УДК 532.6

АЛГОРИТМЫ ИДЕНТИФИКАЦИИ ВЕЩЕСТВ ПО ИК-СПЕКТРАМ 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]

Описаны алгоритмы идентификации веществ, реализованные в информационной системе ИСМОЛ — базе данных гибридного типа по молекулярным постоянным соединений сырой нефти и продуктов ее переработки. Работа выполнена при финансовой поддержке РФФИ-ОБЬ (код проекта 05–07–98001).

Ключевые слова: идентификация химических веществ, спектральные постоянные, информационные системы, базы данных, молекулярные соединения.

1. Введение. Общая схема базы данных ИСМОЛ (информационной системы с элементами ГИС-технологий) была представлена ранее в работах [1, 2]. В архитектуре базы данных статическая справочная часть и динамическая часть сочетаются в виде интегрированной информационно-вычислительной системы, включающей в себя пакет программ СПЕКТР для решения прямых и обратных задач колебательной спектроскопии [3, 4].

База данных ИСМОЛ снабжена средствами систематизации поиска информации. Помимо поиска документов по интересующим соединениям, реализована процедура идентификации неизвестных спектров по спектрам, имеющимся в составе базы данных.

Процедура идентификации спектров реализована в базе данных ИСМОЛ на двух уровнях. Первый уровень направлен на поиск спектров, непосредственно сходных с заданным; второй уровень — на поиск комбинации (смеси) веществ, спектр которой наилучшим образом приближает заданный.

Основы отбора и распознавания спектров следуют принципам, изложенным в работах [5, 6], где аналогичные процедуры разработаны для спектров, получаемых с фурье-спектрометра. В связи с тем, что спектры базы данных определяются в лабораторных условиях и имеют, как правило, гораздо более высокое качество, нежели спектры, измеряемые в указанных работах, некоторые алгоритмы были изменены в сторону упрощения. Спектры базы данных хранятся в текстовом формате и представляют собой зависимости коэффициента поглощения (или пропускания) от частоты.

Алгоритмы поиска сходных спектров реализованы на базе двух основных подходов: корреляционного и структурно-логического. В первом случае основой классификации спектров является некоторая функция расстояния между спектрами, рассматриваемыми как функции на некотором интервале частот. Во втором случае для каждого из спектров базы данных и для исследуемого спектра строится система признаков (отражающих число и положение линий поглощения), а классификация строится на основе сходства этих признаков. В настоящей статье основное внимание уделяется корреляционному подходу.

2. Корреляционная процедура поиска сходных спектров. При использовании корреляционного подхода спектры базы данных классифицируются по степени сходства с заданным неизвестным спектром. В качестве "расстояния" между исследуемым спектром S и каждым из спектров базы данных S_i $(i = 1, 2, \ldots, N)$ используется функция

$$d_i = d(S, S_i) = \min_k \frac{\|kS - S_i\|}{\|S_i\|}, \tag{1}$$

где $S = S(\nu) = \log \dfrac{1}{\tau(\nu)}$ — оптическая плотность, ν — частота (в см^{-1}) и $\tau(\nu)$ — коэффициент поглощения (в базе данных спектры хранятся в виде зависимости коэффициента поглощения от частоты). Минимум

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Ленинские горы, 119991, Москва; e-mail: kochikov@tm-net.ru

[2] Московский государственный университет им. М. В. Ломоносова, химический факультет, кафедра физической химии, Ленинские горы, 119899, 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, 119899, Москва; e-mail: sharapov@srcc.msu.ru, sharapova@srcc.msu.ru

в (1) достигается при $k = \dfrac{(S, S_i)}{||S_i||^2}$, а скалярное произведение определено как

$$(S_i, S) = \int_{\nu_1}^{\nu_2} S_i(\nu) S(\nu) w(\nu)\, d\nu, \quad \|S\|^2 = (S, S), \tag{2}$$

где $w(\nu)$ — некоторая весовая функция. Диапазон частот (ν_1, ν_2) задается пользователем.

Выражение (1) определено так, чтобы расстояние между спектрами поглощения одного и того же вещества при разных концентрациях равнялось нулю. Максимальное значение этого выражения достигается при нулевом спектре S и равно 1. Вместо величин d_i удобно использовать величину $r_i = 1 - d_i$, которая имеет смысл, сходный с коэффициентом корреляции спектров.

Таким образом, определяется набор веществ, спектры которых имеют наибольшее сходство с исследуемым.

Описанная процедура применима для спектров достаточно высокого качества. Если же исследуемый спектр содержит заметные систематические погрешности (например, медленно меняющуюся подложку), можно непосредственно вычислять коэффициент корреляции спектров

$$r_i = \frac{\langle \tau_i - \overline{\tau}_i, \tau - \overline{\tau} \rangle}{\sqrt{\langle (\tau_i - \overline{\tau}_i)^2 \rangle \langle (\tau - \overline{\tau})^2 \rangle}}, \tag{3}$$

где τ — измеренный спектр пропускания, а τ_i — спектр пропускания i-го вещества из базы данных. Скалярные произведения вычисляются по формулам (2), однако вместо оптических плотностей в (2) используются коэффициенты пропускания τ. В этой процедуре спектры пропускания предварительно подвергаются фильтрации, подавляющей низкочастотную составляющую, — таким образом, коэффициент корреляции, вычисленный по формуле (3), отражает сходство в положении и величине полос поглощения. Данная процедура оказывается менее чувствительной, чем использующая формулу (1), но более устойчивой по отношению к систематическим погрешностям спектров.

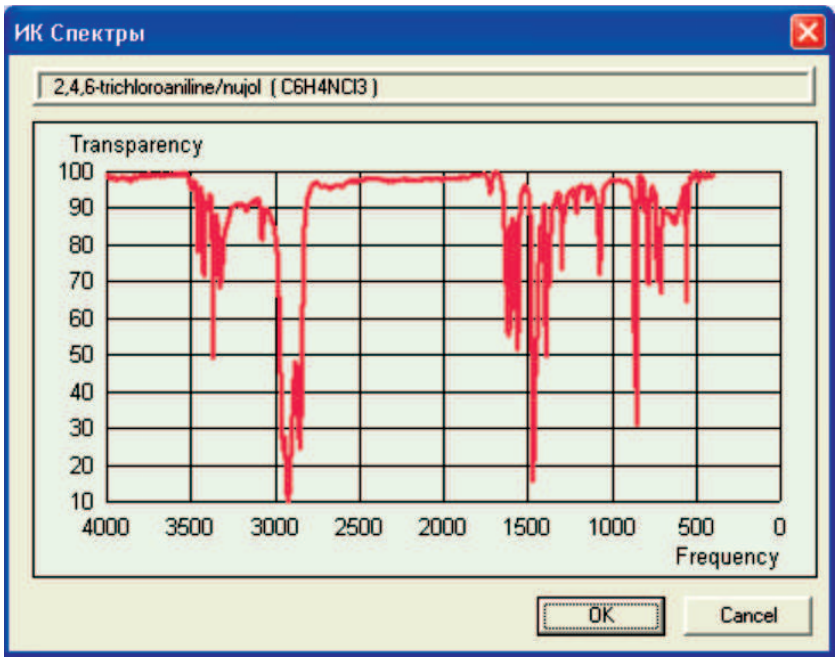

Рис. 1. Идентифицируемый спектр

На рис. 1 и 2 показан поиск сходных спектров для одного из соединений базы данных. Величины в списке соответствуют степени сходства спектров в процентах.

3. Процедура идентификации смеси веществ. Рассмотренная процедура оказывается недостаточной для идентификации спектра, соответствующего смеси различных соединений. Для определения состава неизвестного спектра в этом случае решается задача минимизации функционала

$$\Phi(c_1, ..., c_M) = \int_{\nu_1}^{\nu_2} \left[\sum_{i=1}^{M} \frac{c_i}{c_i^0} S_i(\nu) - S(\nu) \right]^2 w(\nu)\, d\nu, \tag{4}$$

где c_i — концентрации веществ, подлежащие определению, а c_i^0 — известные концентрации веществ, для которых получены спектры базы данных S_i. Индекс $i = 1, 2, \ldots, M$ нумерует вещества базы данных. Минимизация квадратичного функционала (4) ведется с учетом ограничений на неотрицательность коэффициентов c_i, причем ограничения могут быть наложены и сверху, так как при очень больших концентрациях коэффициенты c_i все равно не могут быть определены надежным образом. Фактически вместо интеграла в правой части (4) стоит сумма по числу точек спектра (обычно несколько сотен точек).

Для одного вещества задача минимизации (4) совпадает с алгоритмом, использующим формулу (1), поэтому формально для идентификации спектров можно было бы в любом случае сразу использовать метод идентификации смеси веществ. Тем не менее, на практике предварительный анализ спектров с помощью методов идентификации одного спектра оказывается необходимым элементом анализа. Практика расчетов показывает, что результаты распознавания становятся менее надежными, если алгоритм минимизации (4) применяется одновременно для многих веществ базы данных — при наличии большого количества спектров решение задачи становится неустойчивым. С целью повышения устойчивости процедуры необходимо провести предварительный отбор ограниченного набора спектров базы данных, в котором и осуществляется идентификация смеси.

Процедура отбора спектров имеет итерационный характер. Причина выбора такого подхода состоит в том, что при наличии нескольких веществ корреляции более слабых примесей могут быть сильно искажены основным веществом. Поэтому вначале определяется концентрация главного вещества (т.е. того, спектр которого наиболее сходен с исследуемым). Затем исходный спектр делится на спектр поглощения этого вещества (при найденной концентрации) и процедура повторяется: вновь находится наиболее похожий спектр и т.д.

В результате описанного итерационного процесса для количественного анализа может быть оставлено заранее заданное число веществ. Описанная процедура может привести и к меньшему набору веществ, если наложить требование, чтобы коэффициенты корреляции для них были положительны и превосходили некоторый заданный минимальный уровень. Практика расчетов показывает, что для надежного определения концентраций следует оставлять для анализа лишь 3 – 5 спектров.

При наличии в базе данных большего числа сходных спектров процедура минимизации функционала (4) повторяется для различных ограниченных наборов тестовых спектров. При сравнимых величинах минимумов функционала Ф может быть сделано лишь заключение о вероятном присутствии в исследуемом спектре веществ данного класса, но не каждого вещества индивидуально.

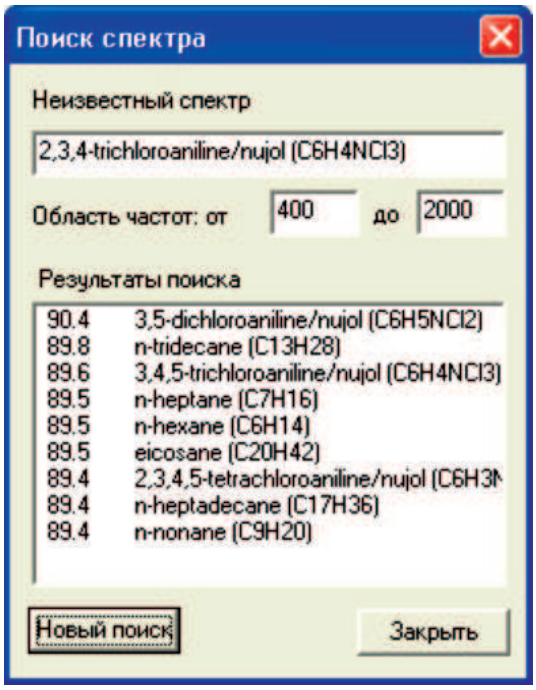

Рис. 2. Пример работы модуля идентификации спектров

4. Включение квантово-химических данных об интенсивностях ИК-спектров в процедуры идентификации веществ. Расширяющиеся возможности квантово-химических расчетов многоатомных молекул позволяют ставить задачу более широкого привлечения теоретических данных в процедуры идентификации веществ. Традиционно сопоставление экспериментальных и теоретических спектров, полученных для данной молекулярной системы на достаточно развитых уровнях теории, ограничивается сравнением величин экспериментальных и рассчитанных частот колебаний. Теоретические данные по интенсивностям ИК-полос поглощения используются для визуализации спектральных кривых и в достаточной степени пассивно используются в непосредственной практике спектрального эксперимента. Однако очевидно, что эти важные сведения также могут присутствовать в идентификационных алгоритмах в качестве активной, а не пассивной компоненты. В БД ИСМОЛ включена процедура, позволяющая восстанавливать теоретическую ИК-кривую по данным об абсолютных интенсивностях, получаемых при квантово-химических расчетах в приближениях гауссова или лоренцева контура, а также их комбинации. Таким образом, возможности идентификации веществ не ограничиваются списком имеющихся экспериментальных спектров.

При сравнении теоретических спектров с экспериментальными возникают проблемы, связанные с тем, что расчет интенсивностей полос поглощения теоретических спектров пока далеко не сопоставим по точности с экспериментом. Кроме того, весьма ограничены теоретически возможности описания формы линий и полос.

В этой связи эффективным оказывается применение структурно-логических методов распознавания. Для этого каждый спектр представляется набором признаков, которые в данном случае состоят в положении, ширине и интенсивности полос поглощения. Расстояние между спектрами (или функция их сходства)

задается в пространстве этих признаков, причем ожидаемая неточность задания ширины и интенсивности полос учитывается весовыми множителями. В результате поиск сходных спектров фактически сводится к поиску качественно сходных полос поглощения, расположенных в одном и том же частотном интервале, что дает возможность идентифицировать спектры, даже если непосредственно вычисленные корреляции оказываются не очень высокими.

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., Самков Л.М., Шарапов Д.А., Шарапова С.А.* Структурирование и формализация информации в базе данных гибридного типа по молекулярным спектральным постоянным (ИСМОЛ) // Вычислительные методы и программирование. 2006. **7**, № 2. 209–214.
3. *Кочиков И.В., Курамшина Г.М., Пентин Ю.А., Ягола А.Г.* Обратные задачи колебательной спектроскопии. М.: Изд-во Моск. ун-та, 1993.
4. *Yagola A.G., Kochikov I.V., Kuramshina G.M., Pentin Yu.A.* Inverse problems of vibrational spectroscopy. Utrecht: VSP, 1999.
5. *Бойко А.Ю., Григорьев А.А., Дворук С.К., Корниенко В.Н., Кочиков И.В., Лелъков М.В., Мацюк Г.В., Морозов А.Н., Павлов А.Ю., Светличный С.И., Табалин С.Е., Шишкин Г.В., Шлыгин П.Е.* Проблема идентификации и определения концентраций загрязняющих веществ с помощью фурье-спектрорадиометра // Вестник МГТУ им. Н.Э. Баумана. Естественные науки. 2004. № 1. 26–41.
6. *Дворук С.К., Корниенко В.Н., Кочиков И.В., Лелъков М.В., Морозов А.Н., Светличный С.И., Табалин С.Е.* Мониторинг загрязняющих веществ в атмосфере с помощью фурье-спектрорадиометра // Оптический журнал. 2004. **71**, № 5. 7–13.

Поступила в редакцию
15.10.2007