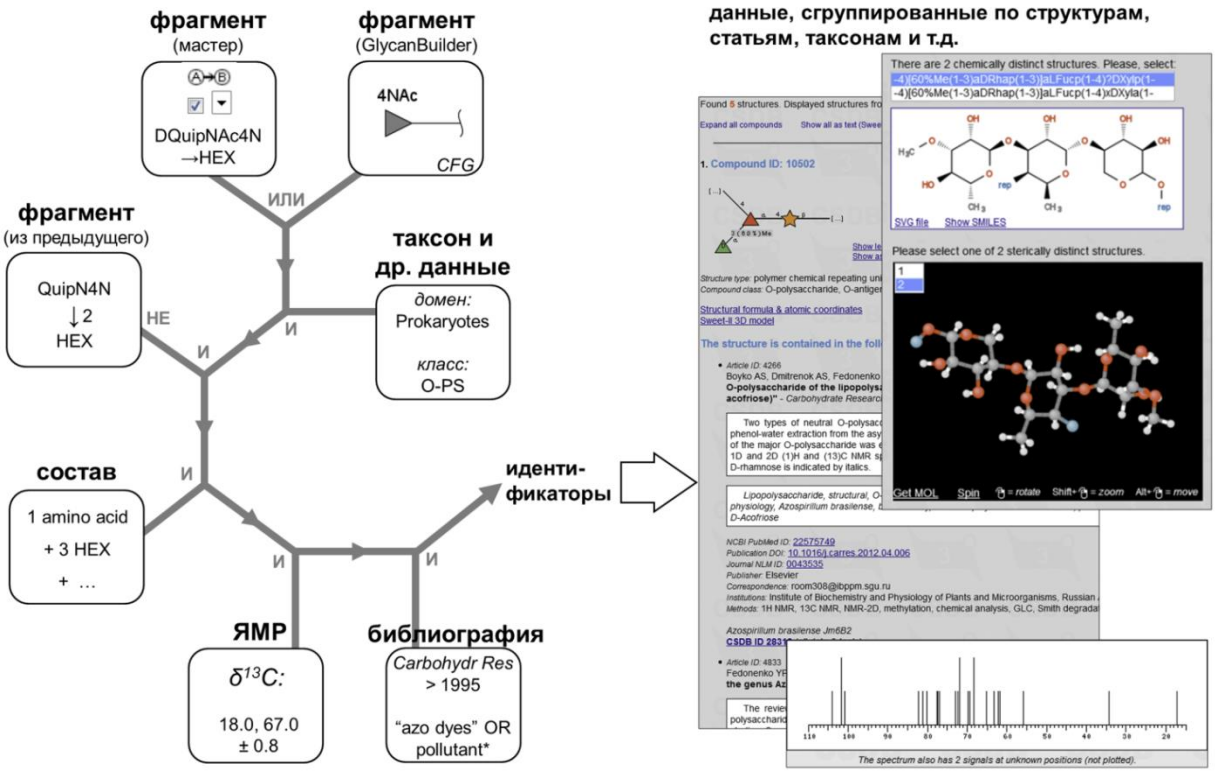


Рис. 3. Пример комбинирования разнородных критериев поиска.

2. *Поиск по мономерному составу* (*Composition*). Этот критерий позволяет найти структуры с указанным полным или частичным мономерным составом. В качестве единиц могут использоваться имена остатков и их суперклассы (напр., «гексоза» или «аминокислота»). Результат и способы его уточнения аналогичны характеристикам структурного поиска.
3. *Поиск по биологической привязке* (*Taxonomy*). Этот критерий позволяет перейти от организмов, из которых выделены структуры, к самим структурам, а также получить записи, соответствующие микроорганизмам, инфицирующим указанный организм-хозяин. Полная или частичная таксономия вводится путём последовательного сужения области поиска с помощью четырёх меню: домен, род, вид и штамм. Подвиды и серогруппы могут быть введены в свободнотекстовом виде с поддержкой символов-заменителей. Альтернативным способом идентификации является прямой ввод NCBI Tax ID. Результатом поиска являются соответствующие запросу организмы вместе с сопутствующей биологической информацией и списком выделенных из них структур, с привязкой к библиографии.
4. *Поиск по библиографии* (*Bibliography*) позволяет найти статьи, главы и книги, удовлетворяющие заданным условиям, чтобы затем перейти к опубликованным структурам и орга-

низмам. Поисковый запрос учитывает любую комбинацию следующих критериев: фамилии и инициалы авторов, наличие заданного термина в названии, реферате или ключевых словах, название журнала, ограничения на год публикации, номер тома и номер страницы. Для ввода авторов используется авторский указатель. Текст обрабатывается с поддержкой логических операций И, ИЛИ и НЕ, неразделимых терминов, символов-заменителей и группировки терминов. Возможна фильтрация результатов по факту установления новой структуры в публикации и по таксономии на уровне царств. Результатом поиска являются соответствующие запросу публикации и их метаданные, а также список описанных в них структур, с привязкой к таксономии.

5. *Поиск сигналов 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 сходства. Возможна фильтрация по признаку нахождения атомов в пределах одного остатка. Результат аналогичен структурному поиску и содержит дополнительные ЯМР-данные, включая таблицы отнесения сигналов и метрику соответствия искомого фрагмента спектра экспериментальному спектру ЯМР.
6. *Поиск по идентификаторам (CSDB IDs).* Этот тип поиска позволяет найти записи, структуры, публикации, организмы и спектры по их идентификаторам в 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 на одном из распространённых языков кодирования биохимических данных для автоматической обработки.

Интерфейс поисковых запросов описан в справочной системе (http://csdb.glycoscience.ru/help/usage.html). Поиск активностей гликозилтрансфераз реализован в виде отдельного модуля, позволяющего перейти к описанным ферментам, синтезируемым ими структурам, используемым субстратам и другим данным, присутствующим в CSDB, и к записям фермента и кодирующего его гена в протеомных и геномных базах данных. Взаимосвязь поисковых критериев с получаемым результатом представлена на Рис. 4.

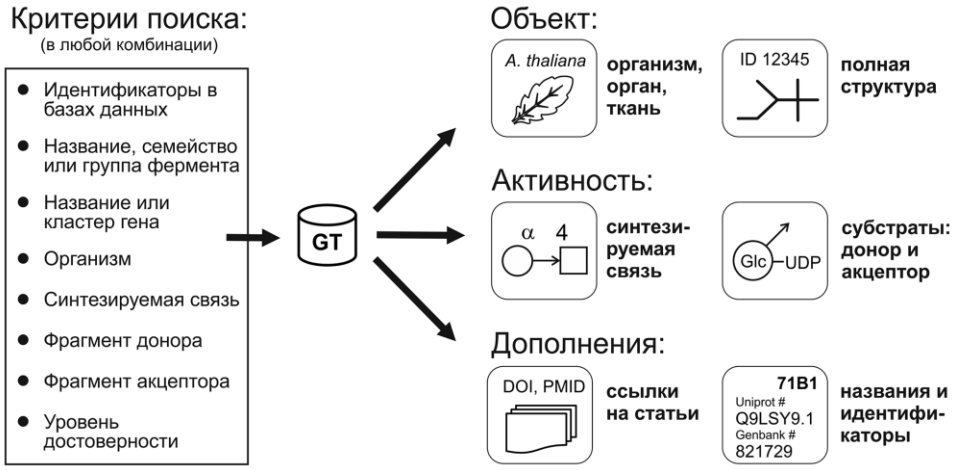


Рис. 4. Получение информации о биосинтезе углеводов с помощью модуля CSDB GT.

1.3 Описание углеводных структур

1.3.1 Кодирование структур. Был проведён анализ 11 углеводных и 5 общехимических языков по критериям полноты, однозначности, человекочитаемости, машиночитаемости, поддержки неопределённостей в структуре. В процессе работы над организацией заполнения

дампа и операций ввода-вывода была разработана нотация CSDB Linear, исправляющая недостатки конкурентов. Она представляет собой язык описания гликанов и их производных (включая экзотические случаи) в строковой форме. Пользователю знание этого языка требуется лишь для создания сложных запросов или добавления данных в CSDB. В рамках интерфейса CSDB возможен перевод с CSDB Linear на языки IUPAC Extended (для визуализации), SNFG (для визуализации), WURCS и GlycoCT (для взаимодействия с другими базами), SMILES (для перехода к атомарному описанию и молекулярной геометрии), GLYCAM (для моделирования конформаций), а также обратный перевод с языка GlycoCT, используемого несколькими другими проектами, на CSDB Linear.

CSDB Linear использует распространённый в гликоинформатике подход, кодирующий молекулы в виде направленных графов, в которых остатки соответствуют вершинам, а связи между ними – рёбрам. Графы записываются в виде текста и используются для аннотирования структур в дампе, импорта и экспорта, контроля ошибок и прямого ввода сложных структур.

Структура биогликанов 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.). Эти наименования комплектуются жёстко типизированными префиксами и суффиксами (аномерная и абсолютная конфигурация, размер цикла или признак полиола; модификации, связанные с окислением, восстановлением или функционализацией, Рис. 5А). 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. *Примеры*: `aDTalfA` = α-D-талофуранозуроновая кислота; `?XKdop` = кетодезоксиманноктоновая кислота в пиранозной форме с неизвестной аномерной конфигурацией; `xDManN-ol` = 2-амино-D-маннитол; `xRPyr` = пируват, приобрётший при связывании новый стереоцентр в R-конфигурации. Словарь мономеров, их классификация, атомарные свойства и систематические названия по IUPAC формализованы в виде специального сервиса (http://csdb.glycoscience.ru/database/core/residues.php).

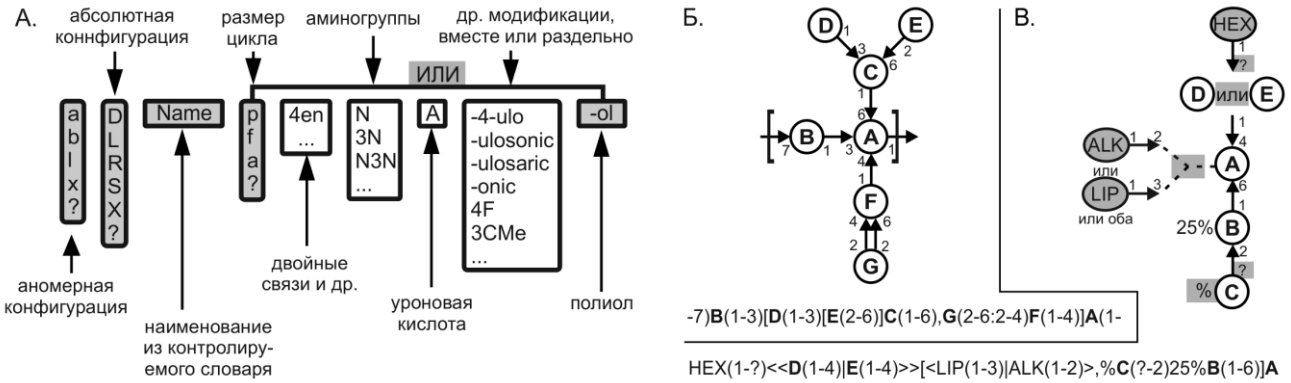


Рис. 5. Возможности нотации CSDB Linear. **А.** Компоненты кодировки остатков (обязательные показаны серым). **Б.** Топология и связность (A,B – основная цепь полимера; A,C – точки разветвления; E,D,G – терминальные остатки; G и F связаны двумя связями). **В.** Пример кодировки неоднозначностей вне остатков: суперкласс (HEX = любая гексоза), неизвестные позиции замещения (?), альтернативные фрагменты (D и E); альтернативные связи (ALK и LIP); нестехиометрические компоненты (B,C).

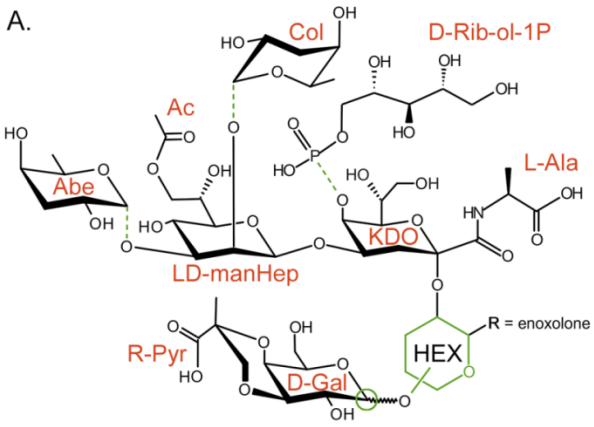
На топологическом уровне (Рис. 5Б) одна цепь остатков всегда является основной (в полимерах её выбор очевиден, для олигомеров разработаны правила сортировки цепей по старшинству). Боковые цепи перечисляются через запятую в квадратных скобках слева от остатка.

На уровне связности (Рис. 5Б) каждый остаток, кроме восстанавливающего конца, 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. 5 (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 предусмотрен специальный синтаксис для остатков фосфорной и серной кислот, для моновалентных агликонов на восстанавливающем конце, для C-C-связей между остатками в C-гликозидах, для связей в нестандартные положения, не кодируемые нумерацией углеродного скелета (напр., 3'), для второй связи с тем же остатком (напр., пируват или бифосфат). Если углеродный скелет остатка имеет части, не соединённые C-C-связями, можно предположить связь двух остатков 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 Ac(1-2)bDGlcpN, но не bDGlcpNAc) позволяет лучше формализовать первичную структуру и на два порядка снизить объем требуемого словаря мономеров. Значительная доля установленных природных структур содержит неоднозначности (Рис. 5В). В простых случаях они кодируются знаками «?» вместо конфигураций мономеров или положений связей. На уровне мономеров предусмотрены 15 суперклассов (напр., PEP = любая аминокислота) и возможность создавать собственные обозначения для неподдерживаемых или неявно заданных остатков. Предусмотрен специальный синтаксис для нестехиометрических боковых цепей и модификаций, а также для обозначения альтернативных фрагментов структуры (угловые скобки на Рис. 5В).

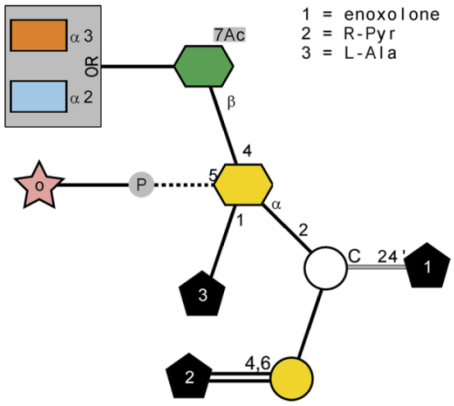
Структурные особенности, не поддерживаемые в CSDB Linear (чередование повторяющихся и уникальных звеньев; присоединение цепи к неизвестному узлу топологии; гетерогенность боковых цепей; указание степени полимеризации; циклические полимеры и др.), кодируются в дополнительных полях. Возможности (см. также Табл. 2) и синтаксис CSDB Linear документированы в справочной системе (http://csdb.glycoscience.ru/help/rules.html). Для демонстрации возможностей на Рис. 6 приведена структура со множеством характерных для биогликанов особенностей (А) и её кодировка в нотации CSDB Linear (Б).

1.3.2 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: текстовый формат IUPAC extended или графический CFG. Для текстового отображения CSDB использует модифицированный формат SweetDB (пример на Рис. 6В), зарекомендовавший себя в CarbBank. Он представляет собой формализованный псевдографический вариант IUPAC extended, в котором в рамках CSDB был повышен уровень строгости описаний мономеров и добавлены другие возможности.

А.



Г.



Б.

```
<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
```

В.

```
                                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
```

Рис. 6. А. Модельная структура, содержащая различные типы остатков (показаны красным) и неопределённостей (показаны зелёным). Пунктиром обозначены нестехиометрические связи. Б. Кодировка структуры на языке CSDB Linear. В. Визуализация структуры в формате SweetDB. Г. Визуализация структуры в формате SNFG.

Графическая нотация CFG, отображающая остатки в виде пиктограмм, была разработана в 1970-х годах и быстро набрала популярность, особенно в биологических и медицинских статьях. В настоящее время около 30% публикаций гликохимической тематики содержат структуры в этом формате. Однако с открытием множества новых моносахаридов и распространением гликомики в область углеводов прокариот возможностей нотации CFG стало не хватать. В коллаборации с другими членами консультативной группы по гликоинформатике при NCBI автором диссертации была разработана третья версия нотации, получившая название Symbol Nomenclature for Glycans (SNFG, пример для той же структуры см. на Рис. 6Г). В настоящее время её поддержали основные проекты гликоинформатики и рекомендовали к использованию ведущие углеводные журналы; появились компьютерные сервисы для перевода на SNFG с других углеводных языков (в том числе в проекте CSDB).

SNFG стандартизирует пиктограммы для 75 моносахаридов и 12 суперклассов, подобранные так, чтобы обеспечивать обратную совместимость с CFG. Наиболее распространённые N-ацетилпроизводные представлены как отдельные остатки, остальные модификации указываются текстом рядом с пиктограммами. Для каждого моносахарида существует конфигурация по умолчанию, а в случае если абсолютная конфигурация, состояние восстановления или размер цикла отличаются, это указывается символами внутри пиктограмм. Связи между остатками обозначаются линиями 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. Версия SNFG, использованная в CSDB, расширена по отношению к канонической для обеспечения возможности визуализировать любые структуры из CSDB, в том числе содержащие неопределённости. Описание SNFG дано на странице NCBI, посвящённой номенклатуре углеводов (https://www.ncbi.nlm.nih.gov/glycans/snfg.html).

1.3.3 Атомарное описание. Возможность использования общехимического программного обеспечения для природных углеводов долгое время ограничивал медленный ручной перевод семантического описания в атомарное. На платформе CSDB реализован алгоритм перевода семантических описаний в наиболее распространённый формат хемоинформатики (SMILES) с учётом неопределённостей в структурах. Основные шаги этого процесса представлены на Рис. 7.

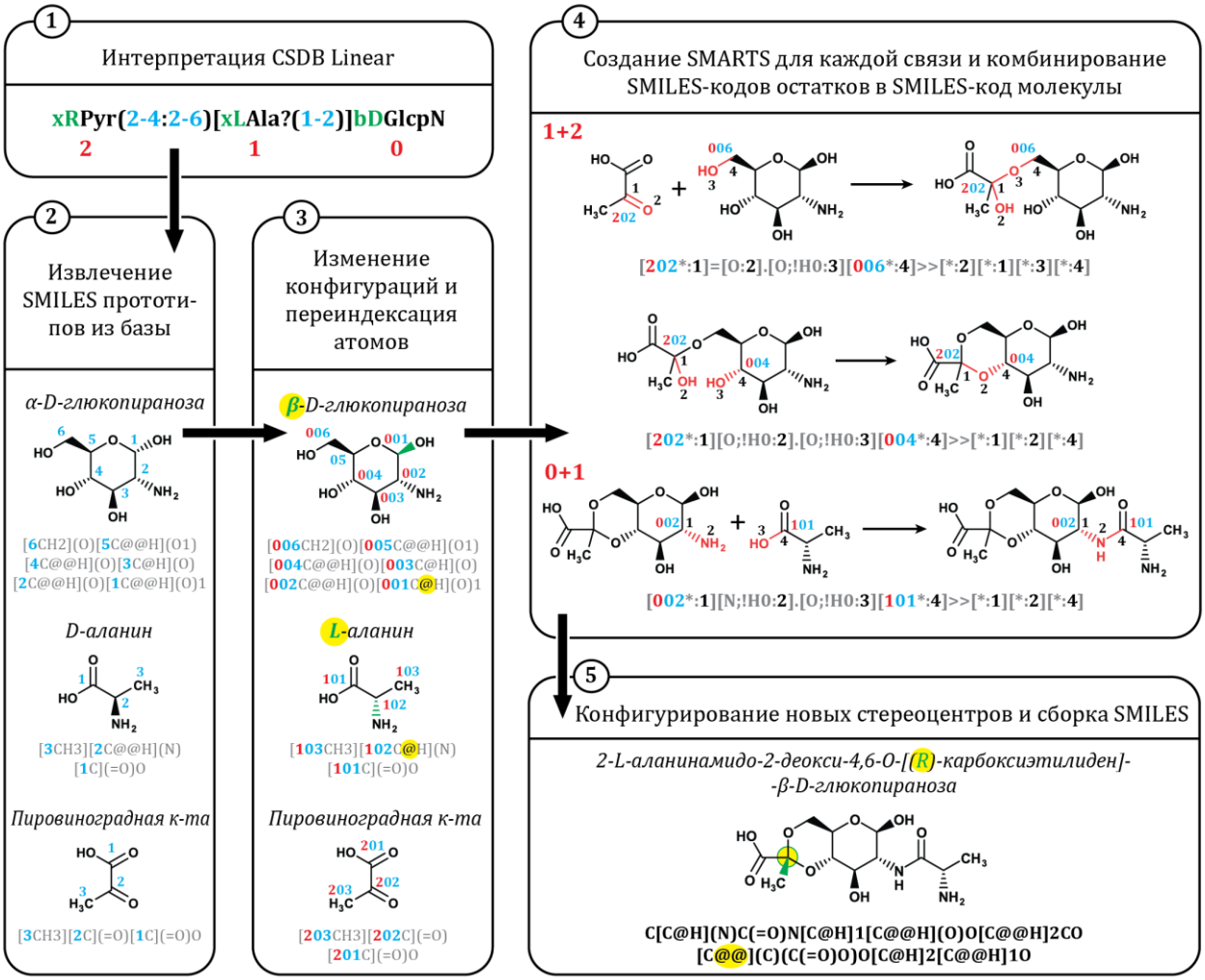


Рис. 7. Получение поатомного описания (SMILES) из семантического (CSDB Linear) на примере R-пирувата β-D-глюкопиранозамина, амидирующего L-аланин. Красными числами пронумерованы остатки, голубыми – атомы в остатках, черными – положения в реакциях SMARTS. Конфигурации остатков обозначены зелёным, изменяемые стереоконфигурации атомов – жёлтым.

Остатки, поддерживаемые в CSDB Linear, сводятся к 943 прототипам (немодифицированным остаткам в D-форме, со свободными аминогруппами и в определённой конфигурации). Атомарные описания прототипов, включая каноническую нумерацию атомов, получены заранее и сохранены в базе данных. Для перевода произвольной структуры в атомарное описание код CSDB Linear подвергается синтаксическому анализу и интерпретируется в псевдоструктуру, где каждый остаток представлен отдельным объектом, содержащим всю информацию о конфигурациях и связях с другими остатками. К каждому такому объекту добавляется код SMILES, полученный модификацией прототипа с учётом конфигураций стереоцентров. Атомы перенумеруются (Рис. 7, 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 с учётом позиций, в кото-

рых каждый остаток образует связь. Если при этом возникает новый стереоцентр, его конфигурация берётся из исходной 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, либо когда неопределённость подразумевает изомерию 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. Возможности обработки структурных характеристик природных углеводов и их производных перечислены в Табл. 2. Они систематизированы с использованием четырёх уровней детализации: остатков, связей, топологии и неопределённостей. Разработанная схема позволяет адекватно перевести в атомарное описание те из них, которые поддерживаются в CSDB Linear.

Табл. 2. Возможности кодировки CSDB Linear и её перевода в атомарное описание.

Особенность	*кодируется*	*перевод в SMILES*	*Примеры в нотации CSDB Linear*	*Комментарии*
Уровень остатков				
Моносахариды	+	+	aDGlcpN, aXKdop, aXLDmanHepp	234 прототипа остатков *
Пиранозы, фуранозы и ациклические формы	+	+	aDGlcp, aDGlcf, aDGlca	
Полиолы, инозитолы и альдоновые кислоты	+	+	xDGro, xDGlcN-ol, xXmyoIno; myoIno = *мио*-инозитол	63 прототипа остатков
Фосфаты и сульфаты	+	+	xXEtN(1-P-P-5)[P-4)]aXKdop, aDGlcp(1-P-5)xDRib-ol; EtN = этаноламин	как терминальные, так и в цепи
Жирные кислоты	+	+	lXPam, lX3HOLau; Pam = пальмитиновая к-та, 3HOLau = 3-гидроксилауриновая к-та	167 прототипов остатков (включая 25 сфингозидов)
Аминокислоты	+	+	xLLys, xXPmN2, xXSRCetLys; SRCetLys = S,R-карбоксиэтиллизин, PmN2 = диаминопимелиновая к-та	42 прототипа остатков
Прочие неуглеводные остатки	+	+	xSPyr, xXCho, xXSuc; Pyr = пировиноградная кислота, Cho = холин, Suc = янтарная к-та	70 прототипов остатков (включая 9 нуклеотидов)
Агликоны и подстановки свободного текста	+	+/–	aDGlcp(1-3)Subst // Subst = enoxolone	представлены в SMILES как изотопно-меченные псевдоатомы
Суперклассы остатков	+	+/–	LIP(1-3)xDGro(1-P-2)HEX	28 суперклассов; представлены в SMILES как изотопно-меченные псевдоатомы

Уровень связей				
Модификации остатков	+	+	Ac(1-2)[Me(1-3)]aDGlcp	алкилирование, ацетилирование и др.
Две связи между остатками	+	+	xSPyr(2-4:2-6)aDGalp	бифосфаты, О-пируваты, 1,1-связанные ацетали моносахаридов
Связи C-C и C-N	+	+/–	Me(1C-3)aDGlcp	C-гликозиды, N-гликаны; (кроме C-C-связей с неопределённой позицией связывания)
Нестехиометрические связи	+	+/–	-4)[30%Ac(1-3),xXEtN(1-%P-6)]aDGlcp(1-	30% глюкозы ацетилировано, неизвестная её доля фосфорилирована; SMILES генерируется для структур со 100% стехиометрией
Сложноэфирные и амидные связи	+	+	Ac(1-2)xLLys(1-2)aDGalpN	
Уровень топологии				
Олигомерные структуры	+	+/–	aDGlcp(1-2)bDFruf	
Регулярные полимеры	+	+/–	-9)[Ac(1-5)]aXNeup(2-	границы повторяющегося звена представлены псевдоатомами
Циклические полимеры	+	–	CYCLO -4)bDGlcp(1-	поддержка в отдельном поле CSDB «тип молекулы»
Нерегулярные полимеры	–	–		
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	
Неизвестные абсолютные конфигурации	+	+	a?Rhap	для остатков с единственным хиральным атомом атом попадает в SMILES неопределённым; для мультихиральных остатков генерируются два энантиомера (D/L)
Неизвестный тип циклизации	+	+	bDGal?	генерируются возможные изомеры (разные структуры)
Неизвестное положение связи	+	+	aDGlcp(1-?)aLRhap	генерируются все химически разрешённые структуры
Альтернативные фрагменты	+	+	<Ac(1-2)|Me(1-3)>aDGlcp; -3)<<aDGlcp(1-4)|aDGalp(1-4)>>aLRhap(1-	поддерживается логика «ИЛИ» и «ЛИБО», число альтернатив не ограничено; в SMILES генерируется несколько структур
Неизвестный узел присоединения боковой цепи	–	–		
Известен только мономерный состав	–	–		

1.3.4 Молекулярная геометрия. Коды SMILES обрабатываются модулем моделирования 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. 8, а интерфейс веб-инструмента - на

Рис. ***9***.

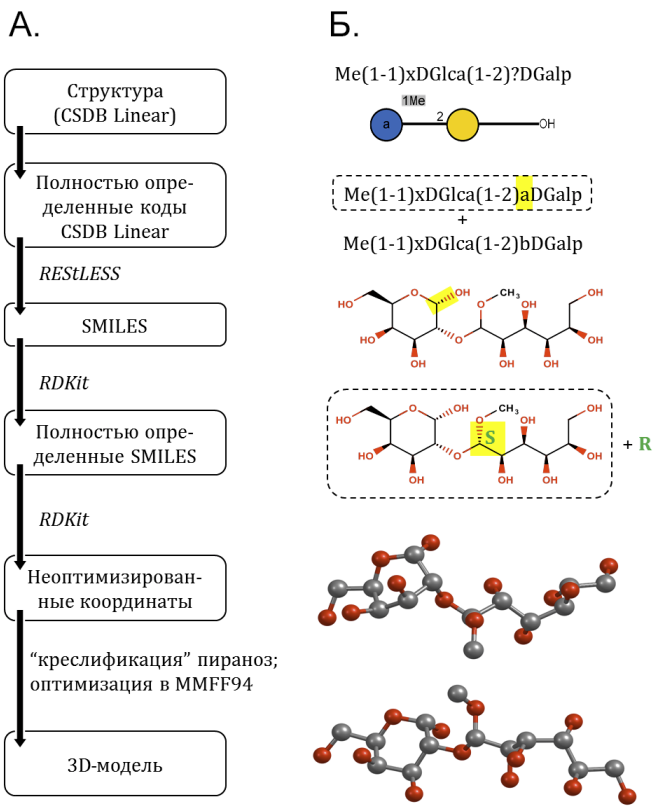


Рис. 8. (↑) А. Алгоритм получения пространственной модели на основе семантического описания. Б. Объекты, используемые на каждом шаге на примере одной из четырёх возможных структур 1-О-метил-D-глюкозил-2-D-галактопиранозы. Объекты, используемые на следующем шаге, обведены. Жёлтым обозначен выбор одной из двух стереоконфигураций (α/β и R/S).

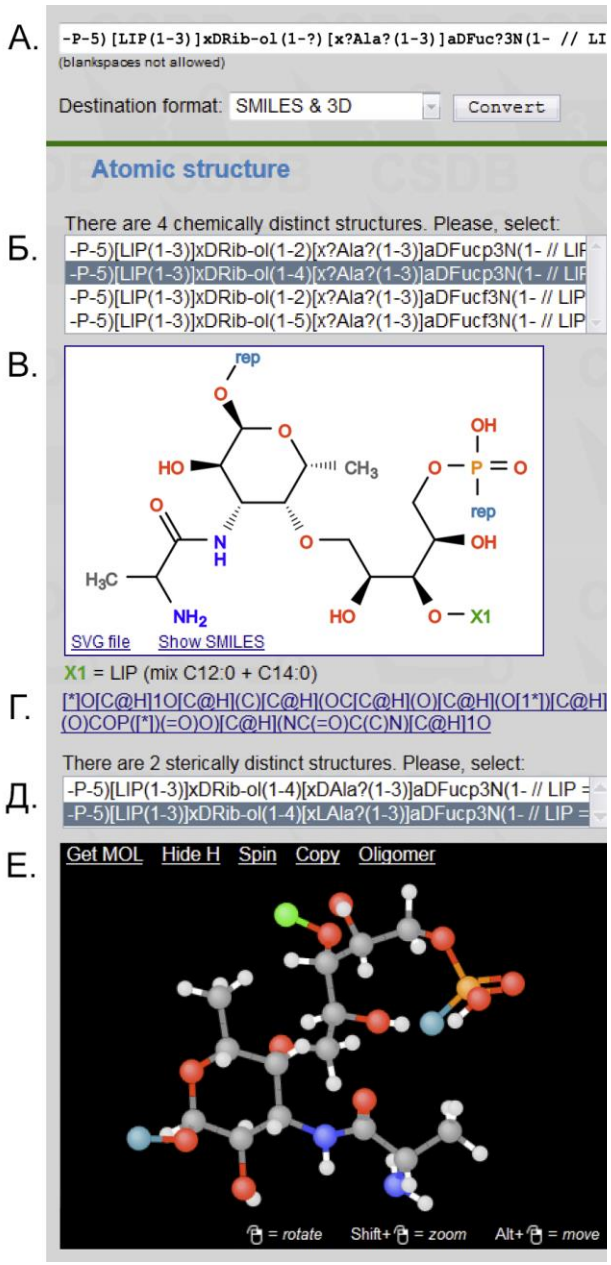


Рис. 9. (→) Интерфейс модуля моделирования геометрии. А. Структура, записанная в CSDB Linear. Б. Выбор из химически различных вариантов структуры. В. Сгенерированная формула выбранного варианта. Г. SMILES выбранного варианта. Д. Выбор из возможных стереомеров. Е. Оптимизированная модель выбранного стереомера и инструменты работы с ней.

Коды SMILES трансформируются в наборы молекул с полностью определёнными конфигурациями стереоцентров, и для каждой молекулы геометрия предсказывается и записывается в формате MOL. Было обнаружено, что существующие программы для потокового моделирования молекулярной геометрии ошибаются в конформациях пираноз в составе более сложных структур, прогнозируя «ванну», твист-форму или инвертированное «кресло». Для устранения проблем, привносимых неправильной начальной конформацией, 381 моносахарид в пиранозной форме был обсчитан с помощью высокотемпературной молекулярной динамики. 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 на молекулярно-динамической траектории. После внедрения в MOL правильных конформаций мономеров они объединяются друг с другом в полную структуру, которая затем оптимизируется релаксацией в силовом поле MMFF94. Этот инструмент предназначен для получения начальных геометрий для последующих ресурсоёмких расчётов молекулярно-механическими или квантово-механическими методами.

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 их валидации на основании экспериментального NOE (ядерный эффект Оверхаузера) значения торсионных углов, соответствующие минимумам конформационных карт, будут использоваться для моделирования взаиморасположения остатков в каждом фрагменте полной структуры. Наличие нескольких минимумов в конформационных картах подразумевает параллельные расчёты на основании многих начальных геометрий.

Для природных структур, 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 числа возможных структур. Для решения проблемы своевременного предоставления моделей структур в CSDB их варианты с полностью определёнными конфигурациями (43910 молекул) были рассчитаны заранее и сохранены. Для произвольных структур, введённых пользователем, расчёт проводится только первый раз, после чего результаты также помещаются в кэш, из которого извлекаются при последующих запросах.

1.4 Обработка данных и прогнозирование

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

1.4.1 Моделирование спектров ЯМР. Для создания инструментов помощи экспертам в интерпретации спектров природных углеводов были улучшены существующие и разработаны новые подходы к теоретическому расчёту химических сдвигов ЯМР. Были проанализированы эмпирические, статистические, квантово-механические, регрессионные и нейросетевые подходы к предсказанию наблюдаемых данных ЯМР. Первые два были выбраны как имеющие наибольший потенциал для повышения точности расчётов в химии углеводов.

Эмпирическая схема расчёта спектров ЯМР ^{13}C углеводов, известная более 25 лет и доведённая до практического использования в рамках кандидатской диссертации автора, была расширена на все классы природных углеводов и родственных соединений, дополнена данными по теоретическим эффектам замещения и химическим сдвигам в олигомерных фрагментах, дополнена модулем оценки достоверности модели и снабжена веб-интерфейсом. В современной реализации (BIOPSEL) она представляет собой инкрементную схему, основанную на 9-13 дескрипторах уровня остатков и учитывающую отклонения от аддитивности химических сдвигов, привнесённые стерическим влиянием соседних заместителей. Для расчётов используются спектры олигосахаридов и эмпирические эффекты замещения, полученные усреднением наблюдаемых данных при различных комбинациях дескрипторов. Эти данные извлечены из литературы и включают 215 эффектов замещения, спектральные характеристики 434 мономеров и 2245 димеров и тримеров.

Появление обширной и регулярно пополняемой базы данных CSDB, содержащей более 9400 спектров 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, 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. Разработанная схема предсказания химического сдвига конкретного атома подразумевает следующие шаги:

1. Выделение из структуры фрагмента, содержащего остаток, включающий предсказываемый атом, и соседние остатки. Биохимический смысл термина «остаток» (часть структуры, соединяющаяся с другими частями в результате реакций с отщеплением воды) в большинстве случаев совпадает с ЯМР-спектроскопическим смыслом (изолированная протон-углеродная спиновая система).
2. Многошаговое последовательное обобщение структурных характеристик (дескрипторов) фрагмента, начиная с наиболее удалённых от предсказываемого атома, происходящее по пути 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 основании суммарного веса проведённых обобщений и дисперсии значений из базы.

Для получения полного спектра структуры её атомы моделируются независимо. При разбиении структуры на фрагменты образуются подструктуры, содержащие центральный остаток (включающий предсказываемый атом) и соседние остатки (пример показан на Рис. 10А). Взаимное влияние фрагментов исключается их достаточным размером 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.

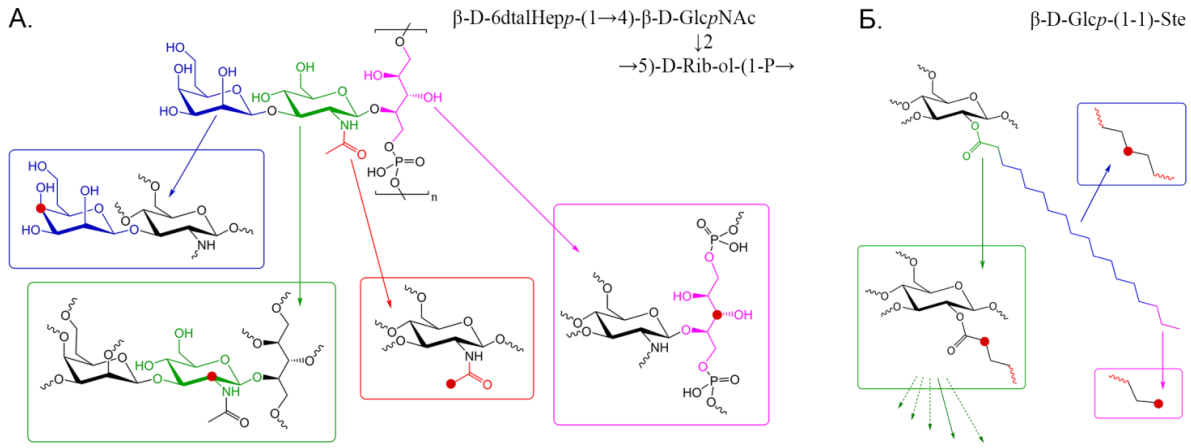


Рис. 10. А. Фрагментация структуры. «Центральные» остатки выделены цветом. В каждом фрагменте показан один из предсказываемых атомов (красная точка). Б. Разбиение алифатических цепей на структурно различные области.

Полученные фрагменты обобщаются для поиска содержащих их структур. Обобщением называется изменение набора структурных дескрипторов, после которого ему соответствует большее число структур. Например, превращение β-D-Fuc*p* в D-Fuc*p* является обобщением дескриптора «аномерная конфигурация», так как первому случаю соответствует одна структура, а второму – две (α и β). Обобщаемые дескрипторы центрального и прилежащих остатков включают:

- стереоконфигурации и типы всех атомов углерода;

- 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).

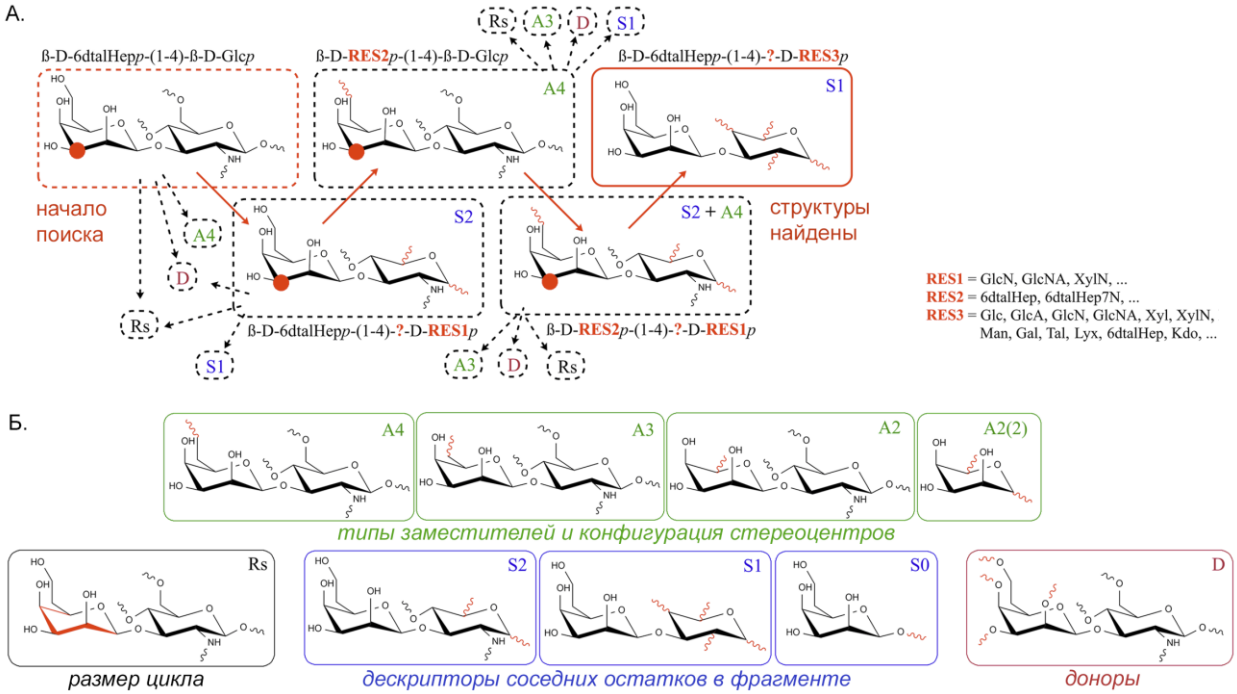


Рис. 11. А. Найденный путь генерализации (красные стрелки) на примере моделирования сигнала C3 6-дезокситалогептозы в структурном фрагменте β-D-6dTalHep*p*-(1→4)-β-D-Glc*p*NAc. Пунктирными стрелками показаны другие возможные пути. Б. Элементарные обобщения. Обобщаемые дескрипторы показаны красным. A, R, S, D – тип дескриптора, число после символа – удалённость дескриптора от предсказываемого атома либо от точки образования связи с остатком, содержащим дескриптор.

Пример многошагового обобщения фрагмента с использованием разных дескрипторов показан на Рис. 11. 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. Атомы типизированы в 11 категорий, отражающих гибридизацию, связанный гетероатом, и пребывание в составе функциональной группы. Стереоконфигурация может принимать одно из пяти значений (D или L по Фишеру, ахиральный атом, неизвестная, неизвестная для всего экзоциклического «хвоста»).

Когда центральный остаток образует связь с другими остатками, их дескрипторы также обобщаются. Для ускорения предсказаний обобщение параметров этих остатков происходит ступенчато (Рис. 12). На каждом шаге обобщается определённая доля дескрипторов остатка, оказывающих наименьшее влияние на химический сдвиг (имеющих наименьший вес).

Каждому параметру фрагмента, который может быть подвергнут обобщению, ставится в соответствие эмпирический весовой фактор («вес»), отображающий влияние этого параметра на предсказываемый атом. Вес обобщения зависит от числа связей между положением де-

скриптора и предсказываемым атомом, 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. Например, имея спектры структур α-D-Man*p*-(1→4)-α-D-Glc*p* и α-D-Tal*p*-(1→4)-β-D-Glc*p*, требуется предсказать химический сдвиг атома C1 глюкозы в дисахариде α-D-Tal*p*-(1→4)-**α**-D-Glc*p*. Использование для предсказания второго дисахарида приведёт к заметной ошибке, в то время как использование маннозосодержащего дисахарида даст хороший результат, т.к. его единственным отличием от целевой структуры является конфигурация удалённого от Glc C1 четвёртого атома остатка талозы. 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 был использован генетический алгоритм ABC («искусственная пчелиная колония») с отбором по средней ошибке моделирования на большой выборке структур с известными спектрами.

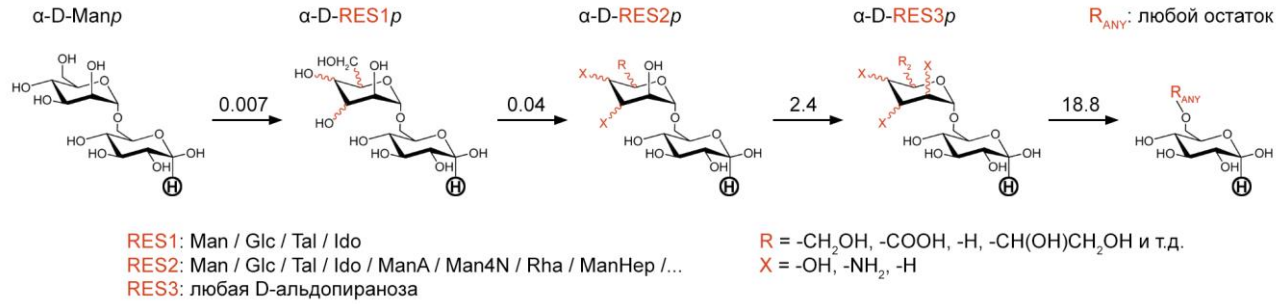


Рис. 12. Пример обобщения остатка-донора в фрагменте α-D-Man*p*-(1→6)-β-D-Glc*p*. Обобщаемые дескрипторы показаны красным. Веса приведены над стрелками для случая предсказания протона Glc H1. Шаги обобщения: 1 - стереоконфигурации удалённых атомов донора, 2 – типы заместителей при этих атомах, 3 – конфигурация и тип ближайшего к связи атома, 4 – весь остаток.

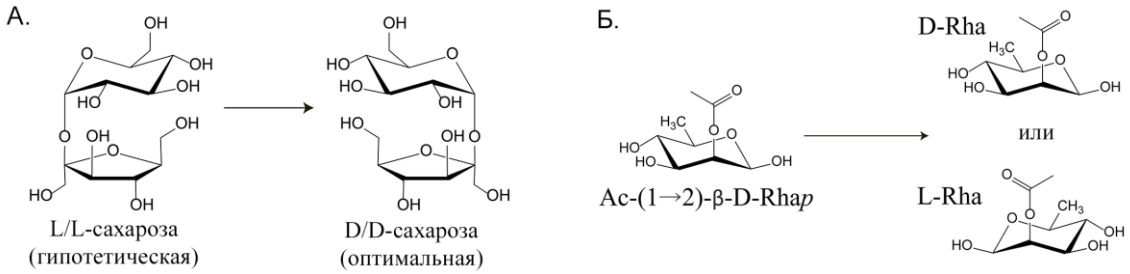


Рис. 13. Примеры предварительных обобщений с нулевым весом: А. Инверсия абсолютных конфигураций всех остатков в фрагменте. Б. Обобщение абсолютных конфигураций остатков, не имеющих хиральных заместителей.

При обобщении фрагментов требуется соблюдать баланс между максимизацией числа подходящих структур в базе данных и минимизацией влияния обобщений на результирующий химический сдвиг. Спектры ЯМР энантиомеров в ахиральном окружении совпадают, и инверсия абсолютных конфигураций каждого остатка в фрагменте может увеличить число подходящих структур без уменьшения точности предсказания. Один из двух альтернативных наборов абсолютных конфигураций фрагмента выбирается на основании данных о встречаемости остатков (Рис. 13А). Если все соседние остатки оптически неактивны (напр., остаток уксусной кислоты на Рис. 13Б), обобщение абсолютной конфигурации остатка не повлияет на спектры фрагмента (т.е. имеет нулевой вес), повышая при этом число подходящих структур в базе.

Природные гликоконъюгаты часто содержат остатки с большим углеродным скелетом, что приводит к лавинообразному нарастанию числа возможных обобщений и критическому снижению скорости моделирования. В то же время остатки, состоящие из одинаковых многократно повторяющихся групп атомов, не требуют тонко настроенной для углеводов схемы обобщений. Для наиболее распространённых случаев – жирных кислот, спиртов и сфинголипидов – был разработан специальный алгоритм обобщений. Остатки разбиваются на три сектора (Рис. 10Б): «голову» (ближайшие два атома), «хвост» (концевые группы из трёх 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. 10Б). Обобщения остатка для остальных атомов игнорируют остаток, связанный с «головой», и алгоритм ищет в базе фрагменты, включающие атомы в пределах двух связей от предсказываемого.

После поиска 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 явных и скрытых отклонений с помощью модифицированного критерия Шовене. Проблема калибровки шкалы исключается единым стандартом химических сдвигов в CSDB.

Большинство исследований природных углеводов проводится в водных растворах, поэтому по умолчанию моделирование предполагает воду в качестве растворителя. Для моделирования спектров растительных гликоконъюгатов, часто регистрируемых в других растворителях, введён параметр, ограничивающий растворитель при выборке данных. Точность статистического моделирования зависит от заполненности базы данных спектрами, снятыми в конкретном растворителе. Наиболее распространёнными растворителями в CSDB являются вода и пиридин; подробная статистика доступна по ссылке «Coverage» на странице ЯМР-модуля (http://csdb.glycoscience.ru/goddess.html). Дополнительные параметры позволяют задать ограничения на кислотность среды и температуру.

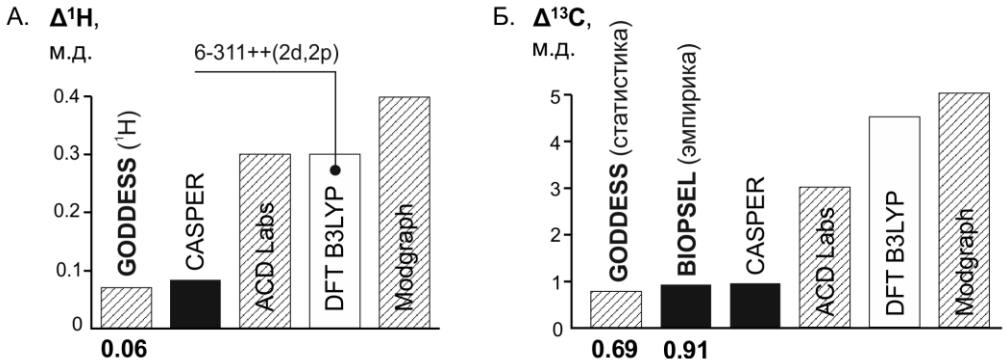


Рис. 14. Сравнение точности GODDESS и BIOPSEL с другими алгоритмами ЯМР-моделирования. Эмпирические подходы показаны черным, статистические – штриховкой. Средняя точность разработанных алгоритмов на природных углеводах (м.д.) продублирована в нижней строке. А. Моделирование спектров ^{1}H; Б. Моделирование спектров ^{13}C.

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 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 методов. По результатам этого сравнения можно сделать вывод о существенном преимуществе методов, оптимизированных для углеводов, перед методами, основанными на использовании HOSE, нейронных сетей и квантово-механических расчётов на высоких уровнях теории в больших базисных наборах, которые считаются достаточными для ЯМР-моделирования биоорганических соединений (Рис. 14). Из оптимизированных методов CASPER демонстрирует сравнимую точность, но поддерживает ограниченный набор структурных особенностей, не покрывающий разнообразия углеводов прокариот.

Поддержка большинства структурных особенностей природных биогликанов является отличительной характеристикой GODDESS. Так, с помощью инструмента GlyNest удалось предсказать спектры ЯМР только для глюкозы и модельных структур; CASPER имеет ограниченный набор поддерживаемых компонентов структуры; инкрементный подход, основанный на BIOPSEL, моделирует только спектры ЯМР ^{13}C и только в воде. Как эмпирические методы общего назначения, так и методы *ab initio* не поддерживают предсказания спектров ЯМР полимеров; кроме того, последние работают в 10000-100000 раз медленнее. Предсказание спектра ЯМР типичной природной структуры методом GODDESS занимает около одной минуты.

Характерные примеры суперпозиции предсказанных и экспериментальных сигналов приведены на Рис. 15. В качестве модельных структур выбраны бактериальные гликополимеры, структура которых была установлена автором ранее. В случае химических сдвигов ^{1}H совпадение сигналов находится в пределах экспериментальной погрешности. По углеродной оси в HSQC визуально заметные несовпадения наблюдаются для атомов, сигналы которых зависят от pH раствора (C2/H2 лизина, C5/H5 галактуроновой кислоты) и конформационно подвижного C6/H6 4,6-дизамещенной глюкопиранозы, подверженного стерическим эффектам.

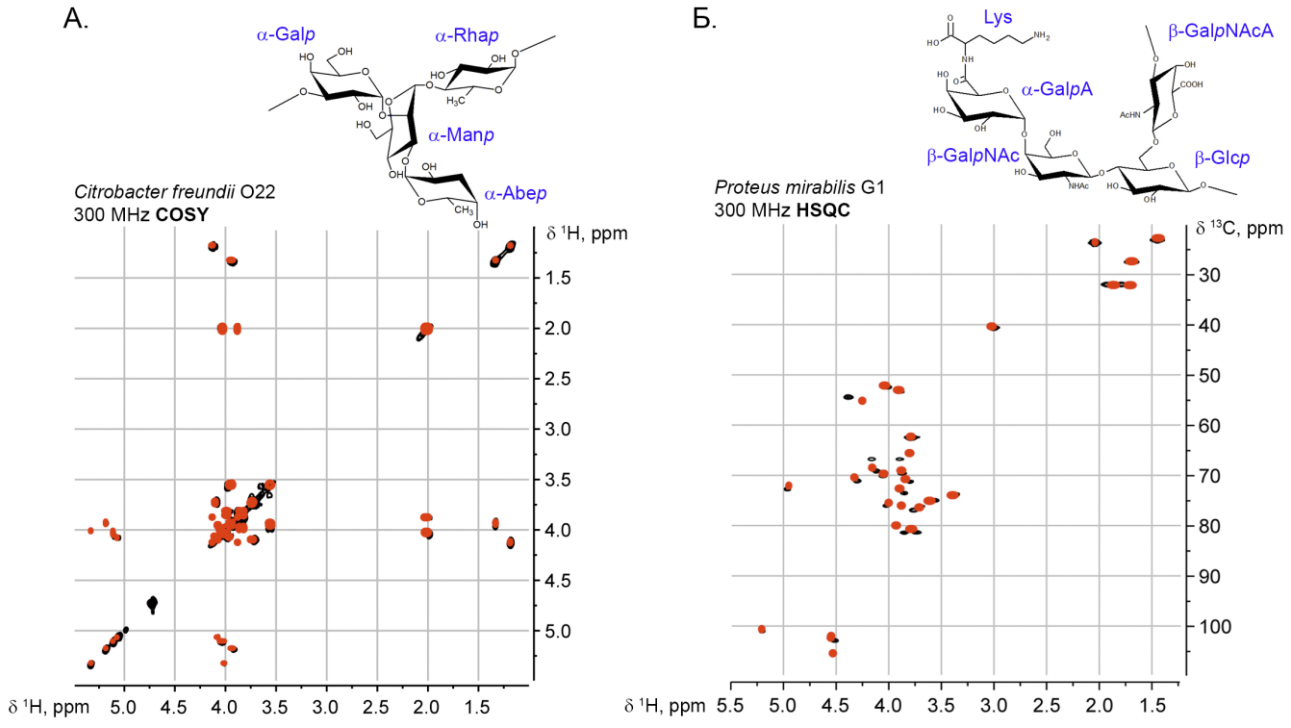


Рис. 15. Суперпозиция модели (красная) и экспериментального спектра (чёрный) на примерах А. COSY полисахарида *Citrobacter frendii* O22 (D_2O, 30 °C, 600 МГц); Б. ^{1}H,^{13}C HSQC полисахарида *Proteus mirabilis* G1 (D_2O, 45 °C, 500 МГц).

Для статистического анализа точности моделирования из базы CSDB было отобрано 36385 химических сдвигов ^{13}C ЯМР, моделирование которых поддерживается как эмпирическим, так и статистическим методами, и 40441 химических сдвигов ^{1}H ЯМР. Критерием отбора было отсутствие неопределённостей в опубликованной структуре и доступность экспериментальных спектров ЯМР в водных растворах. Если спектр, с которым сравнивалась полученная модель, присутствовал в CSDB, эта запись временно удалялась из базы данных для предотвращения искажения результатов из-за «правильного подбора» примера.

На Рис. 16 приведены распределения абсолютных отклонений предсказанных значений от экспериментальных, а также 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. Эмпирический подход (только ^{13}C) продемонстрировал среднюю

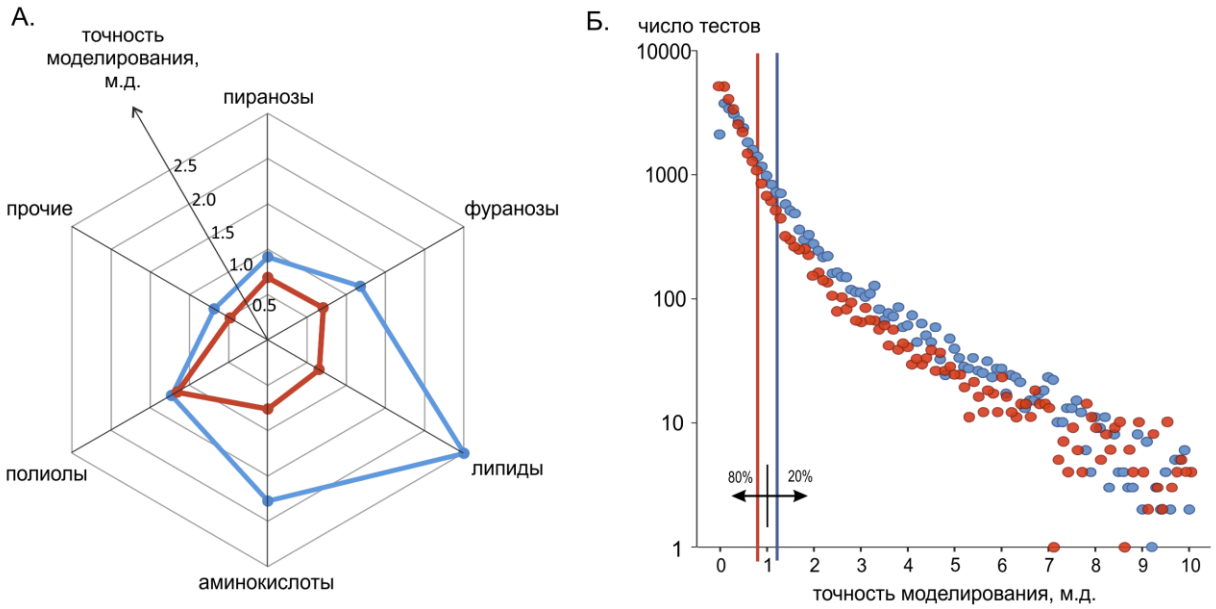


Рис. 16. А. Точность моделирования углеродных химических сдвигов для остатков разных классов. Голубые точки и линии – эмпирическое моделирование, красные – статистическое. Б. Распределение точности моделирования ~36000 атомов в природных структурах. Вертикальные линии показывают уровень, разделяющий 80% лучших и 20% худших моделей.

точность 0.91 м.д., для 95% предсказаний ошибка лежала в пределах 3.0 м.д. 80% химических сдвигов ^{13}C было предсказано с отклонением, меньшим чем 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 и нехватки данных по эффектам замещения плохо поддаются теоретическому анализу: фураноз, липидов и аминокислот (Рис. 16А).

Достоверность предсказания оценивается для каждого атома и нормируется с получением числа от 0 до 100. В эмпирических расчётах она зависит от того, использовались ли сохранённые или теоретические эффекты замещения для данного структурного окружения, и от числа пермутаций дескрипторов, которое потребовалось для нахождения в базе химического сдвига или эффекта. При использовании статистического метода достоверность зависит от суммарного веса применённых обобщений и от размера и дисперсии выборки химических сдвигов. Эта зависимость формализована в виде суммы полиномов с коэффициентами, итеративно подобранными по критерию максимизации корреляции между величинами достоверности и

наблюдаемой ошибки предсказания. Для перевода значения достоверности в ожидаемую ошибку моделирования с помощью регрессии получены линейные зависимости.

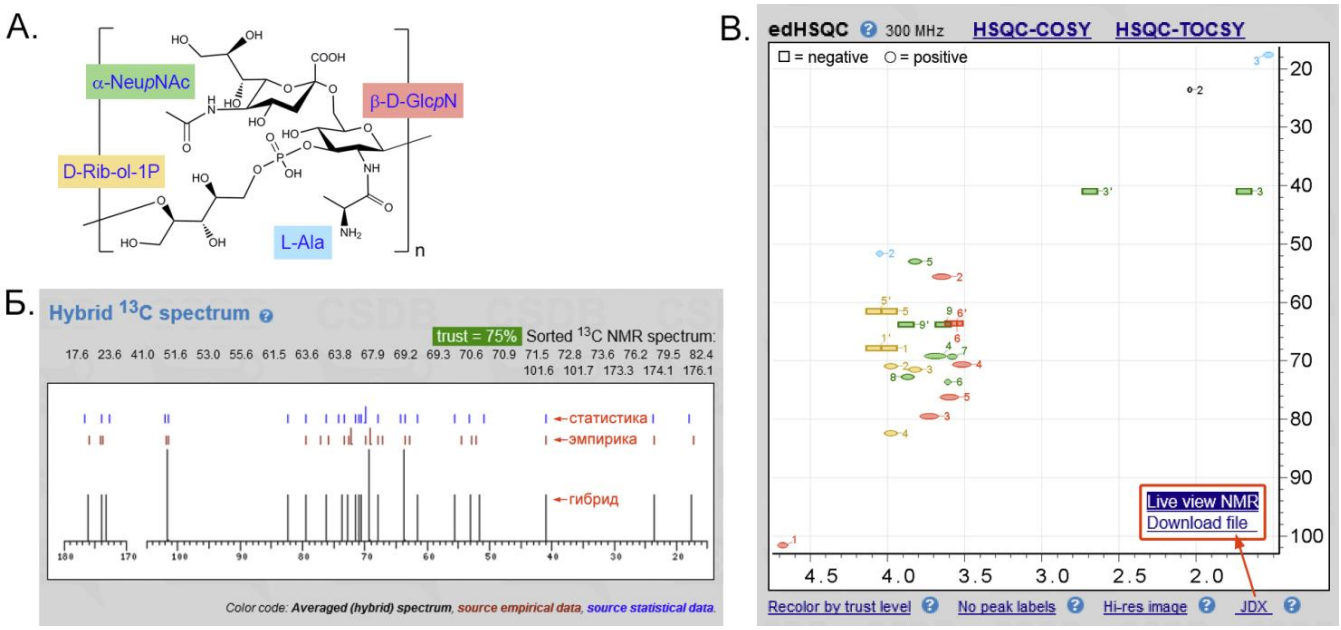


Рис. 17. Вывод предсказанных спектров для структуры, содержащей остатки типичных для биогликанов классов. А. Структура и цветовой код для отнесения сигналов. Б. Одномерный спектр ЯМР ^{13}C. В. ^{1}H,^{13}C edHSQC как один из предсказанных двумерных спектров, отображённый в режиме отнесения. Цвет кодирует остаток, число – номер атома в остатке.

Из-за неравномерной полноты баз данных атомы в одной и той же структуре предсказываются каждым из подходов с разной достоверностью в зависимости от химического окружения. Для устранения связанных с этим погрешностей была разработана гибридная модель углеродных спектров. Она подразумевает линейное смешивание эмпирических и статистических предсказаний на основании значений химических сдвигов и их достоверностей, сгенерированных каждым из методов. Гибридная достоверность учитывает, в какой степени одна модель подтверждает или опровергает другую. Выведены и обоснованы аналитические формулы для получения гибридных данных. Пользователю предоставляются все три варианта углеродных химических сдвигов и достоверностей моделирования (Рис. 17Б). В качестве характерного примера можно привести моделирование спектра ЯМР ^{13}C общего антигена энтеробактерий (Рис. 18). Несмотря на распространённость этого трисахаридного повторяющегося звена в гликоме прокариот, инфицирующих млекопитающих, в других структурах такие комбинации остатков встречаются редко, поэтому соединение является сложным случаем для статистического предсказания. Гибридное моделирование позволило получить максимальную корреляцию с экспериментальным спектром ЯМР ^{13}C (Рис. 18Б). 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.

Предсказанные химические сдвиги визуализируются в виде таблиц отнесения, одномерных углеродных и двумерных спектров. Разработанный модуль визуализации поддерживает основные гомо- и гетероядерные спиновые корреляции, востребованные в исследованиях углеводов, кроме NOESY/ROESY. Этот набор спектров доступен для работы в браузере, включая

экспорт и наложение предсказанных и экспериментальных спектров. Он позволяет химикам проверять структурные гипотезы и делать отнесение сигналов в спектрах сложных объектов с минимальными усилиями. Для предсказания размеров кросс-пиков используется эмпирическая оценка ширины сигнала по протонной оси, основанная на моделировании констант спин-спинового взаимодействия (КССВ), исходя из торсионных и валентных углов между предсказываемым атомом и соседними протонами. Сумма этих КССВ определяет ширину кросс-пика с учетом частоты спектрометра.

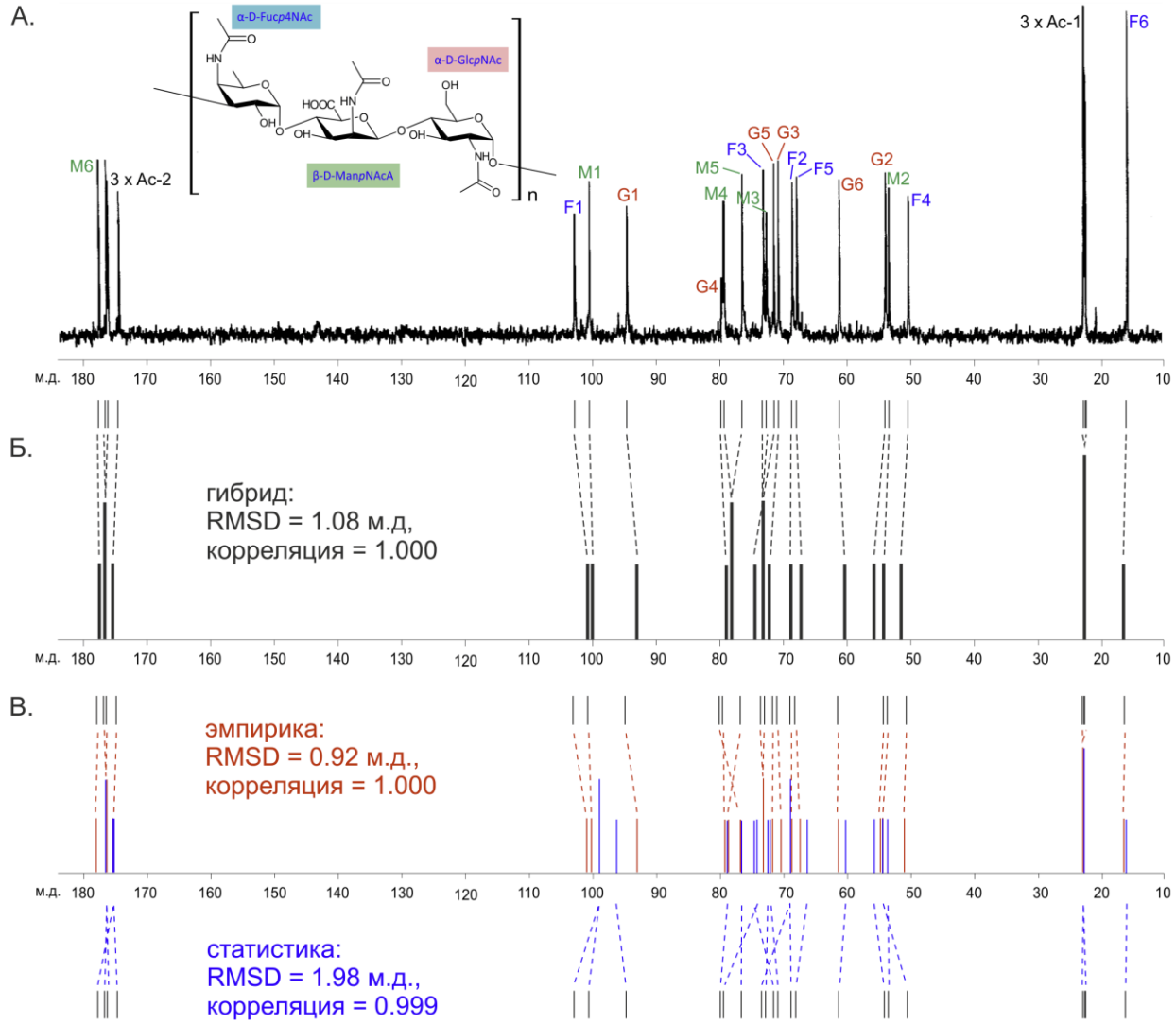


Рис. 18. Моделирование спектра ЯМР ^{13}C общего антигена энтеробактерий. А. Структура (из *Proteus penneri* 17), экспериментальный спектр в D_2O и отнесение сигналов. Б. Гибридная модель. Соответствие сигналов экспериментальным данным показано пунктиром. В. Эмпирическая (красный цвет) и статистическая (синий цвет) модели. *(RMSD - среднеквадратичное отклонение.)*

1.4.2 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 предсказания структур биогликанов на основании данных ЯМР и других экспериментов. Предсказание включает следующие шаги (Рис. 19):

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, и их ранжирование по степени соответствия между неотнесённым экспериментальным спектром ЯМР ^{13}C и моделью, полученной быстрым эмпирическим методом.
3. Моделирование спектров для 500 лучших гипотез относительно медленным, но более точным статистическим методом (GODDESS), оценка достоверности модели и её соответствия эксперименту, окончательное выявление наиболее вероятных структурных гипотез.

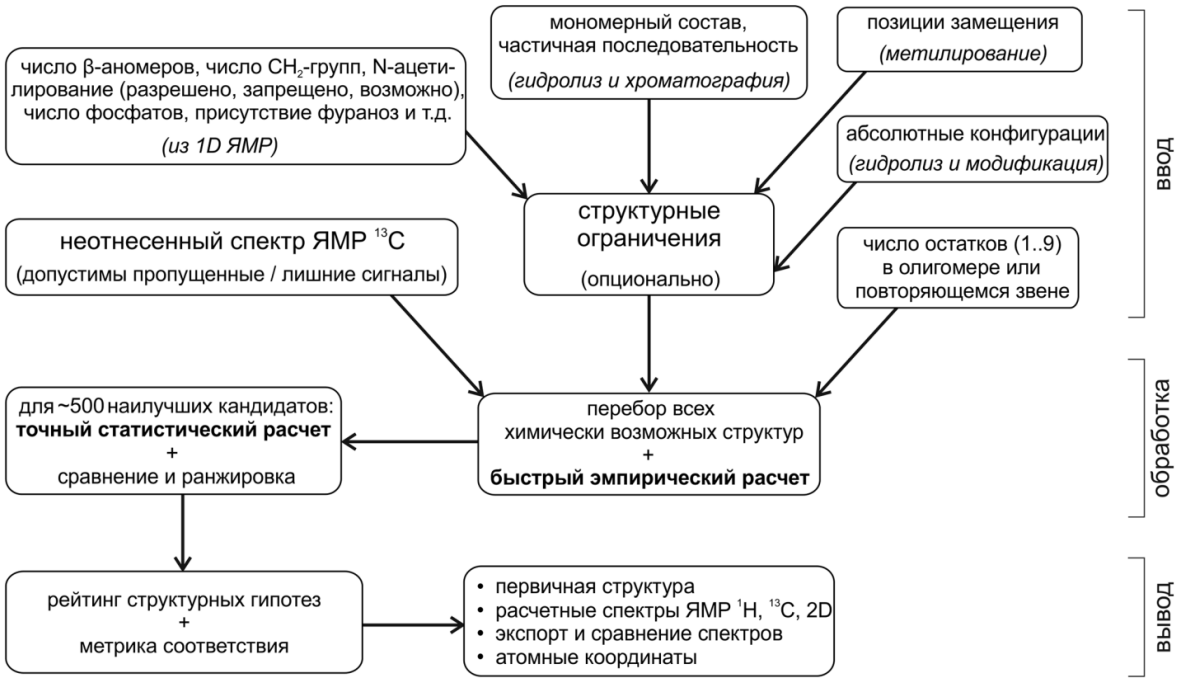


Рис. 19. Входные данные («ввод»), выходные данные («вывод») и последовательность шагов («обработка») алгоритма GRASS.

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. Типичным вариантом использования GRASS является установление особенностей строения (топологии структуры, последовательности остатков, аномерных конфигураций), с трудом поддающихся анализу без полного отнесения спектров ЯМР, на основании информации, 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 квадратных 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 или классы остатков (напр., «любая гексоза») [ГХ, ГЖХ, ВЭЖХ, МС].
- Структурная единица: олигомер или повторяющееся звено [хроматография при выделении].
- Общее число β-моносахаридов в структурной единице [ЯМР ^{1}H или анализ прямых КССВ J_{CH}]. Аномерная конфигурация любого из остатков также может быть задана в явном виде.
- Наличие или отсутствие фураноз [обзорный спектр ЯМР ^{13}C]. Способ циклизации любого остатка (пираноза, фураноза или линейная форма) также может быть задан в явном виде.

- Ограничения на ацетилирование аминогрупп каждого остатка: «обязательно», «возможно» или «запрещено» [ЯМР ^{13}C де-О-ацетилированного образца; HSQC при изменённом pH].
- Позиции замещения каждого остатка или общее число его заместителей [метилирование]; подтверждённое положение остатка на восстанавливающем или невосстанавливающем конце.
- Абсолютные конфигурации остатков [ГХ модифицированных продуктов гидролиза].
- Общее число групп CH_2 при неполном мономерном составе [ЯМР APT, DEPT-135].
- Известные фрагменты последовательности остатков [анализ продуктов гидролиза].
- Количество остатков фосфорной кислоты [ЯМР ^{31}P].
- Глубина поиска - обзорная или детальная [на основании здравого смысла].

Обзорная глубина поиска исключает из перебора экзотические структуры, накладывая дополнительные ограничения: при интерпретации суперклассов используются только распространённые остатки (>20 вхождений в базе), для каждого мономера используются только характерные для него абсолютные конфигурации и размеры циклов, исключаются сильноразветвлённые топологии, объёмные боковые цепи и другие редкие структурные факторы.

Для получения структурных гипотез используется пятиступенчатый итеративный поиск (Рис. 20). 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. Второй шаг включает получение всех возможных топологий для известного числа остатков. Топология – это направленный граф, отражающий связность остатков без учёта их природы. Правила кодировки и интерактивный список топологий доступны в виде веб-сервиса (http://csdb.glycoscience.ru/biopsel/topology.php). Для каждой топологии, удовлетворяющей структурным ограничениям, вырабатываются операторы транспозиции (Т-операторы). Например, топология разветвлённого трисахарида A-[B-]C не изменится, если поменять местами узлы A и B, поэтому для неё существует Т-оператор, меняющий местами эти узлы («T1» на Рис. 20). 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 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. 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 на основании коэффициента линейной корреляции, среднеквадратичного отклонения и степени достоверности, усреднённой по всем сигналам, предсказан-

ным для данной структуры. Так как экспериментальные спектры могут содержать труднообнаруживаемые сигналы четвертичных углеродных атомов, а также иметь нарушения аддитивности интегральных интенсивностей при совпадении нескольких сигналов, разработанная метрика учитывает степень совпадения размеров спектров, внося «штраф» за пропущенные или лишние сигналы. Это сводит к минимуму влияние неточной оцифровки спектров на предсказательную силу ранжировки.

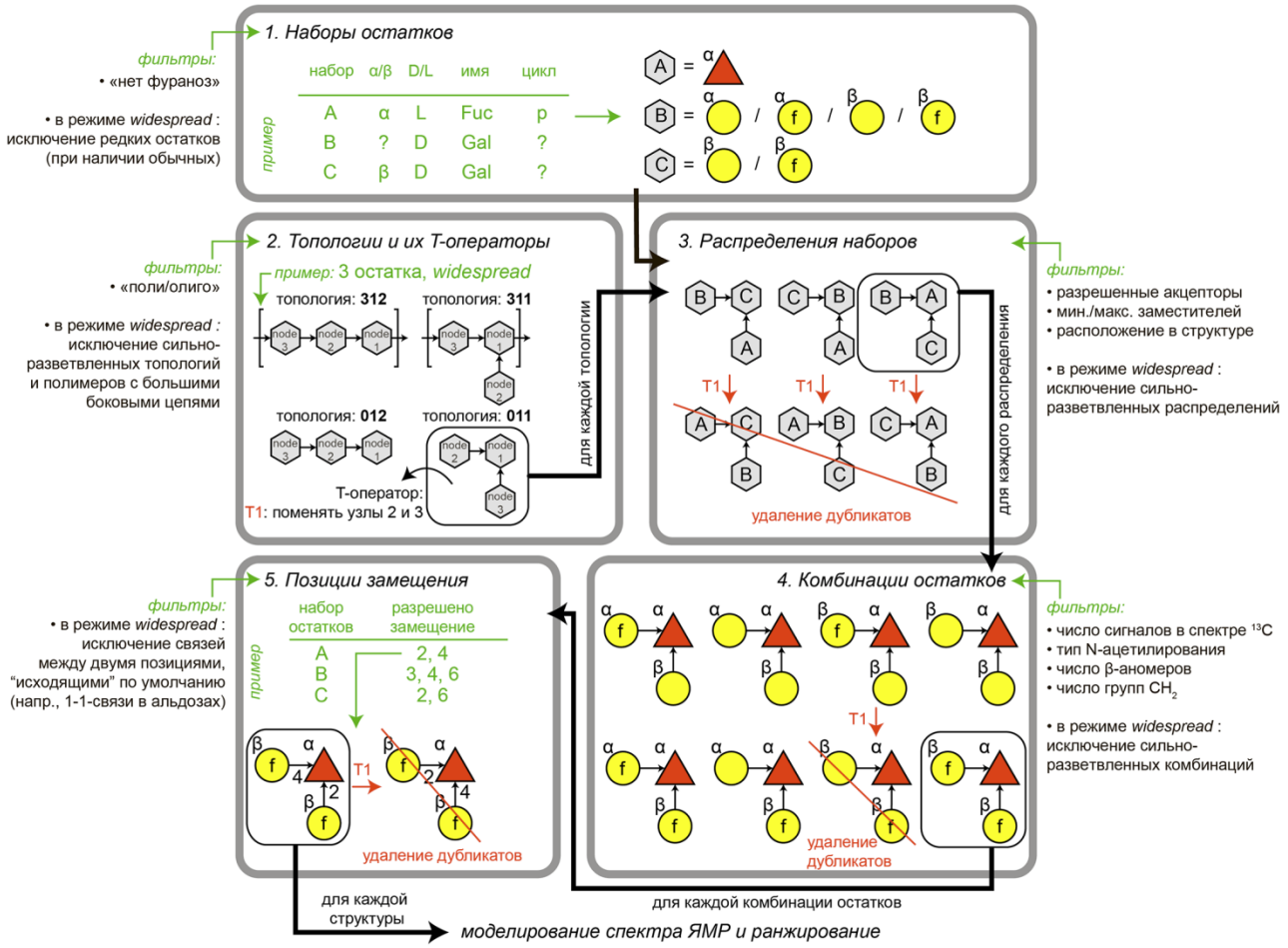


Рис. 20. Алгоритм генерирования разнообразия структур на примере трисахарида. Ограничения приведены зелёным. Остатки показаны в нотации SNFG. Наборы остатков обозначены шестиугольниками. На каждом шаге показана только одна из возможных ветвей алгоритма. Она соответствует объекту, обведённому рамкой на предыдущем шаге.

Для тестирования созданного подхода использовалась случайная выборка из 556 полностью определённых структур биогликанов с опубликованными спектрами ЯМР ^{13}C. В случае присутствия тестируемой структуры в CSDB точность предсказания оказалась бы завышенной по сравнению с предсказанием произвольной структуры из-за нахождения идентичных молекулярных фрагментов при моделировании спектров, поэтому тестируемые структуры временно удалялись из базы. Точность предсказания была протестирована в восьми режимах с разной глубиной поиска и соотношением известных и предсказываемых параметров. Во всех режимах известными считались мономерный состав, типы циклизации, абсолютные конфигурации, тип структурной единицы (полимер или олигомер) и, для олигомеров, остаток на восстанавливающем конце.

Рис. 21А отражает вероятность попадания правильной структуры в рейтинг в зависимости от размера рейтинга в трёх характеристичных режимах. Наиболее распространённый режим 1, подразумевающий установление последовательности остатков и их аномерных конфи-

гураций, 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, 7 моносахаридов и 2 моновалентных остатка, Рис. 22А) в режиме 1 заняла пятое место в рейтинге и переместилась на первое место, когда аномерные конфигурации фукозы и галактозамина были в явном виде указаны как α.

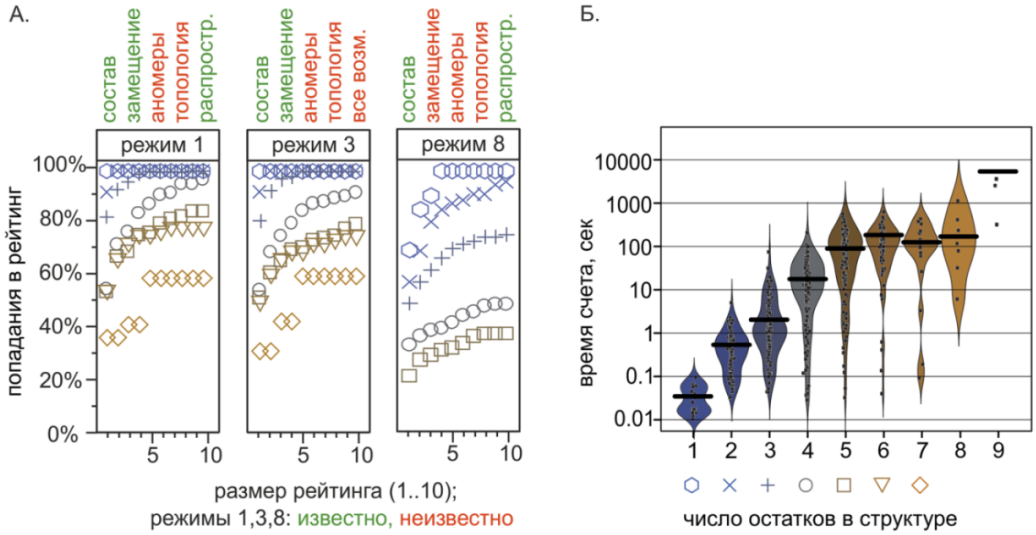


Рис. 21. Валидация GRASS на выборке из 556 структур с известными спектрами. А. Зависимость числа попаданий правильных структур в рейтинг от размера рейтинга. Три режима отличаются количеством ограничений: зелёным перечислены известные параметры, красным – предсказываемые. Топология подразумевает также и последовательность остатков. Форма и цвет точек соответствуют числу остатков в структуре (см. Б, ось абсцисс). Б. Производительность расчёта. Средние значения показаны линиями, отдельные измерения – черными точками. Ширина фигуры отражает распределение времён счёта.

Неуказание общего числа β-аномеров и числа заместителей остатков не оказало существенного влияния. Неуказание положений связей в остатках вызвало среднее уменьшение точности и десятикратное уменьшение производительности. Установление строения больших молекул в слабоограниченных режимах является фундаментальной проблемой, связанной с появлением большого числа структурных гипотез с похожими спектрами. Поэтому для однозначного предсказания строения тетрасахаридов и бóльших структур требуется указание позиций замещения, полученных из эксперимента по метилированию. Использование слабоограниченных режимов оправдано для небольших структур. Так, при отсутствии ограничений правильная структура D-галактана I (Рис. 22Б) была предсказана как наиболее вероятная из ~300000 проверенных; расчёт занял 28 минут.

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 ЯМР-моделированию, не являются препятствиями для работы алгоритма, лишь количественно влияя на достоверность модели. Например, гексасахаридное повторяющееся звено глюкуроноксиломаннана *Cryptococcus neoformans* серотипа А (Рис. 22В) предсказано как наилучшая гипотеза даже при детальной глубине поиска (режим 3). Подобные структуры, содержащие остатки в близком химическом

окружении, традиционно являются сложной задачей для ЯМР-моделирования, особенно при наличии замещения в соседних положениях, стерически влияющего на конформации гликозидных мостиков, а следовательно, и на химические сдвиги (напр., маннозы на Рис. 22В).

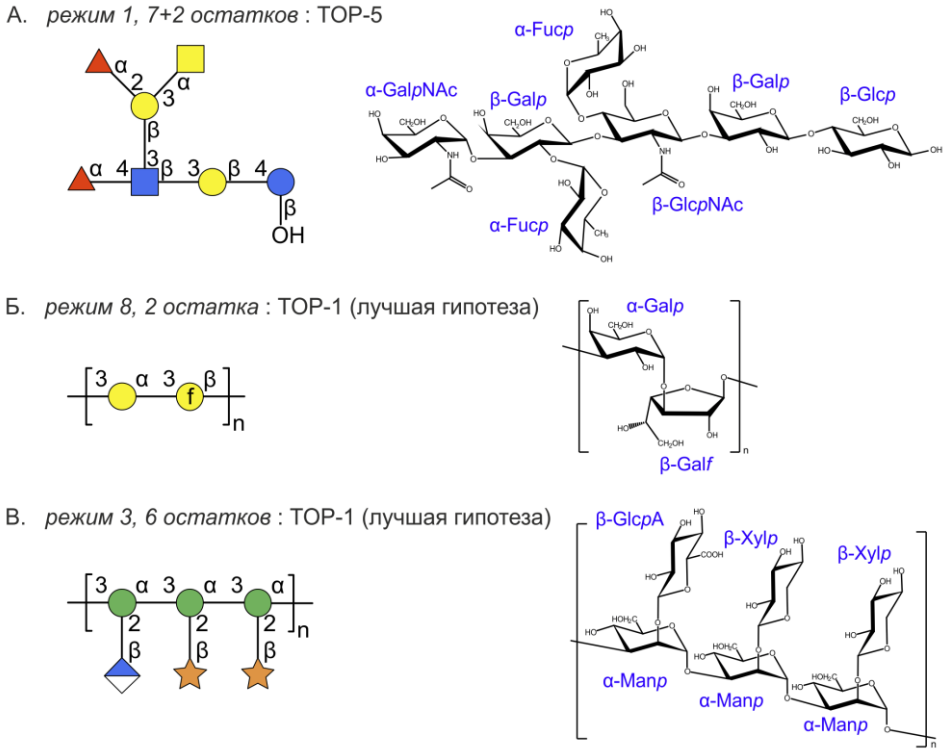


Рис. 22. Результаты предсказания модельных структур по экспериментальным спектрам и данным ГЖХ: А. Антиген группы крови А человека. Б. D-галактан I . В. Ксиломаннан *Cryptococcus neoformans* серотипа А. В примерах А и В известен состав и позиции замещения; предсказана последовательность и аномерные конфигурации; в примере Б известен только состав, предсказано все остальное, включая размеры циклов.

По результатам сравнения способов ЯМР-моделирования углеводов можно заключить, что общехимические подходы к автоматизации связи «структура – спектр» неприменимы даже к простым сахаридам. Из специализированных подходов только CASPER обладал заметной предсказательной силой, но его использование ограничивается набором структур, характерных для биогликанов млекопитающих. Для структур, поддерживаемых обоими методами, GRASS показал превосходство над CASPER, особенно для сложных случаев. В частности, две модельные структуры, представленные на Рис. 22 А и В, не попали в десятку лучших гипотез CASPER, а структура на Рис. 22Б в принципе не поддерживается итератором CASPER.

Созданный инструмент генерирования и оценки качества структурных гипотез в гликохимии востребован при установлении первичной структуры природных 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. Алгоритм и инструмент GRASS подробно документированы (http://csdb.glycoscience.ru/help/nmr.html#grass).

1.4.3 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» создало основу для хемотаксономической классификации организмов, использующей специфичность синтезируемых ими углеводов. Это особенно востребовано для микроорганизмов, чьи

иммунологические свойства часто определяются углеводными антигенами. Для решения этой задачи был создан инструмент статистического анализа содержимого CSDB, позволяющий изучить распределение встречаемости 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 таксономической группы более высокого ранга (например, поиск фрагментов, уникальных для определённого рода в пределах типа, включающего данный род);
- наличие в фрагментах неопределённых конфигураций или размеров циклов;
- наличие в фрагментах агликонов, моновалентных и неуглеводных заместителей;
- считать ли разными фрагменты, отличающиеся только аномерной конфигурацией моносахаридов.

Предлагаемые области применения созданных алгоритмов и программ включают поиск характерных углеводных маркеров в пределах таксонов, в частности, специфических антигенных детерминантов, и исследование активностей гликозилтрансфераз в различных таксономических группах. Созданный инструмент был применён для сравнительного анализа моно- (Рис. 23) и дисахаридных (Рис. 24) строительных блоков в гликанах бактерий, растений и грибов, а также для выявления наиболее распространённых димеров с неуглеводными остатками в растительных гликозидах.

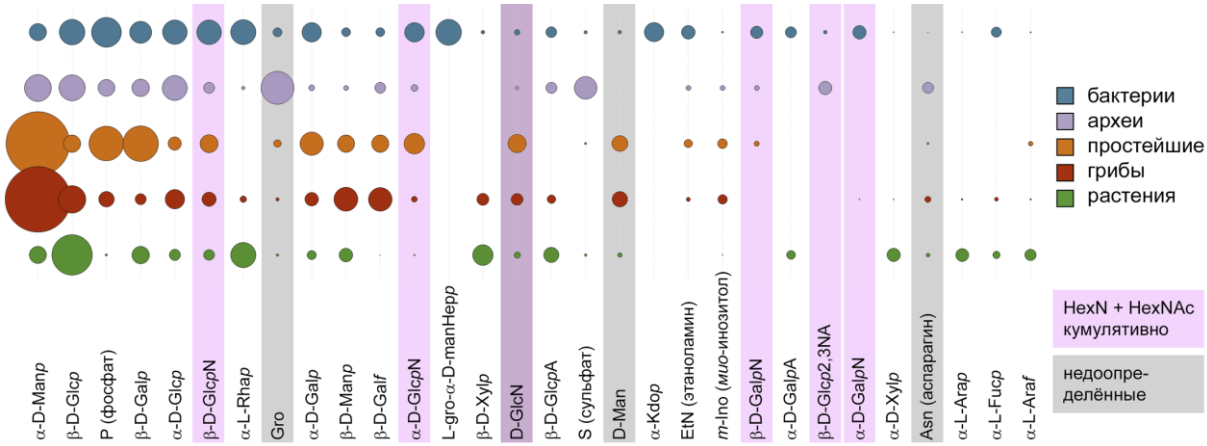


Рис. 23. 30 самых распространённых мономеров в углеводах пяти таксономических групп. Размер кругов соответствует нормализованной частоте встречаемости мономера в домене. Остатки аминосахаров включают ацетилированную и неацетилированную формы и выделены розовым; остатки с неизвестной конфигурацией (аномерная, абсолютная, размер цикла) выделены серым.

Частоты встречаемости фрагментов получены делением числа мономеров и димеров на общее число структур из организмов, относящихся к соответствующему царству (или группе); распространённые неспецифичные моновалентные остатки (метанол, уксусная кислота и т.д.) не принимались во внимание. Возможные настройки статистического анализа подробно документированы (http://csdb.glycoscience.ru/help/stat.html). На Рис. 23 приведены наиболее распространённые мономерные остатки, характерные для гликанов организмов из пяти доменов, представленных 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 конкуренцией и иммунной системой организма-хозяина. К наиболее распространённым мономерам бактерий, исключая компоненты липида А, относятся L-глицеро-D-манногептоза и 3-дезокси-D-маннооkт-2-улозоновая кислота, которые присутствуют в липополисахаридах грамотрицательных бактерий и которых нет в углеводах организмов из других доменов.

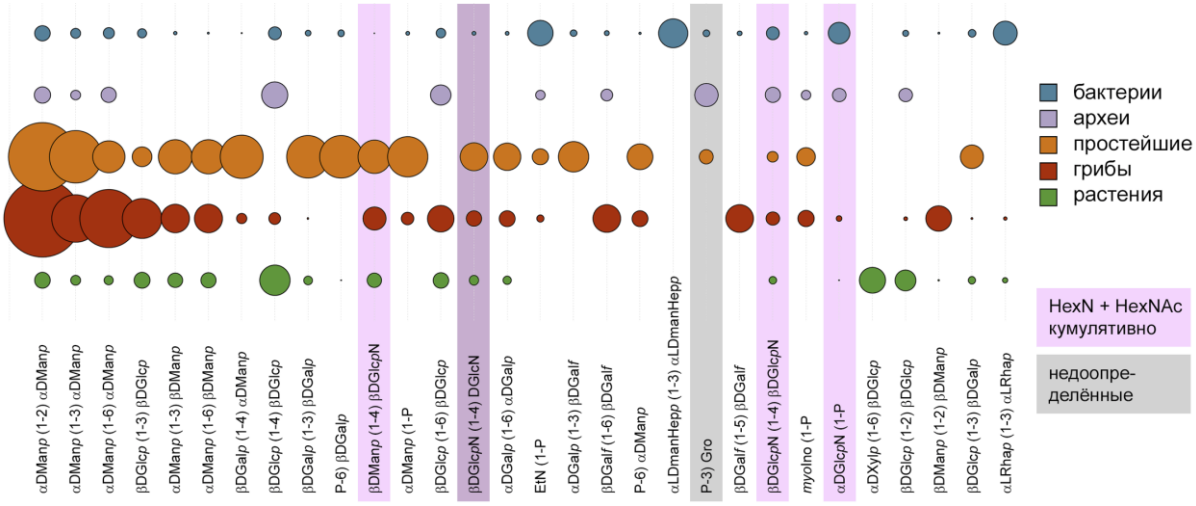


Рис. 24. 30 самых распространённых димеров в углеводах пяти таксономических групп. Размер кругов соответствует нормализованной частоте встречаемости димера в домене. Димеры, содержащие аминосахара, включающие ацетилированную и неацетилированную формы, выделены розовым; димеры, содержащие остатки с неизвестной конфигурацией, выделены серым. Полиолсодержащие димеры чаще всего являются аналитическими артефактами расщепления по Смиту и не включены в рассмотрение.

Димерные фрагменты, уникальные для таксогруппы, могут отражать особенности взаимодействия входящих в неё организмов с окружающей средой, реализуемого за счёт активности специфических гликозилтрансфераз; растительные гликозиды, содержащие агликоны тритерпеновой, стероидной, флавоноидной и фенольной природы, обладают биологической активностью и представляют интерес с медицинской точки зрения. Растительные и прокариотические углеводы демонстрируют большее разнообразие гликозидных связей по сравнению с другими доменами. Доменоспецифичные ферменты синтезируют фрагменты структуры, уникальность которых в растениях связана с положением связи (например, α-L-рамнопиранозил-2-β-D-глюкопираноза, Рис. 24), а не с мономерными строительными блоками (как, например, дигептозные фрагменты в структурах бактериального кора). Анализ распределения димеров по классам патогенов с последующим сравнением с набором 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 выявления характерных признаков и оценки выборки необходимых строительных блоков (защищённых моносахаридов) было проведено сравнение разнообразия гликомов бактерий с разнообразием гликомов млекопитающих по следующим параметрам:

- 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 моно- и димерных фрагментов, уникальных для изучаемой группы организмов,
- распределение модификаций моносахаридов (фосфорилирование, О-ацетилирование и др.),
- 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).

Одним из способов получения биогликанов является ферментативный синтез в бактериях с модифицированным геномом. Был проведён анализ доступности субстратов гликозилтрансфераз в бактериях различной таксономии. Эта информация востребована для выбора микроорганизмов, способных синтезировать углеводы, применяемые в биотехнологии и медицине. Кроме того, карты встречаемости углеводных димеров выявляют структурные признаки, отсутствующие в гликомах тех или иных организмов, что позволяет отбирать микроорганизмы для экспрессии рекомбинантных гликозилтрансфераз с заданной активностью для целенаправленной наработки продукта.

1.4.4 Углеводная фенетика. Инструмент таксономической кластеризации предназначен для выявления групп таксонов, объединённых по признаку схожести синтезируемых ими углеводов. Он впервые позволил гликобиологам группировать произвольные таксономические ранги на основании схожести гликозилтрансфераз. Используемый подход рассчитывает встречаемость моно- и димерных фрагментов в углеводных структурах, присутствующих в заданных таксогруппах и удовлетворяющих фильтрам отбора. На основании сравнения полученных паттернов встречаемости методом Хамминга генерируются матрицы схожести для наборов структур, принадлежащих к конкретным таксонам. Далее таксоны нормализуются по степени изученности и кластеризуются в родственные группы по характерным признакам (блок-схема на Рис. 25). Были протестированы шесть существующих алгоритмов кластеризации, применимых к задачам биологической группировки, основанной на родстве биомолекул. Они доступны в рамках веб-инструмента, интегрированного в CSDB. Результаты кластеризации представляются в виде фенетических деревьев и могут быть экспортированы в популярные филогенетические форматы.

Данный инструмент может быть использован для выявления связей между углеводными структурами и ферментативными активностями, вовлеченными в их синтез и процессинг: организмы, синтезирующие сходные дисахаридные фрагменты, должны обладать ферментами со сходными активностями. Подобный подход ускоряет исследования углевод-активных ферментов, экспериментальное подтверждение функций которых сталкивается со значительными трудностями. В первую очередь это относится к ферментам бактерий, которые представляют интерес с биотехнологической точки зрения и углеводные структуры которых наиболее полно представлены в CSDB. Для иллюстрации качества кластеризации таксонов приведена дендрограмма (Рис. 26), построенная на основании анализа встречаемости димерных фрагментов в гликанах организмов, принадлежащих к родам, наиболее представленным в базе CSDB.

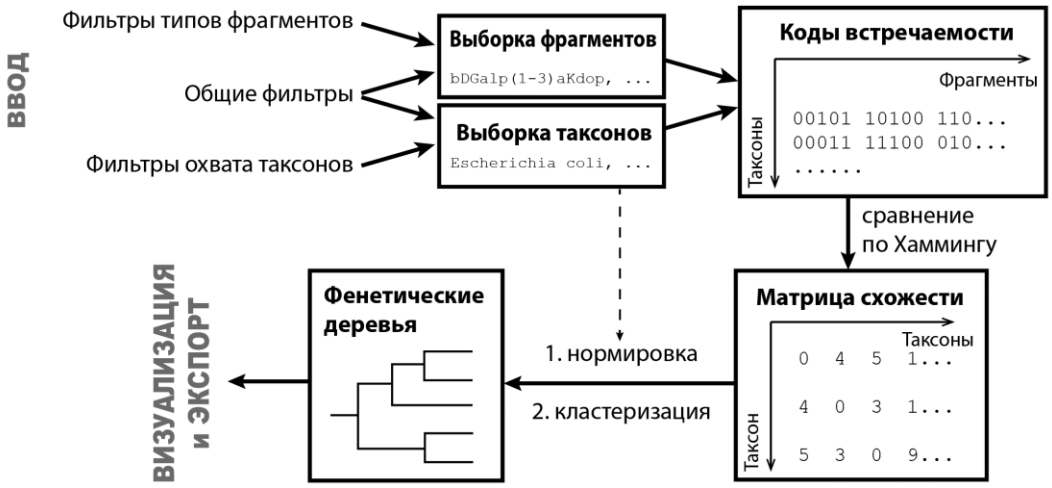


Рис. 25. Схема работы инструмента таксономической кластеризации.

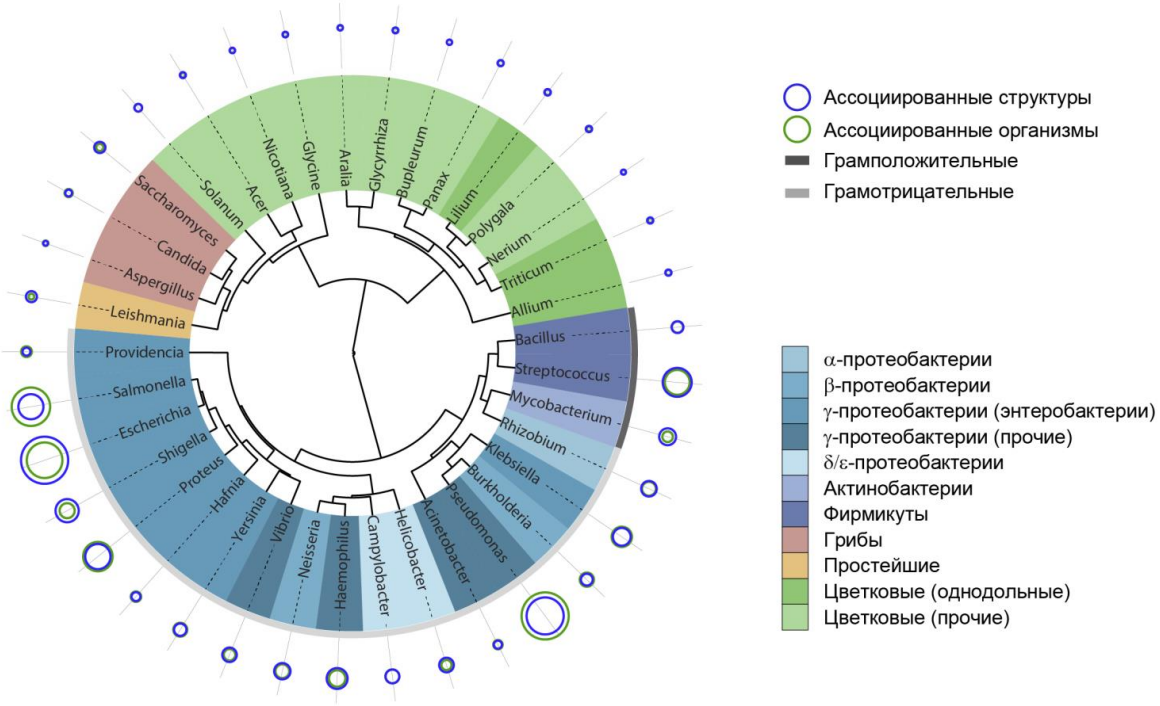


Рис. 26. Фенетическое дерево, построенное по результатам кластеризации наиболее изученных родов. Оттенками синего показаны таксономические группы бактерий, оттенками зелёного – растений, красным – грибов, оранжевым – простейших. Цвет внешней дуги отражает реакцию по Граму. Размер кругов соответствует нормализованной представленности данного рода в CSDB с точки зрения организмов (зелёный) и структур (синий). В случаях, когда зелёный круг не виден, его размер совпадает с синим.

На Рис. 27А представлены 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 (44-53% в зависимости от алгоритма кластеризации) совпадают с классическим «деревом жизни», построенным на основании последовательностей консервативных рибосомальных РНК (Рис. 27Б). Различия в генах углевод-активных ферментов отражают 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. Таким образом, различия между бактериальными фенетическими деревьями, построенными на основании последовательностей сахаридов и рРНК, отражают тот факт, что углеводы являются основным инструментом взаимодействия

бактерий с окружающей средой, и продуцирование определенных структур соответствует определённой среде обитания. Так, *Neisseria gonorrheae* и *Haemophilus ducreyi*, расположенные далеко друг от друга на «дереве жизни» рРНК, но обладающие сходными гликанами, вызывают заболевания, передающиеся половым путем, обитают в мочеполовых путях человека и сталкиваются с близкими факторами среды. Близость видов на углеводном фенетическом дереве отражает сходство их жизнедеятельности и позволяет планировать дальнейшие экспериментальные исследования в условиях слабой изученности механизмов бактериального патогенеза.

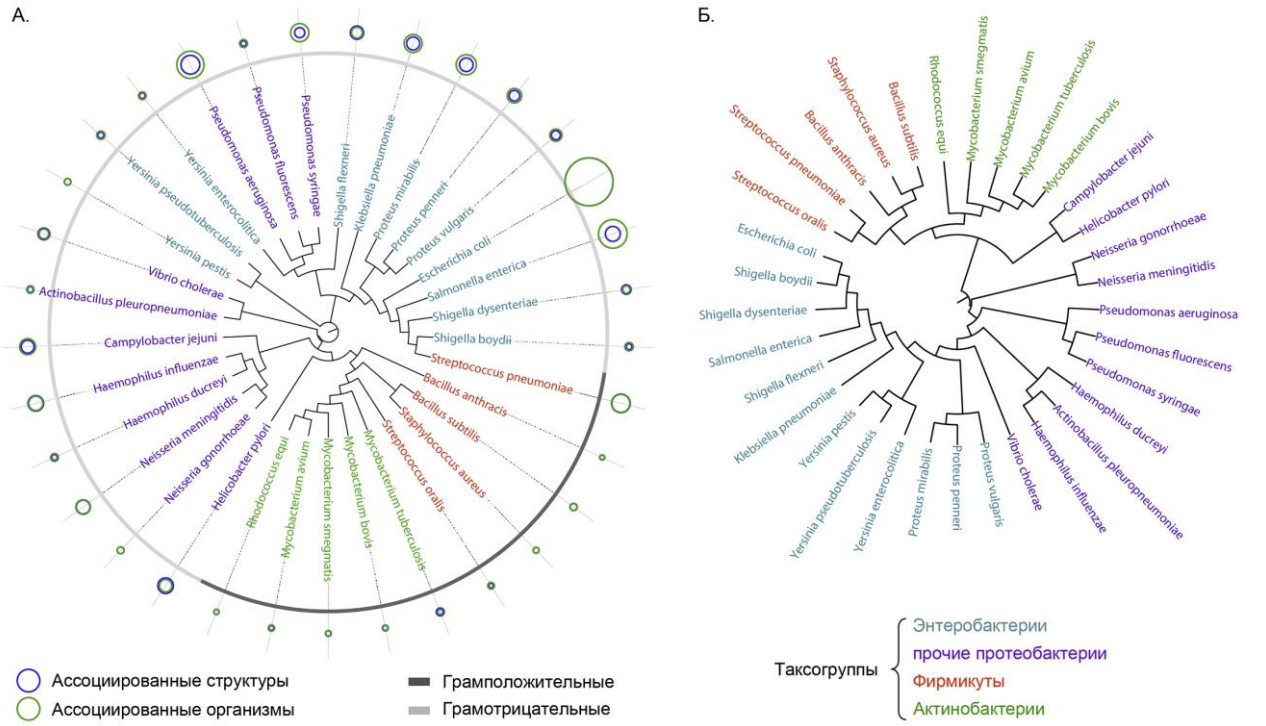


Рис. 27. Фенетические деревья наиболее изученных бактериальных видов. Цвет отражает таксогруппу бактерий. Использован алгоритм кластеризации BioNJ. А. Кладограмма, основанная на распределении дисахаридов, присутствующих в биогликанах. Б. Кладограмма, основанная на последовательностях рибосомальной РНК.

1.5 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, и существующие проекты были изолированы друг от друга из-за недостаточной стандартизации форматов и протоколов. Первая попытка связать углеводные проекты, обеспечив использование двух баз данных в инструментах поиска и прогнозирования, реализованных в каждой из них, была предпринята автором более 10 лет назад. В рамках этого 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.

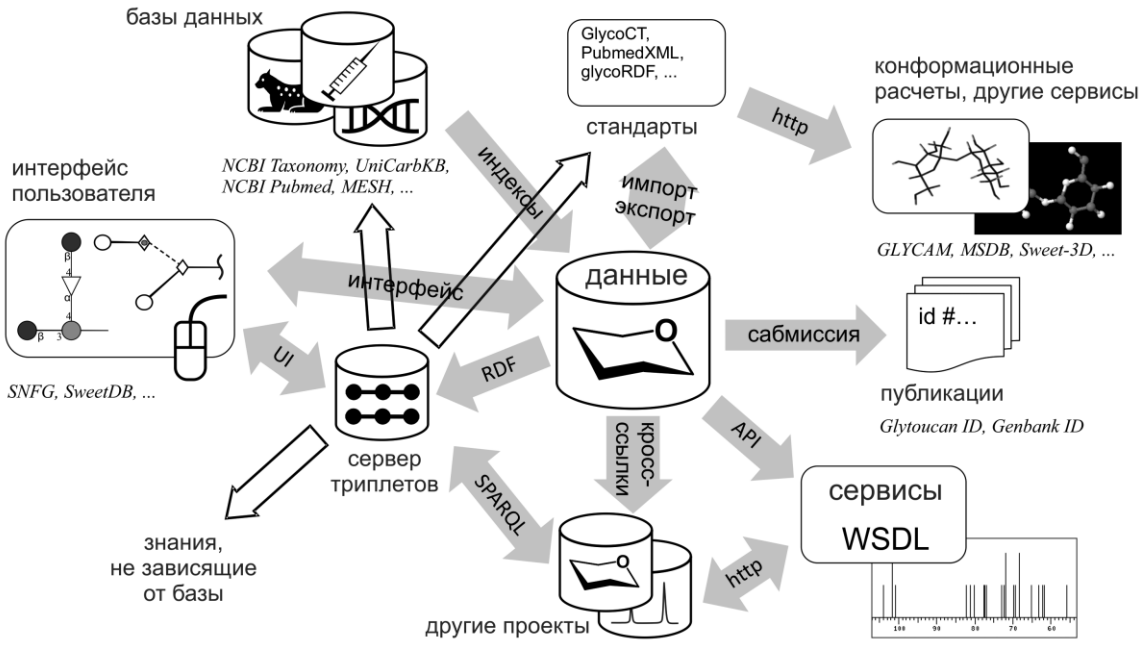


Рис. 28. Интеграция CSDB с другими проектами (показаны курсивом). Серые стрелки - протоколы обмена данными. UI = интерфейс пользователя; API = автоматический программный интерфейс.

В той мере, в которой эти наработки реализованы в CSDB (Рис. 28), они 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 коллаборатorами в 2015 году. Ссылки на идентификаторы GlyTouCan автоматически генерируются при любом выводе структур в CSDB, когда это возможно (приблизительно для 50% структур). Вопрос о включении этих идентификаторов в публикации в обязательном порядке (по аналогии с идентификаторами GenBank в геномике) требует поддержки со стороны научных издательств и лоббируется Консорциумом по гликоинформатике в настоящее время.
3. Возможность перевода информации о структуре между основными углеводными языками. CSDB позволяет импортировать структуры в нотации GlycoCT и экспортировать структурную и сопутствующую информацию в нотациях GlycoCT, Glyde-II, LinUCS, SNFG, SweetDB (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 и Glytoucan (структуры), MSDB (моносахариды), NCBI Pubmed и DOI (публикации), NCBI Taxonomy (организмы). Отдельные типы данных стандартизированы для аналогичной интеграция с международным каталогами ICD-11 (заболевания, ткани, органы) и MeSH (термины и методы анализа).
5. Стандартизированные, сертифицированные и признанные научным сообществом средства ввода и редактирования углеводных структур и обеспечение их совместимости с логикой каждого из проектов. В CSDB для этой цели адаптированы Java-апплет GlycanBuilder (разработка объединённого коллектива нескольких европейских институтов), браузерные при-

ложения SugarSketcher (разработка коллабораторов из Швейцарского института биоинформатики) и Structure Wizard (собственная разработка), а также библиотека распространённых либо биологически значимых структур.

6. Наличие точек входа (API), работающих по стандартным протоколам информатики и их углеводным расширениям в каждом из проектов, и их формальное описание на языке WSDL для того, чтобы другие проекты могли посылать запросы и получать ответы в автоматическом режиме, без необходимости подстраиваться под формат данных CSDB.

Интеграция проектов гликохимии и гликобиологии в единое информационное поле курируется международным Консорциумом по гликоинформатике, членом которого является автор.

Новейшим направлением обработки данных в естественных науках и получения неявно заданных знаний является так называемая семантическая паутина, представляющая данные в модели Resource Description Framework (RDF) в виде триплетов *объект-предикат-субъект*. Она позволяет интегрировать знания из разных проектов автоматически и допускает распределённые запросы к базам данных с минимальным знанием форматов и интерфейсов каждой из них. Адаптация модели RDF к углеводам была начата в сотрудничестве с другими группами в 2014 году. Мощь этого подхода можно продемонстрировать следующим модельным примером. Предположим, необходимо найти белок-носитель для произвольного гликана из CSDB. Это нельзя сделать напрямую, так как структуры в CSDB не связаны ссылками с белковыми базами. Но записи в CSDB имеют ссылки на идентификаторы в GlyTouCan. Как GlyTouCan, так и еще одна база, UniCarbKB, могут экспортировать структуры в формате GlycoCT. Наконец, записи в UniCarbKB имеют ссылки на белковую базу UniProt. Эти факты позволяют предположить, что, сопоставив идентификаторы CSDB и UniCarbKB и используя GlyTouCan, можно получить идентификаторы UniProt из UniCarbKB для каждого идентификатора CSDB. Неуниверсальное решение, реализованное вручную для конкретного объекта исследования, тривиально, но в общем виде эта задача решается только с помощью RDF. 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. Глобальная онтология знаний об углеводах GlycoRDF была разработана в 2015 году в сотрудничестве с американской и японской группами Консорциума по гликоинформатике. Она включает 246 объектов в 130 классах и стандартизирует 367 взаимосвязей между ними, формализована на языке OWL и снабжена примерами использования. Следуя договорённостям, достигнутым в 2013 г., ведущие мировые проекты гликоинформатики (CSDB, Glycosciences.de, MSDB, GlycomeDB – ныне GlyTouCan, UniCarbKB, GlycoEpitope, GlycoNAVI, GlycoProtDB и др.) перевели собственные форматы и базы данных в модель RDF и предоставили данные для размещения в репозитории триплетов. Репозиторий, поддерживаемый Университетом Сока (Япония) и доступный через Интернет любому проекту для запросов на языке SPARQL, обновляется на основании RDF-данных участников раз в два года.

1.6 Техническая реализация проекта

База данных CSDB построена по реляционному принципу. Диаграмма отношений между данными, пришедшими из научных публикаций, их индексами и таблицами базы доступна в

справочной системе (http://csdb.glycoscience.ru/help/csdb_entities.pdf). Ядро базы представлено следующими группами таблиц: свойства биогликанов на уровне молекулярной структуры, свойства моносахаридов и полная первичная структура, библиографическая 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 и не входят в ядро. Модуль работы с гликозилтрансферазами реализован в виде отдельной базы данных, интегрированной с CSDB на уровне данных о молекулярной структуре.

Все данные, вносимые аннотаторами, хранятся и редактируются в текстовых дампах, которые одновременно служат резервными копиями. Формат дампа и архитектура базы подробно документированы (http://csdb.glycoscience.ru/help/dbdocs.html). Управляющие программы платформы CSDB написаны на языках PHP, SQL, R и Python. Новые функции появляются несколько раз в год, все обновления проходят тщательное тестирование. Взаимодействие с пользователем реализовано с помощью веб-интерфейса. Интерфейс поддерживает отрисовку структурных формул, двумерных спектров, интерактивную визуализацию трёхмерных молекулярных моделей и имеет графический редактор углеводных структур в формате SNFG.

Молекулярно-динамические расчёты моносахаридов проводились с использованием пакета TINKER, силового поля MM3-2001 и 1-наносекундных траекторий при 1000 K. Для оптимизации структур в потоковом режиме использовалась молекулярная механика в силовом поле MMFF94.

Для кластеризации таксонов, сравнения фенетических деревьев и построения дендрограмм использовались язык R, библиотека Ape («анализ филогенетики и эволюции»), веб-сервис Compare2Trees, экспорт CSDB в формате Nexus в пакет iTOL и известные алгоритмы кластеризации: UPGMA, Ward.D2, BIONJ, fastME и «полная связность».

Проект работает на выделенном физическом Windows-сервере. Предусмотрена балансировка нагрузки при ресурсоёмких расчётах. 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 документация бесплатно доступны пользователям сети Интернет по адресу http://csdb.glycoscience.ru.

2. Использование разработок в гликохимии и гликобиологии

Научная значимость разработанной платформы, включающей собственно базу данных углеводных структур и её надстройки, подтверждается её применением в различных исследованиях.

2.1 Примеры решения модельных задач

Типовые задачи, решаемые с помощью CSDB, и примеры её использования в гликохимических исследованиях подробно разобраны в специальном веб-учебнике (http://csdb.glycoscience.ru/help/examples.html) и в монографии. Эта документация содержит иллюстрированное пошаговое руководство пользователя CSDB в 12 примерах, моделирующих задачи, с которыми исследователь углеводов сталкивается в своей ежедневной научной практике. Задачи специально сформулированы в качестве обучающих моделей. Все поисковые примеры объясняют сложные составные запросы, так как простые одноступенчатые запросы интуитивно понятны и не требуют специальных разъяснений. Данный раздел справочной системы подразу-

мевает 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, опубликованные после 2005 года в связи с антигенной активностью.
3. Найти гликаны, полученные из растений рода *Паслён* и содержащие соланидиновый компонент.
4. Найти углеводы, кроме октозосодержащих, имеющие в спектре ЯМР ^{13}C сигнал вблизи 34 м.д.
5. Найти публикации *Книреля* или *Шашкова* по бактериальным гликанам, включающим 4-хиновозамин, амидированный любой N-ацетилированной аминокислотой.
6. Найти структуры, построенные из любых ноноз одного типа (моносахариды или гомополимеры).
7. Промоделировать спектры ЯМР 3-O-абеквозил-6-дезокси-β-D-манногептопиранозил-(D-рибит-1)-фосфата в воде 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 и хроматографии.
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.

2.2 Использование знаний, полученных из CSDB, в других исследованиях

Практически полное покрытие по бактериальным углеводам и многочисленные инструменты позволяют использовать CSDB для решения различных химических и аналитических задач. Среди наиболее востребованных применений следует отметить поиск характерных мотивов и эпитопов в гликополимерах патогенных бактерий, проверку новизны структуры, моделирование и отнесение спектров ЯМР, исследование влияния структурных параметров на спектральные.

CSDB востребована также в биохимических, молекулярно-биологических и медицинских исследованиях. С её помощью выявляют характерные элементы в углеводах микроорганизмов с целью изучения иммунного ответа организма-хозяина и использования таких элементов в составе вакцин. Статистический анализ характеристик и разнообразия углеводов, представленных в CSDB, привлекается в качестве обоснования исследований углеводных стимулов, посредством которых бактерии взаимодействуют с иммунной системой. Например, в работе проф. Ю.А. Книреля и коллег с помощью CSDB был проведён поиск эпитопов системы групп крови АВН в бактериальных углеводах с целью создания массива структур для скрининга вза-

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

Репозиторий углеводных структур также востребован 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 О.Г. Овчинниковой и коллег, посвящённой новому семейству гликозилтрансфераз, переносящих остатки β-Kdo, CSDB была использована для поиска структур, содержащих эти остатки, что позволило связать структурные данные с последовательностями генов и предсказать активность β-Kdo-гликозилтрансферазы, которая впоследствии была охарактеризована биохимически. В работе по фосфорилазе 3-O-α-D-глюкопиранозил-L-рамнозы из *Clostridium phytofermentans* Нихира и коллеги использовали CSDB для выявления данного субстрата в гликанах микроорганизмов.

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-сайт составляет около 300 уникальных посетителей ежемесячно, исключая роботов и пользователей, посетивших только одну страницу.

2.3 Выявление ошибок в базах и публикациях

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 программа поиска отклонений химических сдвигов в спектрах ЯМР от ожидаемых на основании ЯМР-моделирования. Было обнаружено около 600 несовпадений, превышающих 6 м.д. для сигналов ^{13}C или 1 м.д. для сигналов ^{1}H. Анализ этих случаев с привлечением оригинальных публикаций и известных зависимостей спектров 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, но отмечались как подозрительные.
3. 12%. Ошибки в структурах, взятых из предыдущих статей, в которых структура установлена неверно при том, что существуют более поздние статьи, в которых структура уточнена на основании новых данных или более тщательного анализа, в том числе с помощью CSDB. Ошибки исправлены с указанием ссылок на уточняющие публикации, пути миграции ошибок отслежены и описаны.
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; 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, неправильный перенос данных из спектра в таблицу отнесения сигналов или из оригинальной публикации в последующую). Ошибки исправлены.

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, 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, ради сохранения возможности поиска в базу попадали как опубликованные данные со специальной пометкой, так и правильные. Кроме ошибок в статьях с помощью повторного аннотирования было исправлено несколько тысяч ошибок в записях, импортированных из CarbBank, что ограничило миграцию этих ошибок в современные проекты, использующие данные CarbBank (включая CSDB).

Возможность выявлять противоречия между спектрами и структурой с помощью CSDB используется и другими коллективами. Например, сотрудники группы чл-корр. Н.Э. Нифантьева нашли ошибки в отнесении галактофурананов, из-за которых в публикациях других авторов были неправильно идентифицированы критические для структуры сигналы C1 и C6 β-D-Gal*f*, 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 (установлена Э. Катцелленбоген и коллегами) с химическими сдвигами C1 и C5 4-дезоксиарабиногексозы позволила М. Вангу и коллегам 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, была позже исправлена авторами и другими группами, но при этом были внесены другие ошибки. Окончательный правильный вариант получен на основании рассуждений сотрудников группы проф. Ю.А. Книреля и анализа химических сдвигов коллективом CSDB. Подобные примеры не учтены в приведённой статистике, так как к моменту начала систематического скрининга несоответствий они уже были исправлены в базе CSDB по результатам аннотирования поздних статей, включая собственные. Исправление ошибок в отнесении опубликованных спектров и исключение из рассмотрения заведомо ошибочных данных позволили увеличить точность статистического предсказания химических сдвигов модулем GODDESS.

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, созданы способы объединения эмпирических и статистических ЯМР-моделей и оценки их достоверности. 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 при-

знана в качестве рекомендованного стандарта ведущими углеводными научными журналами.

7. Создана уникальная база данных по подтверждённым активностям гликозилтрансфераз (~1200 установленных и ~600 предсказанных функций ферментов), 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*).
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.

Основные публикации по теме работы

Главы в книгах:

1) 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.
2) 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.
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.

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 P. V., Knirel Y. A. New database of bacterial carbohydrate structures // Glycoconjugate Journal. – 2005. – Т. 22. – С. 216-217.
6) 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.

7) 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. – T. 8. – C. ID 35.

8) Toukach P. V. Bacterial carbohydrate structure database 3: principles and realization // Journal of Chemical Information and Modeling. – 2011. – T. 51, № 1. – C. 159-170.

9) Egorova K. S., Toukach P. V. Critical analysis of CCSD data quality // Journal of Chemical Information and Modeling. – 2012. – T. 52, № 11. – C. 2812–2814.

10) 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.

11) 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. – T. 4, № 1. – C. ID 39.

12) Toukach P. V. CSDB and other carbohydrate databases // Glycoconjugate Journal. – 2013. – T. 30. – C. 347-349.

13) 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. – T. 23, № 12. – C. 1422-1424.

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 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. – T. 5, № 1. – C. ID 5.

15) 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. – T. 54, № 9. – C. 2594-2611.

16) Egorova K. S., Toukach P. V. Expansion of coverage of Carbohydrate Structure Database (CSDB) // Carbohydrate Research. – 2014. – T. 389. – C. 112-114.

17) 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. – T. 31, № 6. – C. 919–925.

18) 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. – T. 25, № 12. – C. 1323-1324.

19) 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.
20) 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.
21) 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.
22) 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.
23) 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.
24) Egorova K. S., Toukach P. V. CSDB_GT: a new curated database on glycosyltransferases // Glycobiology. – 2017. – Т. 27, № 4. – С. 285-290.
25) Kapaev R. R., Toukach P. V. GRASS: semi-automated NMR-based structure elucidation of saccharides // Bioinformatics. – 2018. – Т. 34, № 6. – С. 957-963.
26) 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.
27) 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.
28) 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.
29) Egorova K. S., Knirel Y. A., Toukach P. V. Expanding CSDB_GT glycosyltransferase database with Escherichia coli // Glycobiology. – 2019.10.1093/glycob/cwz006. – ePub ahead of print.

Тезисы докладов на конференциях и публикации, не участвующие в защите

По теме работы представлен 51 доклад (для 41 из них опубликованы тезисы), преимущественно на международных конференциях (International Carbohydrate Symposium, European Carbohydrate Symposium, International Glycoconjugate Symposium, Baltic Meeting on Microbial Carbohydrates и другие аналогичные мероприятия). По результатам исследований строения природных углеводов в 2007-2019 гг. с использованием наработок проекта CSDB опубликовано также 16 статей в соавторстве с другими коллективами. Так как основная тематика этих статей связана с конкретными объектами гликобиологии, а не с гликоинформатикой в целом, они не представлены в качестве публикаций по защищаемой работе. Полный список докладов и публикаций доступен на сайте автора (http://toukach.ru/rus/publist.htm).

Автор выражает благодарность проф. Ю.А. Книрелю, проф. Т. Люттеке, проф. К. Аоки-Киношита, проф. К.-В. фон дер Лиету, проф. А.С. Шашкову, к.б.н. К.С. Егоровой, д-ру Р. Рандцингеру, д-ру С. Хергету, д-ру М. Кэмпбеллу, к.х.н. Н.А. Калинчук, Р.Р. Капаеву, И.Ю. Чернышову за плодотворное сотрудничество.

Список сокращений

Сокращения, появившиеся в рамках данной работы, показаны **жирным** шрифтом.

API	Application Programming Interface (автоматический программный интерфейс)
BIOPSEL	BIOPolymer Structure ELucidation (установление структуры биополимеров)
CFG	Consortium for Functional Glycomics (Консорциум по функциональной гликомике)
CSDB	**Carbohydrate Structure DataBase** (углеводная структурная база данных) – *название проекта, базы данных и платформы*
COSY	COrrelation SpectroscopY (корреляционная спектроскопия)
DEPT	Distorsionless Enhancement by Polarization Transfer (улучшение без искажений за счёт переноса поляризации)
DOI	Digital Object Identifier (цифровой идентификатор объекта)
GlycoCT	Glyco Connection Table (углеводная таблица связности)
GODDESS	**Glyco-Optimized Database-Driven Empirical Spectrum Simulation** (оптимизированное для углеводов эмпирическое предсказание спектров на основании базы данных) – *название алгоритма*
GRASS	**Generation, Ranking and Assignment of Saccharide Structures** (генерирование, ранжирование и отнесение сахаридных структур) – *название алгоритма*
HOSE	Hierarchical Organization of Spherical Environment (иерархическая организация сферического окружения)
HSQC	Heteronuclear Single-Quantum Coherence (гетероядерная одноквантовая когерентность)
IUPAC	International Union for Pure and Applied Chemistry (Международный союз по фундаментальной и прикладной химии)
NCBI	National Center for Biotechnology Information (Национальный центр биотехнологической информации)
MMFF94	Merck Molecular Force Field 1994 (молекулярное силовое поле Мерка, 1994)
MSDB	MonoSaccharide DataBase (база данных моносахаридов)
NOE	Nuclear Overhauser Effect (ядерный эффект Оверхаузера)
REStLESS	**REsidues as SMILES and LinkagEs as SMARTS** (остатки в SMILES + связи в SMARTS) - *название модуля перевода семантического описания в поатомное*
RDF	Resource Description Framework (модель описания ресурсов)
SMARTS	SMILES Arbitrary Target Specification (произвольное соединение кодов SMILES)
SMILES	Simplified Molecular-Input Line-Entry System (упрощённая линейная система для ввода молекулярной структуры)
SNFG	**Symbolic Notation For Glycans** (символическая углеводная нотация) – *название углеводной нотации*
WSDL	Web Service Description Language (язык описания веб-сервисов)
WURCS	Web3 Unique Representation of Carbohydrate Structures (уникальное представление углеводных структур для Web 3.0)
ЯМР	Ядерный магнитный резонанс
ГЖХ	Газо-жидкостная хроматография