УДК 532.6

БАЗА ДАННЫХ (ИНФОРМАЦИОННАЯ СИСТЕМА) ГИБРИДНОГО ТИПА ПО МОЛЕКУЛЯРНЫМ СПЕКТРАЛЬНЫМ ПОСТОЯННЫМ (ИСМОЛ)

И. В. Кочиков[1], Г. М. Курамшина[2], Л. М. Самков[3], Д. А. Шарапов[4], С. А. Шарапова[4]

Спроектирована и реализована база данных (информационная система) гибридного типа с элементами ГИС-технологий по молекулярным постоянным соединений сырой нефти и продуктов ее переработки, включающая в себя статический справочно-информационный блок и динамический блок, представляющий собой информационно-вычислительную интегрированную систему. Работа выполнена при поддержке грантов РФФИ–ЮГРА (код проекта № 03–07–96842) и РФФИ–ОБЬ (код проекта № 05–07–98001).

Ключевые слова: молекулярные постоянные, базы данных, нефть, информационно-вычислительные системы, колебательная спектроскопия.

1. Введение. Автоматизация физико-химических исследований, использующих инструментальную базу колебательной спектроскопии, выдвигает на первый план вопросы обработки и корректной интерпретации данных эксперимента. Создание единой библиотеки данных по строению и колебательным спектрам для всех классов веществ является неосуществимой задачей, поскольку число известных соединений приближается к десяти миллионам и далеко не для всех имеются исследования их спектров. В прагматическом плане более актуальным является создание развитых баз спектральных данных по определенным классам или группам соединений. Для интерпретации современного спектрального эксперимента часто необходим анализ нормальных колебаний молекулярной системы некоторого предполагаемого строения, требующий наличия соответствующих сведений по молекулярным структурам и молекулярным силовым полям. При этом используемые силовые постоянные должны обладать гарантированными свойствами переносимости в рядах родственных соединений.

Расчет силовых постоянных производится с использованием экспериментальных данных о колебательных спектрах, причем в последнее десятилетие, как правило, все больше и больше опирается на результаты квантовохимических расчетов, которые сами по себе становятся неотъемлемой частью спектрального эксперимента.

2. Принципы организации базы данных по спектральным постоянным. Представляемая БД (информационная система ИСМОЛ) [1] создана на химическом факультете МГУ им. М. В. Ломоносова на основе разработанных в последние годы устойчивых численных методов для решения задач молекулярной спектроскопии [2, 3]. Актуальность ее создания определяется необходимостью быстрых и эффективных методов проведения качественного и количественного структурно-группового анализа нефти и продуктов ее переработки. Знание состава нефти и нефтепродуктов является определяющим фактором для правильного выбора метода переработки нефти и расчета технологических процессов. В число наиболее информативных физико-химических методов, используемых для решения этой задачи, входят методы молекулярной, в частности колебательной, спектроскопии (инфракрасного поглощения и комбинационного рассеяния). Автоматизация исследований, использующих инструментальную базу колебательной спектроскопии, выдвигает на первый план вопросы обработки и корректной интерпретации данных эксперимента. Внедрение ЭВМ требует существенного усовершенствования скорости поиска и обработки библиографических данных и, соответственно, более рациональной организации структур данных по молекулярным структурам, колебательным спектрам, термохимическим данным, а также увеличения объемов доступной информации. Наличие развитой теории колебательных спектров многоатомных молекул в сочетании с данными высокоточного эксперимента позволяет получать уникальные данные о строении сложных систем, необходимые для глубокого понимания сути различных физических и химических

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, 119992, Москва; e-mail: kochikov@tm-net.ru

[2] Московский государственный университет им. М. В. Ломоносова, химический факультет, кафедра физической химии, Воробьевы горы, 119992, Москва; e-mail: kuramshi@phys.chem.msu.ru

[3] Югорский научно-исследовательский институт информационных технологий, ул. Мира, д. 151, 628011, г. Ханты-Мансийск; email: saml@uriit.ru

[4] Московский государственный университет им. М. В. Ломоносова, физический факультет, кафедра математики, Воробьевы горы, 119992, Москва; e-mail: sharapov@srcc.msu.ru; sharapova@srcc.msu.ru

процессов, предсказания свойств неустойчивых систем и моделирования различных физико-химических процессов, протекающих в специфических условиях.

В последние годы произошел качественный скачок в развитии квантовохимических (неэмпирических и полуэмпирических) методов, позволяющих с весьма высокой степенью достоверности (в ряде случаев приближающейся к экспериментальной точности) предсказывать строение, спектры и другие физико-химические характеристики многоатомных молекулярных систем. Доступность развитых комплексов программ, таких как системы GAUSSIAN XX и GAMESS, позволяет в анализе экспериментальных данных широко использовать теоретические данные для оперативной оценки ряда молекулярных характеристик.

Существует ряд баз данных в различных областях химии, сфокусированных на отдельных классах веществ и их свойствах. Отметим наиболее известные и успешно развиваемые проекты — такие, как базы данных Национального бюро стандартов США [4], в которых аккумулированы разнообразные сведения по строению и физико-химическим свойствам соединений. В некоторых БД [5] возможен свободный доступ к части информации по термодинамическим данным, по экспериментальным спектрам и теоретическим расчетам большого круга органических соединений. В БД по квантовомеханическим расчетам (Computational Chemistry Comparison and Benchmark Data Base) [5] представлена информация для молекул, в которых число так называемых "тяжелых" атомов (атомов, отличающихся от атома водорода) не превышает шести, а общее число атомов не больше 20. Однако наиболее важная структурная и спектральная информация предоставляется, как правило, на коммерческой основе.

В последние годы получила известность база данных MOGADOC (документация по молекулярным свойствам соединений в газовой фазе) [6, 7], в которой для большого числа органических и неорганических соединений собрана информация по электронографическим данным, микроволновым спектрам и молекулярной радиоастрономии. Существует большое число баз спектральных данных, которые в основном распространяются на коммерческой основе (см. список на сайте [8]), например FDM Reference Spectra Database [9].

Для интерпретации спектра соединения неизвестного строения обычно используют сравнение его спектра с известными спектрами соединений близкого строения. В идеальном случае база данных должна включать в себя картотеку эталонных колебательных спектров, полученных в стандартных или некоторых других условиях регистрации, близких к условиям, в которых получен спектр неизвестного объекта, что позволяет использовать их при идентификации анализируемого соединения. Для интерпретации спектра также используются известные области групповых частот и привлекаются теоретические оценки спектра, для чего необходимы данные о силовых постоянных, обладающих свойством переносимости. Известно, что расчет силовых постоянных по экспериментальным данным является некорректно поставленной задачей [2, 3], что необходимо учитывать при составлении базы данных по силовым полям многоатомных молекул.

Кроме того, современное видение справочных пособий в электронном виде предполагает то, что такое пособие может и должно включать в себя не только информационную базу (опытные, расчетные, модельные и справочные данные), но и комплекс программ для получения различных теоретических оценок, т.е. иметь гибридный характер. В соответствии с этим пособие по молекулярным постоянным должно содержать как статический блок — информационную систему, содержащую экспериментальные расчетные, модельные и справочные данные по колебательным спектрам и равновесным конфигурациям молекул и физико-химическим характеристикам соответствующих веществ, так и динамический блок — комплекс программ для расчета теоретических колебательных спектров в различных приближениях и ряд сервисных программ для оперативного решения сопутствующих задач при интерпретации экспериментальных данных.

Можно попытаться сформулировать ряд основных требований, которым должна удовлетворять современная база данных по спектральным постоянным многоатомных молекул для того, чтобы представлять собой полезный и эффективный инструмент для прикладных и фундаментальных исследований:

— ограничение по кругу объектов, включаемых в БД по классам соединений и по тем или иным дополнительным критериям;

— наличие обширной картотеки эталонных ИК-спектров поглощения, полученных в условиях регистрации, приближенных к реальным условиям существования данного вещества с достаточным разрешением;

— наличие банка данных по силовым постоянным и геометрическим параметрам молекул для возможной оперативной оценки теоретических колебательных спектров молекулы и расчета термодинамических свойств вещества;

— наличие комплекса программ для решения прямых и обратных задач спектроскопии на основе устойчивых численных методов, в котором предусмотрена возможность управления выбором систем обоб-

щенных координат и моделей силовых полей и получения силовых постоянных, обладающих свойством переносимости в рядах родственных соединений с целью последующего включения рассчитанных величин в БД;

— наличие квантовохимических данных, соответствующих определенному уровню теории (или ряду таких уровней в рамках так называемой модельной химии) и обеспечивающих достаточное надежное предсказание свойств молекул рассматриваемых классов;

— возможность идентификации смеси веществ и решения ряда практических задач интерпретации молекулярных спектров.

Перечисленные требования были положены в основу созданной базы данных по молекулярным постоянным ИСМОЛ для соединений, входящих в состав сырой нефти, а также важнейших продуктов их переработки (углеводороды и соединения, содержащие серу, кислород, азот и ряд других сопутствующих органических и неорганических соединений). Кроме того, БД включает данные по геофизическим и физико-химическим свойствам нефтей, добываемых в ряде месторождений Ханты-Мансийского региона. Нами не обнаружены БД, ориентированные именно на спектральные и структурные данные для соединений, входящих в состав сырой нефти и продуктов ее переработки.

Динамический блок БД представляет собой комплекс вычислительных программ, позволяющий оперативно производить необходимые расчеты для получения ряда спектральных характеристик и термодинамических функций индивидуальных веществ, результаты которых могут быть использованы при интерпретации спектров и идентификации соединений, моделирования спектров и свойств сложных соединений.

Поскольку квантовомеханические расчеты молекул стали неотъемлемой частью практически любого спектрального эксперимента, в данную БД включены результаты теоретических расчетов для соединений C1 – C10, проведенных в рамках наиболее используемых в настоящее время подходов — на уровне Хартри–Фока [10], теории возмущений Меллера–Плессе второго порядка [11, 12] и в приближении теории функционала плотности [13, 14].

В базу данных включены таблицы характеристических частот колебаний и наборы силовых постоянных, полученные в рамках унифицированных моделей силовых полей с учетом некорректности соответствующих обратных задач, обладающие гарантированными свойствами переносимости в рядах родственных соединений и позволяющие быстро и эффективно производить полные расчеты колебательных спектров и идентифицировать соединения близкого строения. В число справочных данных, входящих в эту БД, включены величины силовых постоянных, полученных как при квантовомеханических расчетах, так и при решении обратных задач с использованием экспериментальных данных.

3. Общая структура базы данных. Основой представляемой системы является специально разработанная и реализованная программно-технологическая среда, осуществляющая функции хранения, управления и обработки экспериментальных и теоретических спектральных и структурных данных и формирования выходных данных для индивидуальных веществ в виде текстового, табличного и графического материалов. Комплекс программ включает в себя информационную базу, структурированную по различным кластерам (экспериментальные, теоретические и справочные данные по строению, колебательным спектрам, силовым полям и другим молекулярным постоянным, а также физико-химическим свойствам для индивидуальных веществ), которая хранится в специально разработанных таблицах, а также ряд программ для решения прямых и обратных задач молекулярной спектроскопии и моделирования спектров известных и неизвестных соединений с подключением данных квантовомеханических расчетов, для оценки термодинамических свойств методами статистической термодинамики и ряда вспомогательных программ: анализа симметрии молекулы и построения координат симметрии и матрицы приведения по симметрии, визуализации молекулярной структуры и молекулярных колебаний и др.

При проектировании базы данных предусмотрен ее гибридный характер, который предполагает сочетание статической части, представляющей собой справочный инструмент, и динамической части, представляющей собой информационно-вычислительную интегрированную систему. База данных реализована как комплекс программ, написанный на Visual C++. Разработанная система состоит из программы-оболочки, собственно базы данных по молекулярным постоянным, поискового модуля, комплекса программ для распознавания образов, а также комплекса программ для расчета колебательных спектров, молекулярных силовых постоянных, термодинамических свойств веществ. В систему также включен ряд сервисных программ и программ визуализации, средства подготовки и отображения отчетных материалов и модуль экспертной поддержки.

Для создания комплекса использовалась среда программирования Microsoft Developer Studio; программирование выполнено на языке C++ с помощью компилятора Microsoft Visual C++ с использованием

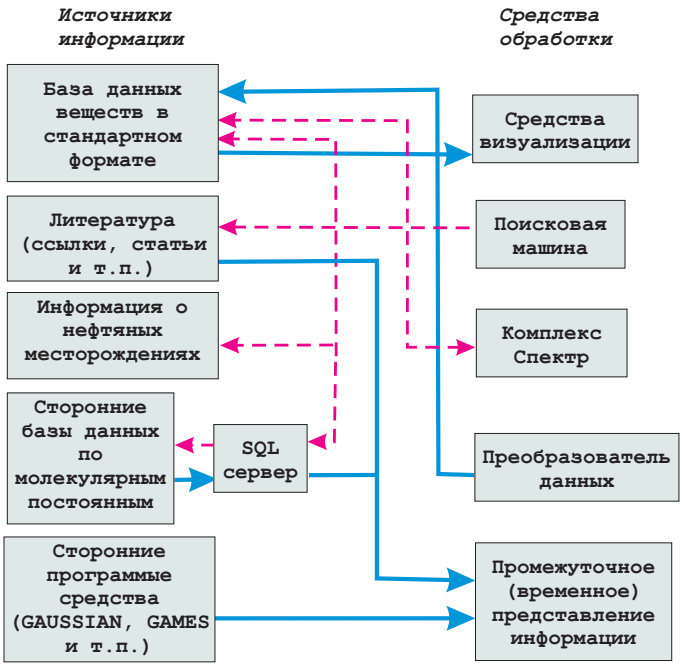
технологии MFC (Microsoft Foundation Classes).

При проектировании программы заложена возможность ее общения с существующими (или новыми) базами данных в стандарте SQL (структурированный язык запросов). Это дает возможность обмена информацией практически с любой стандартной реляционной базой данных как в локальном, так и в сетевом варианте. Программы сторонних производителей (типа GAUSSIAN и GAMESS) не могут быть интегрированы в той же степени, но могут запускаться из-под программной оболочки базы данных. В дальнейшем такие программы работают автономно и передача информации между ними и базой осуществляется через входные/выходные файлы.

Программная оболочка поддерживает анализ и импорт данных из стандартных выходных файлов широко используемых программ квантовомеханических вычислений (в том числе GAUSSIAN, GAMESS, Природа). Поддерживаются также стандартные форматы представления спектральных данных (форматы ASP и JDX).

Общая структура созданного комплекса программ представлена на рисунке.

Существенной частью базы данных является созданный впервые набор утилит, позволяющих преобразовать данные различных видов в единый формат и, таким образом, осуществлять импорт и (если требуется) экспорт материалов. Разработаны способы автоматизированного представления пользовательской документации, созданы способы преобразования структурных разделов исходных документов и предложены соответствующие шаблоны, исследованы вопросы обработки и хранения графической информации, возникающие при разработке систем визуализации данных при учете различных форматов хранения изображений и способов установки текущей палитры экрана. БД построена так, чтобы при необходимости можно было ввести новые поля (например, добавить дополнительные данные какого-либо нового типа физического эксперимента), причем СУБД должна уметь работать со старыми данными (без этих полей) без их модификации. Введение описания структуры данных аналогично структуре файлов графического формата TIFF, где описание каждого изображения может содержать (помимо нескольких обязательных полей) произвольное число дополнительных полей.



Общая схема созданного комплекса программ

Ожидаемый объем данных по молекулярным постоянным — от 50 – 100 кбайт на небольшую молекулу до 500 – 1500 кбайт для больших соединений. Общий объем базы данных составит около 1 Гбайт. Объем документации (статьи, таблицы и т.п.) может составить примерно такую же величину, учитывая, что база данных позволяет хранить и осуществлять доступ как к локально хранимым документам, так и к ссылкам в сети Интернет.

Способ доступа пользователей к базе данных на настоящем этапе предполагается локальным (система устанавливается на компьютере пользователя); в дальнейшем предполагается возможность удаленного доступа к данным с помощью SQL-сервера.

4. Выводы. Создана база данных сложной архитектуры (информационная система) ИСМОЛ с элементами ГИС-технологий по молекулярным постоянным соединений сырой нефти и продуктов ее переработки, представляющая собой оригинальный программный продукт.

База данных имеет гибридный характер, в ней сочетаются статическая справочная часть и динамическая часть. Она может рассматриваться как информационно-вычислительная интегрированная система, которая позволяет выполнять различные вычисления и необходимые для этого преобразования данных. Для этого осуществлено интегрирование базы данных с пакетом программ СПЕКТР, осуществляющим решение прямых и обратных задач колебательной спектроскопии. Созданы программы для визуализации и табличного представления расчета частот колебаний многоатомных молекул. В комплекс программ

включена программа для расчета термодинамических функций веществ методами статистической термодинамики. Проведено включение в БД информации по спектрам и молекулярным постоянным для основных представителей углеводородов — алканов и алкенов в виде табличных, графических и текстовых материалов. Отличительной чертой созданной базы данных является возможность работы с отдельными соединениями, выходящая за рамки систематизации информации, т.е. возможен теоретический анализ колебательных спектров, оценка термодинамических функций вещества, синтез силовых полей многоатомных молекул из отдельных фрагментов. Фактически бóльшая часть процедуры исследования может быть проведена в рамках созданного пакета программ.

В базу данных включены справочные данные о составе нефти, физико-химических свойствах входящих в нее соединений, методах ее переработки и данные о ряде месторождений Ханты-Мансийского автономного округа.

СПИСОК ЛИТЕРАТУРЫ

1. *Kochikov I.V., Kuramshina G.M., Samkov L.M.* A hybrid database on spectral data: reference and research tool // Proc. of the 20th Austin Symposium on Molecular Structure. Austin (USA), 2004.
2. *Кочиков И.В., Курамшина Г.М., Пентин Ю.А., Ягола А.Г.* Обратные задачи колебательной спектроскопии. М.: Изд-во Моск. ун-та, 1993.
3. *Yagola A.G., Kochikov I.V., Kuramshina G.M., Pentin Yu.A.* Inverse problems of vibrational spectroscopy. Zeist (The Netherlands): VSP, 1999.
4. http://www.nist.gov
5. http://www.nist.gov/srd/onlinelist.htm
6. *Vogt J., Vogt N., Schunk A.* Databases in Inorganic Chemistry // Handbook of chemoinformatics: From data to knowledge. Volume 2. N.Y.: Wiley, 2003. 629–643.
7. *Vogt J., Vogt N., Kramer R.* Visualization and substructure retrieval tools in the MOGADOC database // J. Chem. Inform. Comput. Sci. 2003. **43**. 357–361.
8. http://www.lohninger.com/spectroscopy/dbsurvey.html
9. http://fdmspectra.com
10. *Roothaan C.C.J.* New developments in molecular orbital theory // Rev. Mod. Phys. 1951. **23**. 69–89.
11. *Möller C., Plesset M.S.* Note on an approximation treatment for many-electron systems // Phys. Rev. 1934. **46**. 618–622.
12. *McWeeny R., Dierksen G.* Self-consistent perturbation theory. II. Extension to open shells // J. Chem. Phys. 1968. **49**. 4852–4856.
13. *Becke A.D.* Density-functional thermochemistry. I. The effect of the exchange-only gradient correction // J. Chem. Phys. 1992. **96**. 2155–2160.
14. *Lee C., Yang W., Parr R.G.* Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density // Phys. Rev. 1988. **B 37**. 785–789.
15. *Kochikov I.V., Kuramshina G.M., Sharapov D.A., Yagola S.A.* Data base of quantum mechanical and regularized force constants in redundant internal coordinates // Proc. of the 12th Conference on Current Trends of Computational Chemistry. Jackson (USA), 2003. 87–90.
16. *Kochikov I.V., Kuramshina G.M., Sharapov D.A., Yagola S.A.* Self-consistent model for the joint treatment of spectroscopic and electron diffraction data // Proc. of the 12th Conference on Current Trends of Computational Chemistry. Jackson (USA), 2003. 91–94.

Поступила в редакцию
09.11.2005