УДК 532.6

СТРУКТУРИРОВАНИЕ И ФОРМАЛИЗАЦИЯ ИНФОРМАЦИИ В БАЗЕ ДАННЫХ 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]

Рассмотрены принципы представления и структурирования информации в информационной системе ИСМОЛ — базе данных гибридного типа по молекулярным постоянным органических соединений сырой нефти и продуктов ее переработки. Работа выполнена при поддержке гранта РФФИ–ОБЬ (код проекта № 05–07–98001).

Ключевые слова: базы данных, информационные системы, молекулярные постоянные, визуализация, спектроскопия, многоатомные молекулы.

1. Введение. Созданная база данных ИСМОЛ [1] носит гибридный характер. В ней сочетаются статическая справочная часть и динамическая часть, или информационно-вычислительная интегрированная система. Последняя позволяет выполнять различные вычисления и необходимые преобразования данных. Для этого база данных интегрирована с пакетом программ СПЕКТР [2, 3], осуществляющим решение прямых и обратных задач колебательной спектроскопии. В БД также включены программы для визуализации и табличного представления частот колебаний многоатомных молекул, программа для расчета термодинамических функций веществ методами статистической термодинамики. БД содержит информацию 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 (теоретический анализ колебательных спектров, оценка термодинамических функций вещества, синтез многоатомных молекул из отдельных фрагментов и др.), что выходит за рамки одной только систематизации информации. 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, а также данные о ряде месторождений Ханты-Мансийского автономного округа (ХМАО). В настоящей статье рассмотрены основные принципы представления и структурирования информации в ИСМОЛ.

2. Структурирование информационной базы данных. Информационная система состоит из нескольких разделов, представленных на титульной странице базы (рис. 1).

Раздел “Вещества” позволяет войти в базу данных химических соединений и их фрагментов и предназначен для исследовательской работы.

Раздел “Документы” — база документов и ссылок общего назначения (доступ к литературе по отдельным соединениям может быть осуществлен также из раздела “Вещества”).

Раздел “Месторождения” содержит информацию о месторождениях нефти и газа Ханты-Мансийского региона, а также о составе и физико-химическим свойствах добываемой нефти.

Раздел “Настройки” предназначен для связывания базы данных с другими программными продуктами, созданными для анализа нормальных колебаний молекулы, распознавания ИК-спектров поглощения и расчета термодинамических функций или, например, позволяющими вести квантово-химические вычисления.

Таким образом, информационная база хранится в специально разработанных таблицах и структурирована по различным кластерам — экспериментальные данные, теоретические данные, литературные

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

[2] Московский государственный университет им. М. В. Ломоносова, химический факультет, кафедра физической химии, Ленинские 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; e-mail: saml@uriit.ru

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

Рис. 1. Титульная страница базы данных с основными разделами

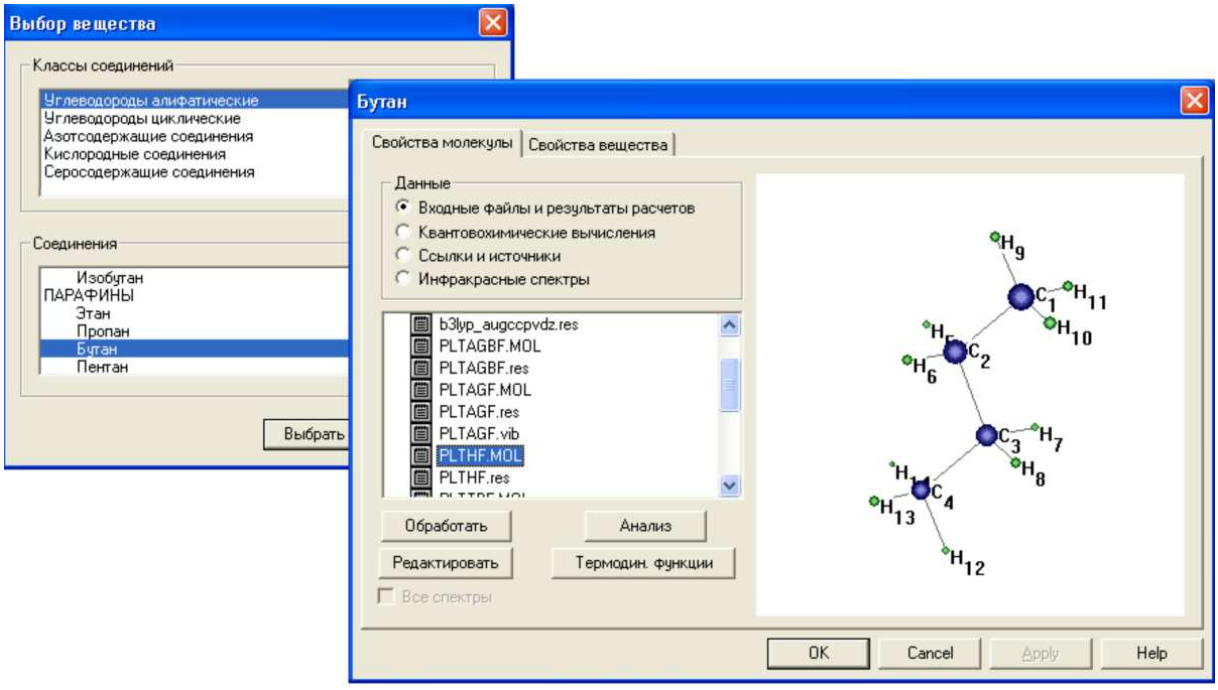

Рис. 2. Структурирование информации по свойствам молекулы и вещества

(справочные) данные о месторождениях, по строению, колебательным спектрам, силовым полям и другим молекулярным постоянным, а также по физико-химическим свойствам индивидуальных веществ.

Исследовательская (статическая) часть БД состоит из нескольких разделов, охватывающих литературные данные как общего характера, так и систематизированные по классам химических соединений. Пользователь может параллельно открывать любые документы, пополнять базу данных, переносить данные из одного раздела в другой и конвертировать форматы представления данных (рис. 2). Пользователь имеет также возможность, выбрав конкретное соединение, перейти к анализу всей наличной информации, классифицированной по молекулярным свойствам и свойствам вещества.

3. Представление информации для единиц <молекула> и <вещество>. После выбора одного из соединений можно перейти на один уровень вниз к каталогу данных для конкретного вещества. На рис. 2 представлен вид окна, к которому переходит пользователь при выборе конкретного вещества. В данном случае на экран выведена структура молекулы этана, получаемая с помощью специальной программы визуализации MOLGRAPH. В этом окне содержатся сведения о имеющихся входных и выходных файлах для расчетов частот колебаний и силовых полей (Prepared Data and Results), о квантово-механических данных (Quantum Chemistry Calculations), сведения об имеющихся публикациях (References and Sources) и инфракрасных спектрах поглощения (IR spectra). После выбора одного из типов данных происходит доступ для проверки, редактирования и вывода информации.

Таблицы для единиц <молекула> и <вещество> имеют следующий примерный вид (определяемый, в частности, наличием той или иной информации).

Единица <молекула>	1) брутто-формула
	2) пространственная структура
	3) геометрические параметры (экспериментальные или теоретические), включая декартовы и естественные координаты, графическое представление, ссылка на источник
	4) экспериментальные ИК-спектры поглощения с указанием условий регистрации, графическое представление, ссылка на источник
	5) экспериментальные ИК-спектры поглощения с указанием условий регистрации, представление в виде таблицы (положение максимумов с указанием интенсивностей), ссылка на источник
	6) квантово-химические данные по частотам колебаний (табличное (положение максимумов, абсолютные интенсивности) и графическое представление), ссылка на источник
	7) набор силовых постоянных в виде таблиц и матриц, ссылка на источник
	8) список литературы
Единица <вещество>	1) брутто-формула
	2) удельный вес
	3) температуры кипения и плавления
	4) термодинамические свойства (экспериментальные и расчетные)
	5) список литературы

Отличительной чертой предлагаемой базы данных является возможность работы с отдельными соединениями, выходящая за рамки систематизации информации, — теоретический анализ колебательных спектров, оценка термодинамических функций вещества, синтез силовых полей многоатомных молекул из отдельных фрагментов.

Вся информация о молекуле (карта молекулы) классифицируется на четыре группы:

— данные, подготовленные для решения прямых и обратных задач с помощью комплекса программ СПЕКТР [9] (файлы, содержащие декартовы координаты, список введенных естественных координат, матрицы силовых постоянных в декартовых или естественных координатах и результаты решения этих задач);

— входная информация для квантово-химических программных комплексов (типа Gamess, Gaussian

и др.) и результаты квантово-химических расчетов;

— табличная, текстовая и графическая справочная информация;

— информация о рассматриваемом соединении, содержащая ИК-спектры, список экспериментальных частот, структурные данные, физико-химические данные о веществе и т.д. и состоящая из документов разного рода и Web-ссылок; документы могут быть размещены как локально, так и на других интернет-серверах.

В настоящее время в БД включена информация для углеводородов различного строения, в частности, для цепочечных алканов C1 – C20, ряда непредельных углеводородов, производных бензола и ряда других соединений. Начаты работы по созданию модуля, позволяющего производить преобразования численных и графических данных в единый формат, предназначенный для использования в динамической части базы данных. Для цепочечных алканов, имеющих число атомов углерода до 10, выполнены квантово-механические расчеты, результаты которых также входят в соответствующие разделы базы данных.

4. Интеллектуальный блок. Информационный (интеллектуальный) блок, содержащий информацию общего характера, содержится в блоке “Документы/Documents”.

В этот раздел включена информация как чисто справочного, так и интеллектуально-обучающего характера. В частности, приведены сведения о составе и свойствах сырой нефти, методах ее переработки, химических и физико-химических свойствах различных классов соединений, входящих в состав сырой нефти, а также получаемых при ее переработке, и т.д.

Пополнение базы данных по соединениям и их фрагментам производится в процессе расчетов, выполняемых пользователем, а также по литературным данным. Этот процесс аналогичен пополнению базы данных по документам. На каждый документ создается информационная структура, включающая основную информацию. Эта же информация в дальнейшем используется при поиске нужного документа.

5. Блок информации о месторождениях нефти в ХМАО. Информация, касающаяся месторождений, располагающихся в Ханты-Мансийском регионе, содержит краткую характеристику залежей, физико-химическую информацию о составе нефти, газа и пластовых вод. При формировании этого блока данных нами были использованы элементы ГИС-технологий. При переходе из основного окна базы данных к разделу “Месторождения” появляется следующее окно, содержащее карту Ханты-Мансийского автономного округа. При переходе к позиции “Карта месторождений” появляется подробная карта месторождений (рис. 3).

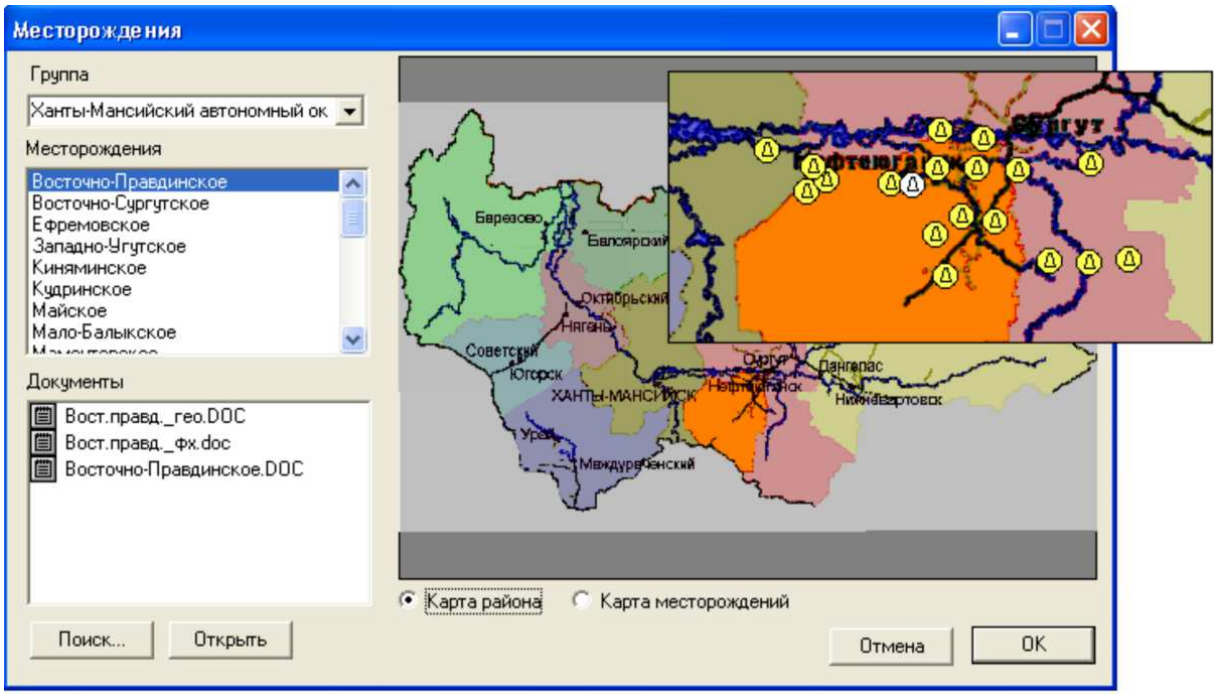

Рис. 3. Карта месторождений Ханты-Мансийского автономного округа

При подводе курсора к отдельному пункту на карте высвечивается название месторождения, а в левом нижнем окне появляется список доступных статей, каждая из которых содержит информацию по общим сведениям о месторождении и его геофизическим и физико-химическим характеристикам.

В настоящее время в систему включены данные по 25 месторождениям, расположенным в Нижневартовском, Ханты-Мансийском, Сургутском, Нефтеюганском районах Ханты-Мансийского автономного округа.

6. Динамический блок БД. Основным блоком динамической части данной информационно-вычислительной интегрированной системы является программный комплекс для теоретического анализа колебательных спектров многоатомных молекул "СПЕКТР" [2, 3], в котором реализованы оригинальные постановки обратных задач и численные методы их решения, основанные на методе регуляризации Тихонова. Применение этого метода связано с тем, что задачи определения параметров силового поля многоатомной молекулы по экспериментальным данным относятся к некорректно поставленным.

Программный пакет для интерпретации экспериментальных данных первоначально был разработан на языке ФОРТРАН для ЭВМ БЭСМ-6 [4 – 6]. В дальнейшем он был доработан для использования на персональных ЭВМ типа IBM PC и переписан на языке C++. Несмотря на то что основной вариант программного комплекса ориентирован на работу в среде семейства 32-разрядных операционных систем Windows, основной расчетный модуль (ядро комплекса) организован так, что может быть без изменений скомпилирован и использован в системах под управлением ОС UNIX.

Программный комплекс включает следующие компоненты.

1) Disp — комплекс программ для обработки спектроскопических данных. Он реализует алгоритм решения обратной спектральной задачи, в которой геометрическая конфигурация молекулы считается известной. Созданный в 1982 – 1990 гг., этот пакет был специально доработан для упрощения использования данных расчетов *ab initio* [7, 8], а также для возможности расчетов с представлением силовых матриц посредством масштабирующих множителей [9]. Возможны расчеты с использованием разнообразных экспериментальных данных, совместного расчета изотопных разновидностей или различных молекул и т.п.

2) ElDiff — комплекс программ для обработки данных газовой электронографии. Этот пакет построен на базе предыдущего и включает возможности одновременного решения структурной задачи и определения параметров силового поля. Пакет использует приближение малых колебаний, приемлемое для широкого круга молекул вследствие учета ангармоничности в первом и втором порядках теории возмущений.

3) Large — комплекс программ, аналогичный ElDiff, но имеющий возможность рассматривать движения с большой амплитудой, несводимые к ангармоническому осциллятору (внутреннее вращение и т.п.).

Для удобства пользователей также создан набор сервисных программ, облегчающих преобразование данных и их визуализацию. В их число входят следующие программы.

1) Symm — программа для определения типа симметрии молекулы по ее геометрической конфигурации и построения координат симметрии в декартовых или внутренних координатах. В настоящее время программа обрабатывает все точечные группы симметрии с осями до шестого порядка включительно, а также группы симметрии линейных молекул.

2) MolGraph — программа визуализации молекул. Наряду с отображением геометрической конфигурации, программа позволяет наглядно представить колебания молекул по данным, рассчитанным предыдущими перечисленными программами. Все иллюстрации молекул, приведенные в данной работе, получены с ее помощью.

3) GssRead — программа чтения выходных файлов квантово-механических программ, извлекающая из них информацию, необходимую для основных программ рассматриваемого пакета, и автоматически формирующую входной файл программ DISP, ELDIFF и LARGE. В настоящее время поддерживается чтение выходных файлов квантово-механических программ GAMESS, "Природа" (разработанной на химическом факультете МГУ) и ряда других квантово-химических комплексов.

Все перечисленные программы базируются на одном формате представления данных (как внутреннем, так и внешнем). Формат данных позволяет естественным образом включать в рассмотрение результаты квантово-механических расчетов. Программа-диспетчер позволяет запускать отдельные перечисленные модули, редактировать исходные файлы и просматривать результаты расчетов.

Наличие в БД геометрических параметров молекулы и регуляризованных силовых постоянных, относящихся к одному классу матриц силовых постоянных и обладающих свойством переносимости в рядах родственных соединений, позволяет с высокой степенью достоверности рассчитать теоретический колебательный спектр. Далее можно оценить методом статистической термодинамики стандартные термодинамические функции как известных химических соединений, так и соединений, которые по своим специфическим физико-химическим свойствам (неустойчивость при обычных температурных условиях, трудности синтеза и т.д.) недоступны для прямых спектральных исследований.

В базу данных включены программы, обеспечивающие визуализацию исходных данных и результатов расчетов (например, колебательного спектра как в табличном, так и в графическом видах).

Предусмотрены следующие возможности:

1) запуск квантово-химических программ непосредственно из оболочки базы данных; при этом как исходные файлы, так и файлы результатов автоматически заносятся в базу, что позволяет систематизировать хранение и поиск выполненных расчетов;

2) импорт данных из выходных файлов квантово-механических программ в единый формат, принятый для базы данных; это позволяет в дальнейшем организовать исходные структуры для запуска программ интерпретации экспериментальных данных и для процедуры идентификации веществ;

3) редактирование молекулярной информации: в частности, пользователь может модифицировать фрагменты исследуемой молекулы (например, замещая атомы на более или менее стандартные фрагменты) и использовать имеющиеся данные по силовым полям из библиотеки молекулярных фрагментов; в результате появляется возможность синтезировать спектры новых соединений.

СПИСОК ЛИТЕРАТУРЫ

1. *Кочиков И.В., Курамшина Г.М., Самков Л.М., Шарапов Д.А., Шарапова С.А.* База данных (информационная система) гибридного типа по молекулярным спектральным постоянным (ИСМОЛ) // Вычислительные методы и программирование. 2005. **6**. 83–87.
2. *Кочиков И.В., Курамшина Г.М., Пентин Ю.А., Ягола А.Г.* Обратные задачи колебательной спектроскопии. М.: Изд-во МГУ, 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. *Кочиков И.В., Курамшина Г.М.* Комплекс программ для расчета силовых полей многоатомных молекул по методу регуляризации Тихонова // Вестник Моск. ун-та. Сер. 2. Химия. 1985. **26**, № 4. 354–358.
5. *Курамшина Г.М., Черник С.И., Пентин Ю.А.* Диалоговая система для обработки колебательных спектров многоатомных молекул. М.: Изд-во МГУ, 1989.
6. *Кочиков И.В., Курамшина Г.М., Пентин Ю.А., Ягола А.Г.* Использование устойчивых численных методов для создания банков силовых постоянных многоатомных молекул // Тезисы докладов на VIII Всесоюзной конференции “Использование вычислительных машин в спектроскопии молекул и химических исследованиях”. Новосибирск, 1989. 70–71.
7. *Kuramshina G.M., Weinhold F.A., Kochikov I.V., Pentin Yu.A., Yagola A.G.* Joint treatment of ab initio and experimental data in molecular force field calculations with Tikhonov’s method of regularization // J. Chem. Phys. 1994. **100**, N 2. 1414–1424.
8. *Kochikov I.V., Kuramshina G.M., Sharapov D.A., Yagola S.A.* Database 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.
9. *Kochikov I.V., Kuramshina G.M., Stepanova A.V., Yagola A.G.* Numerical aspects of the calculation of scaling factors from experimental data // Вычислительные методы и программирование. 2004. **5**, № 2. 162–171.

Поступила в редакцию
26.09.2006