***Ф.В. Тоукач,**
****И.С. Бушмаринов,**
***Ю.А. Книрель**
Институт Органической Химии им. Н.Д. Зелинского РАН
****Высший Химический Колледж РАН***

НОВАЯ ЭЛЕКТРОННАЯ БАЗА ДАННЫХ ПО БАКТЕРИАЛЬНЫМ УГЛЕВОДАМ (BCSDB)

УДК 547.29

Создана открытая структурная база данных по бактериальным углеводам. В работе освещается структура базы данных, ее интерфейс и специально разработанный язык линейной кодировки структур углеводов.

1. ВВЕДЕНИЕ

Углеводы играют важную роль в живой природе, создавая универсальную систему сохранения клеток, регулирования жизненных процессов и информационного межклеточного взаимодействия. Они участвуют в патологии инфекционных и раковых заболеваний и используются для их профилактики, диагностики и терапии. В отличие от геномики и протеомики, гликомика – наука, занимающаяся идентификацией углеводов и анализом их строения и свойств – не располагает универсальной базой данных, необходимой для ориентации в постоянно расширяющемся информационном пространстве данных о природных углеводах.

Целью гликомики является создание каталога всех известных полисахаридов и выявление влияния их структур на здоровье человека и развитие болезней [1,2,3]. Естественно, для решения подобной задачи необходимо накопление статистической информации по структурам и свойствам углеводов.

Установление точной структуры углеводов долгое время являлось препятствием для сбора подобных данных, но развитие современных аналитических методов, таких как масс-спектроскопия и спектроскопия ядерного магнитного резонанса, позволило эффективно и надежно определять строение большинства природных углеводов [4, 5].

Сегодня основной задачей гликомики является обработка уже накопленных сведений об углеводах. Однако следует заметить, что на настоящий момент опубликовано более 30000 структур природных углеводов, и ориентироваться в этом объеме информации практически невозможно. Следовательно, необходимо создание открытых, доступных любому исследователю обновляемых баз данных по различным разделам гликомики.

Наиболее крупным проектом такого рода был CarbBank [6], объединивший данные об углеводных структурах, опубликованных до 1995 года. Однако за прошедшие 11 лет объем накопленной структурной информации об углеводах как минимум удвоился, и данных CarbBank стало недостаточно.

На настоящий момент существует несколько баз данных по структурам углеводов, объединяющих данные из различных источников, в том числе из CarbBank (*GLYCOSCIENCES* [7], GlycoSuite[8], CFG Glycan Database [1]), однако они ориентированы в основном на информацию об углеводах млекопитающих и/или о синтетических углеводах. В то же время, антигены многих бактерий имеют углеводную природу, и распознавание бактерий иммунной системой человека определяется структурой этих соединений. Иммунологическую роль углеводов бактерий трудно переоценить. Таким образом, отсутствие единой базы данных по бактериальным углеводам является серьезным пробелом в информационном пространстве гликомики. В то же время, коллектив лаборатории химии углеводов ИОХ РАН много лет занимается установлением строения и синтезом бактериальных углеводов и в лаборатории был накоплен уникальный библиографический материал именно в этой области.

Целью данной работы было создание единой открытой базы мирового уровня по бактериальным углеводам и их производным (Базы Данных по Структурам Бактериальных Углеводов, далее БДСБУ). Так как языком базы данных является английский, в данном тексте будет встречаться также обозначение BCSDB (Bacterial Carbohydrate Structure DataBase).

1. СТРУКТУРА БДСБУ

Каждая запись в базе данных содержит структурную, ЯМР-спектроскопическую, биологическую, серологическую и библиографическую информацию о конкретном углеводе. Данные в записи разбиты на поля, подробное описание которых приведено в таблице 1.

Таблица 1. Описание содержимого полей в БДСБУ.

Поле	*Значение*
ID	Индекс в БДСБУ.
TH	Указывает, была ли структура установлена в данной статье.
AU	Список авторов.
TI	Название статьи.
JN	Полное название журнала без аббревиатур.
PY	Год публикации.
VL	Том, номер в скобках. Номер в скобках дублируется в поле VL1.
PG	Номера страниц. Указываются первая и последняя страница. Номера первой и последней страниц дублируются в полях PG1 и PG2.
PB	Издательство.
RL	WWW-ссылка на данные статьи.
EA	Электронный адрес автора для корреспонденции.
AD	Участвующие институты.
AB	Аннотация статьи.
ST1	Строковое представление структуры. Однозначная кодировка разработана и документирована (см. Кодировка структуры).
ST2	Тип структуры (повторяющееся звено полимера, биологическое повторяющееся звено полимера, предполагаемое биологическое звено, олигосахарид, моносахарид, повторяющееся звено циклической структуры).
ST3	Молекулярная масса или степень полимеризации
SL	Расположение структуры внутри статьи. Свободнотекстовое поле.
AG	Информация об агликоне.
MF	Молекулярная формула (только для моно- и олигосахаридов).
NMRH	Отнесение спектра ЯМР ^{1}H. Формат этого поля включает названия остатков, химические сдвиги протонов в каждом остатке и последовательность позиций замещения, начиная от невосстанавливающего конца, обеспечивающую однозначную локализацию остатков в структуре,
NMRC	Отнесение спектра ЯМР ^{13}C. Формат поля аналогичен полю HNMR.
NMRS	Растворитель, использованный в ЯМР-экспериментах.
NMRT	Температура образца в ЯМР-экспериментах (в Кельвинах).
SO	Биологический источник без аббревиатур. Сначала род, затем вид, на третьем месте штамм и номер группы. Это три отдельных поля.
NC	(Тривиальное) название соединения.
MT	Названия методов, использованных для установления структуры. Свободнотекстовое поле.
BA	Биологическая активность. Текст, в большинстве случаев ограниченный данными, представленными в статье.
EI	Ферменты, которые синтезируют либо расщеплют структуру.

Продолжение табл. 1.

BG	Наличие биосинтетических или генетических данных. Текст, в большинстве случаев ограниченный данными, представленными в статье.
SY	Наличие данных по синтезу данной структуры. Свободнотекстовое поле.
KW	Ключевые слова, связанные с аннотацией или полным текстом статьи.
NT	Любые комментарии, включая тип соединения, найденные ошибки, нестехиометрические замещения и т. д.
3D	Наличие трехмерной структуры и конформационные данные. Свободный текст.
RR	Перекрестные ссылки идентификаторов записей БДСБУ (ID) с другими структурами из данной статьи.
CA, RN, PT, PR, SD, OT	Перекрестные ссылки на Chemical Abstracts, регистрационный номер вещества в CA, номер патента, белковые БД, спектроскопические БД, прочие БД.
U1, U2, U3, U4	Вспомогательные поля, не показываемые пользователю: имя автора записи, дата последнего обновления и т.д.

2. НАПОЛНЕНИЕ БД

База данных содержит 7645 записей о бактериальных углеводах с известной структурой, что соответствует более чем 99%-ному покрытию по бактериальным углеводам, структуры которых опубликованы на данный момент. Из них около 2500 записей импортировано из базы данных CarbBank, содержащей данные, опубликованные до 1995 г. Остальные записи созданы путем обработки оригинальных источников 1995-2006 гг. с использованием библиографической базы данных коллектива лаборатории химии углеводов ИОХ РАН. Все записи, в том числе импортированные из CarbBank, проверены автоматически и многие проверены вручную, выявленные ошибки исправлены перед наполнением.

Для импорта из формата CarbBank был создан специальный конвертер. Созданы средства автоматической проверки синтаксиса строкового представления структуры, записи имен остатков, определения принадлежности остатков в записи к пространству имен моносахаридов БД. Ошибки записи, не связанные с возможными искажением или потерей информации о структуре, корректируются автоматически, прочие ошибки, выявленные автоматически, отправляются на ручную проверку. Более 50% записей базы прошли подобную проверку вручную после выявления ошибок.

3. ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС

Для базы данных создан дружественный веб-интерфейс, который включает следующие основные элементы:

Поиск по ID. Пользователь указывает один или несколько идентификаторов БДСБУ для получения соответствующей записи.

Библиографический поиск. Эта форма позволяет делать запросы по авторам, названию статьи, терминам из аннотации, названию журналов, году публикации, номерам тома и страниц. Имена авторов проиндексированы с целью предотвращения опечаток. Построение запроса возможно также с использованием ключевых слов и слов, содержащихся в аннотациях к статьям и их заголовках. Язык запросов поддерживает логические операции И, ИЛИ (символы & и | либо AND и OR), группирование терминов (кавычки) и группирование запросов (скобки). Возможно применение поискового фильтра, позволяющего учитывать, была ли структура углевода установлена или уточнена в искомой публикации (рис. 1).

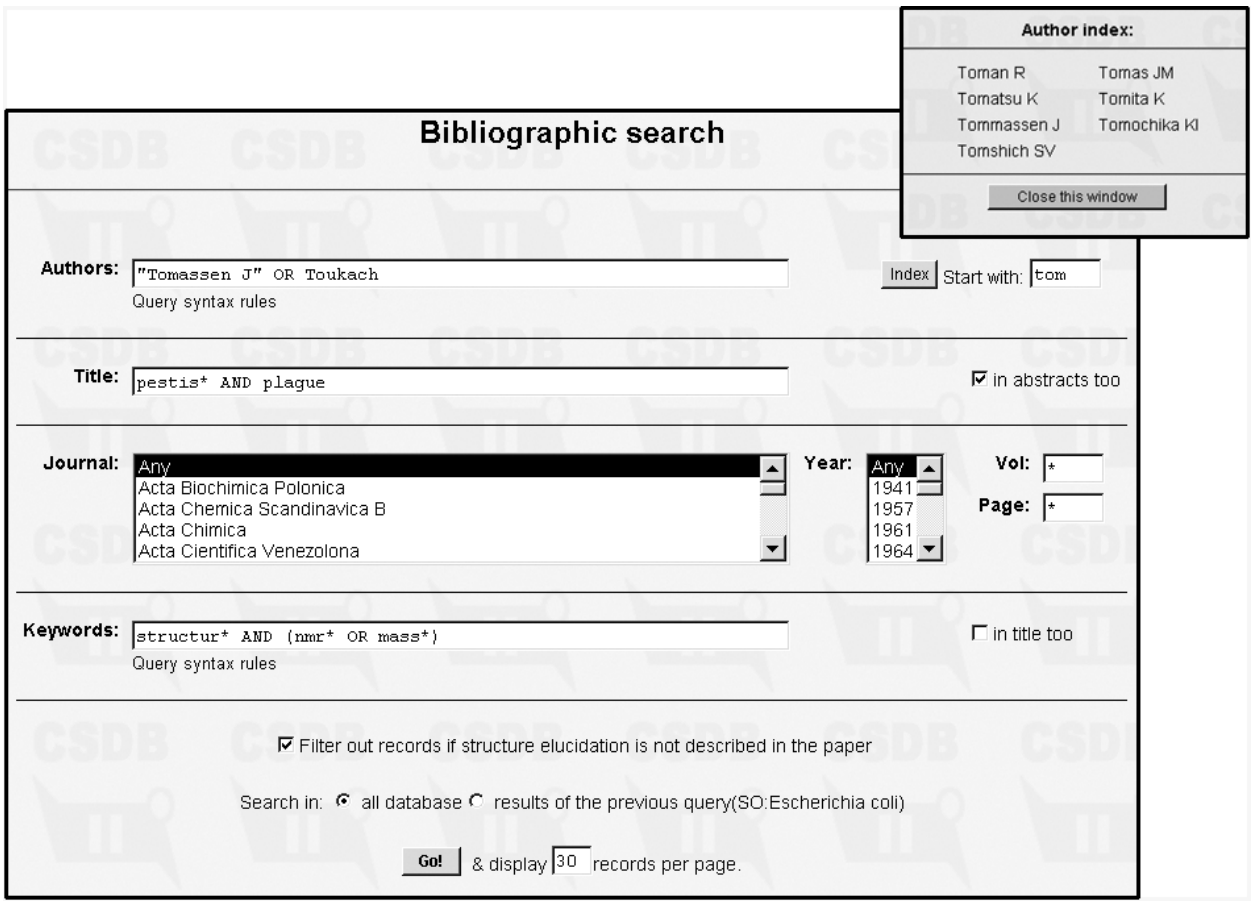

Рис. 1. Поиск по библиографическим данным.

Поиск по биологическому источнику. Эта форма позволяет вести поиск в базе данных, используя в качестве критерия поиска информацию о роде, виде и штамме/серотипе. Все три поля проиндексированы во избежание опечаток, а для поля штамм поддерживается режим использования символов-заменителей (wildcards). После выбора определенного рода микроорганизмов автоматически генерируются списки видов и штаммов/серотипов, соответствующих этому роду (рис. 2).

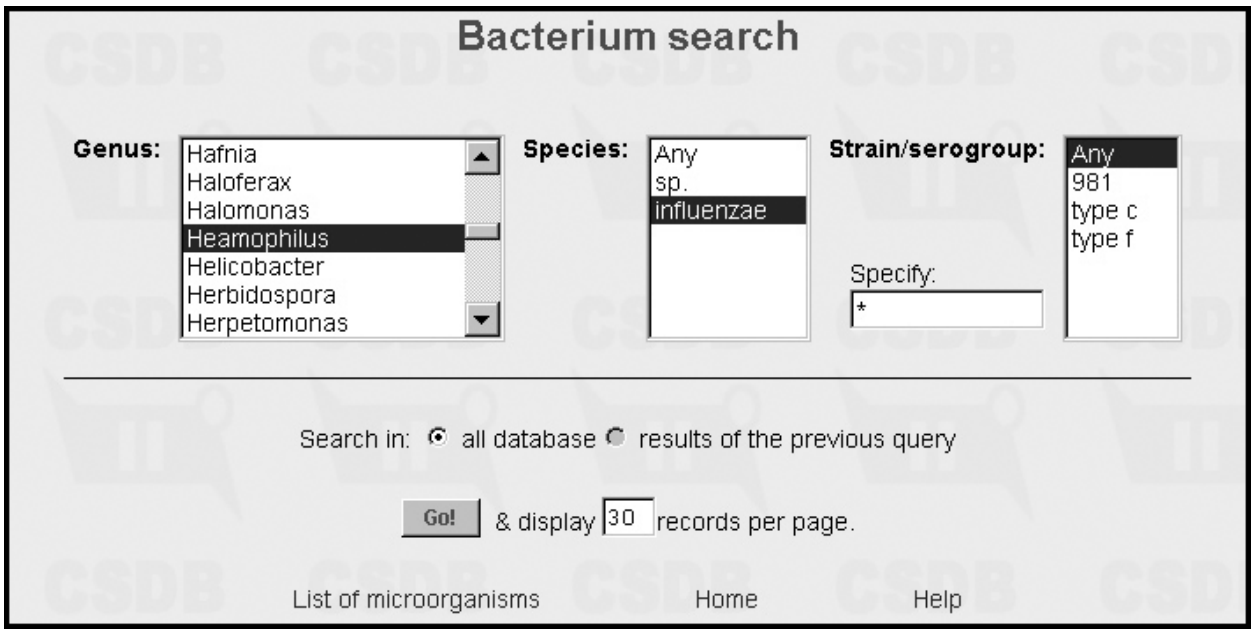

Рис. 2. Поиск по бактериологическим данным.

Структурный поиск. Эта форма позволяет использовать в качестве критерия поиска фрагмент структуры и поддерживает два режима: *эксперт* и *мастер*. В режиме *эксперт* пользователь может самостоятельно указать фрагмент структуры углевода для запуска поисковой программы. Это требует знания правил кодировки структур (см. раздел Кодирование структур), но снимает ограничения поисковых возможностей, накладываемых режимом *мастер* (максимальное число моносахаридных остатков, множественные связи между ними и т. д.). Режим *мастер* позволяет пользователю построить тот же линейный код структуры, используя только визуальные операции с топологией, названиями остатков, конфигурациями, размерами циклов, типами связей, присоединением заместителей и т.д. Этот метод не требует набора символов вручную и знания правил кодировки структур. Все вышеперечисленные параметры могут быть выбраны из выпадающих списков и дочерних окон (рис. 3).

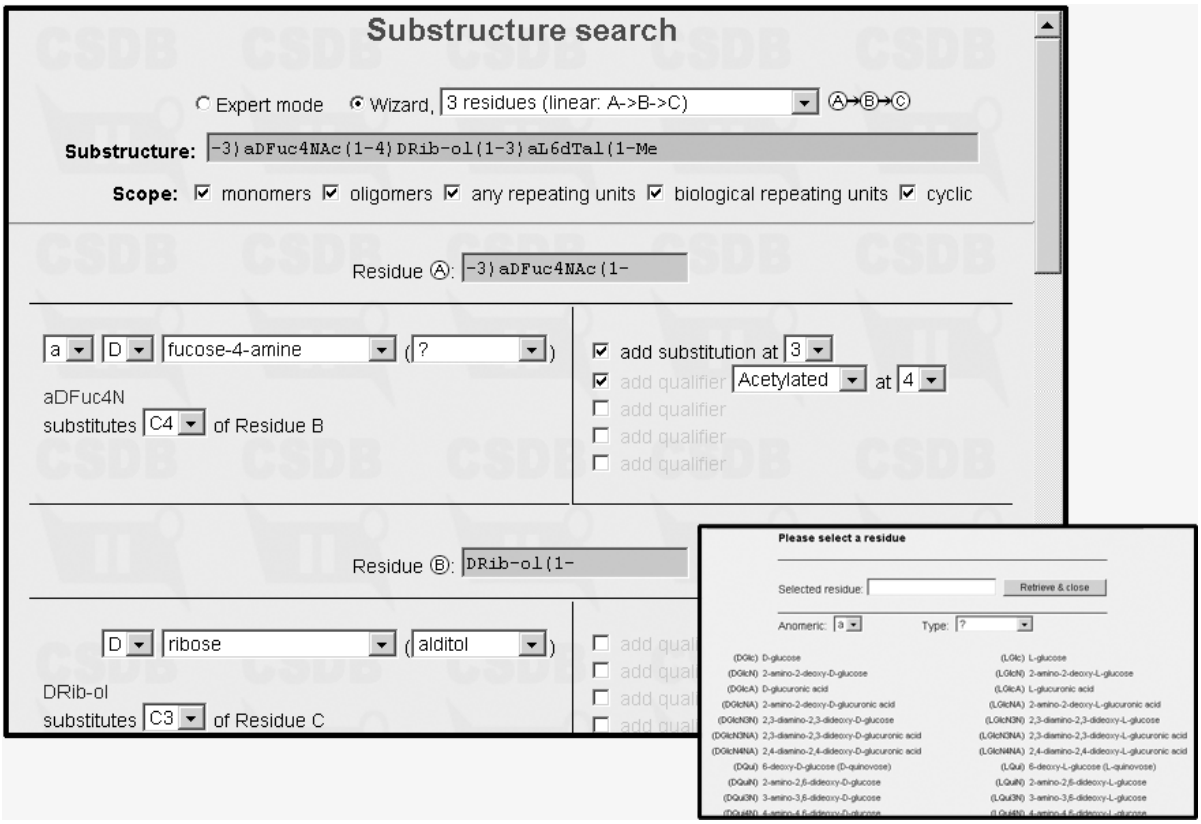

Рис. 3. Мастер структур.

Результат поиска по любому запросу отображается в виде списка записей, каждая из которых представлена в сжатом виде. Эта форма отображает наиболее важную информацию: авторов, название публикации, выходные данные, структуру и биологический источник. Если число найденных записей превышает лимит, указанный в запросе, результат разбивается на несколько страниц, по которым пользователь имеет возможность перемещаться (рис. 4). Расширенная запись, которую можно увидеть, развернув «сжатую», включает все данные, имеющиеся по данному углеводу в БДСБУ.

На этапе формирования запроса может быть указана одна из двух областей поиска: вся база данных или только результаты предыдущего запроса. На этапе отображения результатов запрос может быть уточнен («искать в найденном») дополнительными ограничениями на библиографические данные, структуру или биологический источник.

Пополнение базы данных. Пользователь может предложить пополнение, заполнив специальную форму или загрузив собственный заранее подготовленный файл в документированном формате. Загруженные пользователем данные помещаются в специальную область базы данных и недоступны для поиска, пока администратор не проверит (в том числе с использованием доступных ему автоматических средств) и не подтвердит возможность их включения в БДСБУ.

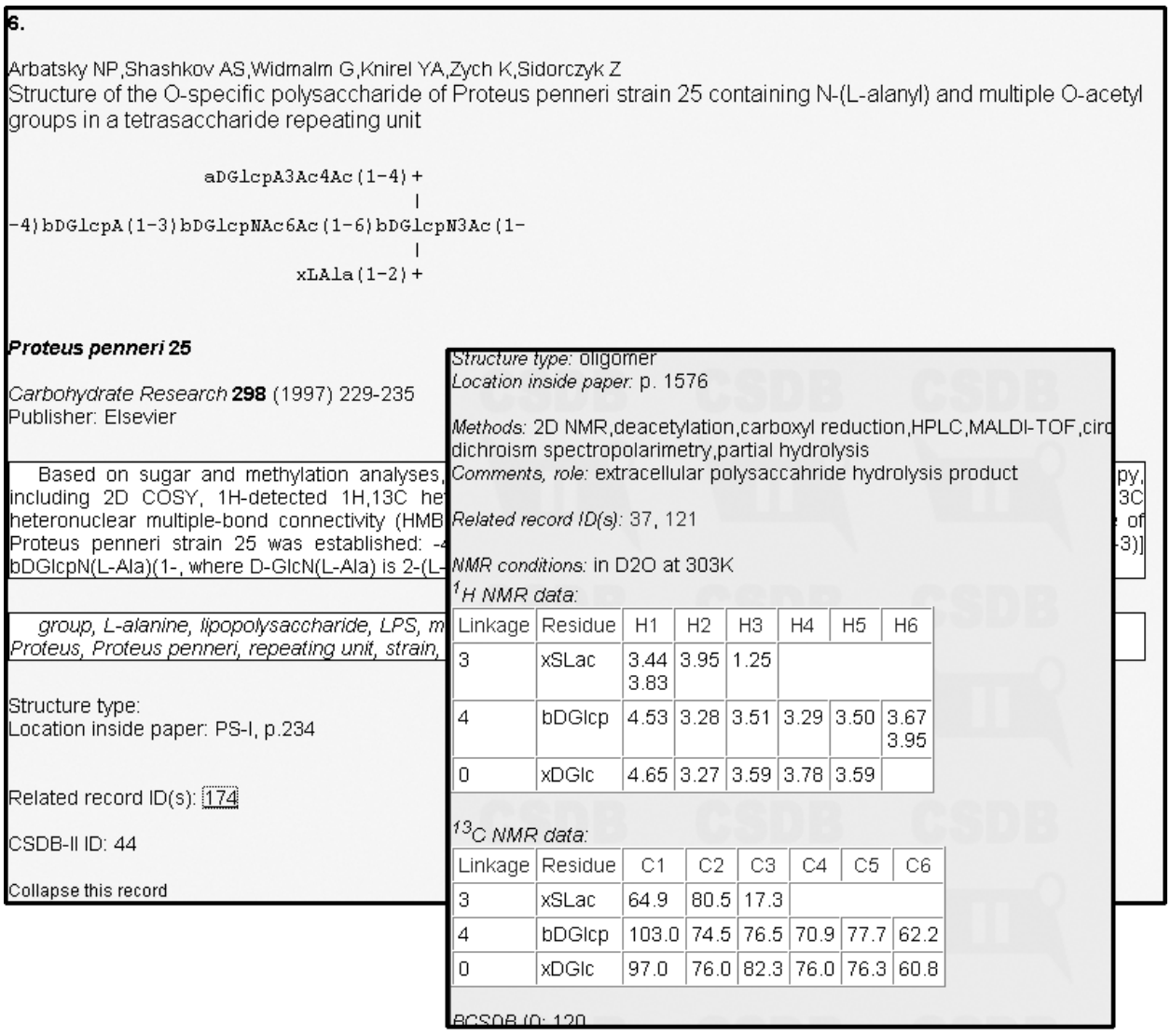

Рис.4. Пример вывода записи базы данных.

4. КОДИРОВАНИЕ СТРУКТУР

Линейная кодировка, разработанная специально для БДСБУ, предназначена для однозначного описания структур любых углеводов. Ниже приведено краткое описание правил записи структур с ее помощью:

Топология. Моносахаридные остатки описываются последовательностью терминов, таких как **<residue name>(<goes by>-<goes to>)**, где **goes by** обозначает положение внутри остатка (номер атома углерода), которым он замещает другой остаток (обычно 1 или 2), а **goes to** – в какое положение он замещает соседний остаток. В случае моносахарида на восстанавливающем конце обозначение в скобках не требуется. Например, **A(1-3)B(1-4)C** – это линейный фрагмент, в котором остаток A замещает своим первым положением остаток B в положение 3, остаток B замещает остаток C в положение 4, а остаток C ничего не замещает. Здесь и далее латинские заглавные буквы обозначают названия остатков. Если данная структура – полимер, крайний левый и крайний правый остатки должны иметь открытые связи: **-2)A(1-3)B(1-4)C(1-** кодирует ту же структуру, что и выше, но с одним отличием: это полимер из повторяющихся звеньев, соединённых между собой 1-2 связью.

Если в углеводе имеются разветвления, то одна из цепей принимается за главную, а остальные – за боковые. Правила для определения главной цепи описаны ниже. Боковые цепи заключаются в квадратные скобки с обозначением связей в круглых скобках. Например, **A(1-3)[B(1-4)]C** – это разветвлённый фрагмент, в котором остаток A замещает остаток C в положение 3, в то время как остаток B замещает тот же остаток C в положение 4. В более сложных случаях несколько боковых цепей,

связанных с одним остатком, разделяются в квадратных скобках запятыми. Боковые цепи также могут быть разветвлённые и тогда предусмотрено вложение квадратных скобок друг в друга, например, **-4)A(1-3)[D(2-6)B(1-4),F(1-3)[G(1-4)]E(1-2)]C(1-** соответствует структуре, приведенной на рис. 5.

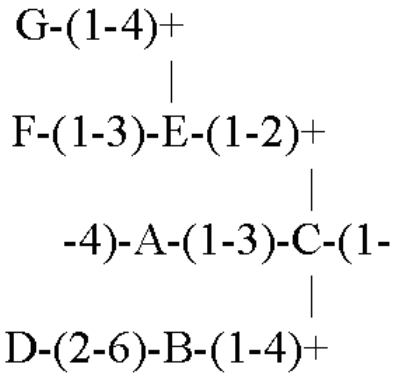

Рис. 5. Пример разветвленной углеводной структуры.

Неуглеводные заместители. Все одновалентные заместители (Ac, Me, Fo и другие остатки, которые сами не могут быть замещены) обозначаются как отдельные остатки. Так, фрагмент **aDGal(1-3)bDGlcNAc** записывается как **aDGal(1-3)[Ac(1-2)]bDGlcN**. Если моновалентный остаток – агликон на восстанавливающем конце, он записывается следующим образом: **aDGlc(1-Me**. В процессе поиска структуры (но не при наполнении базы данных) одновалентные остатки могут быть указаны обычным путём, например **bDQuiNAc3NAc4Ac**. Фосфаты и сульфаты с соответствующими связями заключаются в скобки следующим способом: **aDGlc(1-P-4)bLFuc** (в цепи), **P-4)bLFuc** (на невосстанавливающем конце) или **aDGlc(1-P** (на восстанавливающем конце). Заместители, которые связываются с одним и тем же остатком двумя связями, записываются дважды с разделением двоеточием. Например, запись:

-2)[xRPyr(2-4):xRPyr(2-6]aDGal(1-

соответствует ацеталю пирувата в положениях 4 и 6 остатка галактозы. Если такой остаток находится на невосстанавливающем конце цепи, используется следующая запись: **xRPyr(2-4)[:xRPyr(2-6]aDGal(1-….**

Различия между главной и боковой цепями; порядок боковых цепей. Для однозначности кодировки необходимо определить, какая цепь закодирована как главная, а какие как боковые:

– в случае полимера цепь, формирующая его остов, всегда является главной.

– в случае олигомеров и фрагментов структуры, если одна цепь является моносахаридным или другим неодновалентным остатком, а другая представлена одновалентным заместителем, первая цепь принимается за главную. Например, запись:

...aDGal(1-6)[Ac(1-4)]bDGlc...

является правильной,

а запись:

Ac(1-4)[...aDGal(1-6)]bDGlc... –

неправильной. Если обе цепи являются моносахаридными остатками или обе одновалентными заместителями, главной считается цепь, находящаяся в положении с меньшим номером. Например,

...Ac(1-3)[Ac(1-4)]bDGlc... – правильно, а **...Ac(1-4)[Ac(1-3)]bDGlc...** – неправильно.

– при наличии нескольких боковых цепей, связанных с одним моносахаридным остатком их располагают внутри квадратных скобок в порядке убывания номера положения, в которое они присоединены: **...[Ac(1-6),Ac(1-2)]...,** но не **...[Ac(1-2),Ac(1-6)]...**.

Названия моносахаридных остатков. Полное название каждого моносахаридного остатка состоит из нескольких полей, следующих одно за другим без разделителей в следующей последовательности:

– аномерная конфигурация. Обозначается одной строчной буквой: **a** – альфа, **b** – бета, **x** – данный остаток не имеет аномерных форм или представлен смесью аномеров, **?** – аномерная конфигурация неизвестна. Одновалентные заместители не требуют этого поля.

– абсолютная конфигурация.

Обозначается одной заглавной буквой: **D, L, R** или **S –** соответствует стандартным обозначениям абсолютных конфигураций**, X** – данный остаток не имеет абсолютных форм, **?** – абсолютная конфигурация неизвестна. Если абсолютная конфигурация является частью основного названия остатка (например, а**Abe**), в качестве абсолютной конфигурации указывается **X (aXAbe)**. Одновалентные заместители не

требуют этого поля.

– основное название остатка; если необходимо, включает информацию о дезоксигенировании. Первая буква – заглавная, остальные – строчные (например, 6dTal).

– указатель размера цикла. Обозначается одной строчной буквой: **p** – пираноза, **f** – фураноза, **a** – открытая цепь. Если не указано, остаток может быть в любой форме.

– указатель карбоксильной группы. Заглавная буква **A** обозначает остаток уроновой кислоты (например, aDGlcpA).

– указатель аминогруппы или аминогрупп. Одна или несколько заглавных букв **N**; если положение аминогруппы отличается от 2, это указывается перед заглавной буквой **N** (например, **aLRhap4N)**. Если аминогруппы включены в название остатка, они не указываются (например, **xLLys,** но не **xLLysN6N)**.

– любые другие модификаторы, если таковые присутствуют. Допускаются несколько модификаторов, например, **aDGlcpAN3N**.

– модификатор **-ol**, если остаток является полиолом (если эта информация не содержится в основном названии, например, как в случае **Gro**). Этот указатель не может сочетаться с указателем размера цикла.

Примеры: **aD6dTalA, aXKdo, xLGro, ?DManN-ol, Ac, xRPyr, bDFucN3N**.

5. ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Для управления базой данных используется СУБД MySQL 4, основные функциональные и административные инструменты реализованы на языках Perl 5 и PHP 4. Оформление веб-сайта и интерактивное взаимодействие с пользователем осуществляется с помощью страниц на языках JavaScript и DHTML. Совместимость с браузерами не ограничена, за исключением структурной формы в режиме *мастер*, которая требует браузера Microsoft Internet Explorer 5 или более позднюю версию.

Входная страница БДСБУ доступна по адресу http://www.glyco.ac.ru/bcsdb/.

6. ЗАКЛЮЧЕНИЕ

Была создана открытая база данных мирового уровня по бактериальным углеводам, существенно облегчающая интерпретацию накопленной информации по этим соединениям. Она расширяет возможности специалистов всего мира, исследующих структуру и биологические функции бактериальных углеводов и родственных соединений, включая гликолипиды и гликопротеины. Идеология БДСБУ представлена на XVIII Международном симпозиуме по гликоконъюгатам [9] и III Немецко-русско-польской встрече по бактериальным углеводам [10].

Разработка базы данных БДСБУ поддерживается грантами РФФИ 05-07-90099 и ГПРФ МК-1700.2005.4. Проект был инициирован в рамках работ по гранту МНТЦ 1197р.

ЛИТЕРАТУРА:

1. R. Raman, M.Venkataraman, S.Ramakrishnan, W.Lang, S.Raguram, R. Sasisekharan. Advancing Glycomics: Implementation Strategies at the Consortium for Functional Glycomics. Glycobiology. – 2006. – Vol. 16.– P. 82R-90R.
2. J. Lowe, J.Marth. //Annu. Rev. Biochem. – 2003. – Vol.72. – P.643-691.
3. C. von der Lieth, T. Lutteke, M. Frank: //Biochim Biophys Acta. – 2006. – Vol. 1760. – P. 568-577.
4. D. Harvey. //Expert Rev Proteomics. – 2005. – Vol.2. – P. 87-101.
5. Y.Guerardel, LChang, E. Maes, C. Huang, K. Khoo. //Glycobiology.–2006.–Vol.16.–P. 244-257.
6. S. Doubet, K. Bock, D. Smith, A. Darvill, P. Albersheim. //Trends Biochem Sci. – 1989. –Vol.14. – P. 475-477.
7. T. Lutteke, A. Bohne-Lang, A. Loss, T. Götz, M. Frank, C. von der Lieth: //Glycobiology.– 2006. – Vol. 16. – P. 71R-81R.
8. C. Cooper, H. Joshi, M. Harrison, M. Wilkins, N. Packer. //Nucleic Acids Res. – 2003. – Vol. 31. – P. 511-513.
9. F.V. Toukach, Y.A. Knirel: //XVIII International Symposium on Glycoconjugates, Italy, Florence. – 2005. – P. 216-217.
10. F.V. Toukach, Y.A. Knirel. //III German-Russian-Polish meeting on bacterial carbohydrates, Poland, Wroclaw. – 2004. – P. O2.