АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ

№ 4 1972

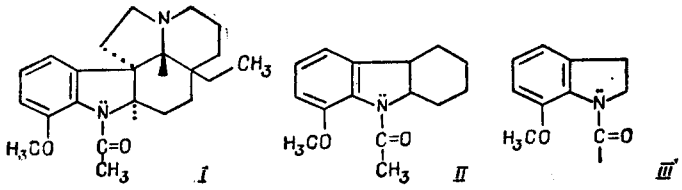
УДК 543.422.6 : 681.3

МАШИННАЯ ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА ДЛЯ ЭЛЕКТРОННОЙ СПЕКТРОСКОПИИ

Широкое внедрение молекулярной спектроскопии в практику химических исследований привело к накоплению столь больших объемов спектральной информации, что для ее эффективного использования становится необходимым создание специализированных информационно-логических систем на базе ЭЦВМ. В предыдущих работах [1, 2] мы сообщали о машинных системах опознания химических соединений по их ИК-спектрам поглощения. Эти системы позволяют быстро идентифицировать вещество, если оно уже было получено ранее и его ИК-спектр опубликован. Если же опознать вещество этим путем не удается, химик вынужден приступить к установлению его строения. При этом он широко использует различные методы спектроскопии молекул с целью получения информации о структурных фрагментах молекулы исследуемого соединения. Большую помощь в этой работе могут оказать машинные системы извлечения структурной информации из молекулярных спектров. В данном сообщении излагаются основы построения подобной системы для электронной спектроскопии, являющейся частью комплексной машинной системы для логической обработки информации по молекулярной спектроскопии, использующей ЦВМ БЭСМ-6.

Электронный спектр поглощения (ЭСП) характеризует поглощение веществом излучения в ультрафиолетовой (200—400 нм) и видимой (400—760 нм) областях и обычно представляется в виде зависимости $\lg \varepsilon = \varphi(\lambda)$, где ε — молярный коэффициент погашения, а λ — длина волны. Поглощение излучения в указанном спектральном диапазоне обусловлено возбуждением электронов кратных связей (π-электронов) и неподеленных электронных пар (n-электронов). Если в молекуле имеется несколько таких фрагментов, причем они непосредственно связаны друг с другом $(A=B)_n-C\overset{\cdot\cdot}{-}(D\equiv E)_m$, что обеспечивает перекрывание орбиталей π- и n-электронов (сопряжение фрагментов), то полосы поглощения смещаются в длинноволновую область и одновременно увеличивается их интенсивность. Подобную комбинацию кратных (двойных и тройных) связей и атомов с неподеленными парами электронов (или вакантными орбиталями) обычно называют хромофорной системой. При структурных исследованиях широко используется то обстоятельство, что соединения, молекулы которых содержат одинаковые хромофорные сис-

темы, имеют очень близкие ЭСП. В качестве примера на рис. 1 приведены ЭСП алкалоида аспидоспермина (*I*) [3] и N-ацетил-8-метоксигексагидрокарбазола (*II*) [3], имеющих одинаковую хромофорную систему (*III*)



Если в молекуле имеется несколько независимых (разделенных двумя или большим числом ординарных связей) хромофорных систем, то наблюдаемый ЭСП представляет собой сумму спектров отдельных хромофоров, т. е. $\lg\varepsilon_{\lambda_i} = \lg\left(\varepsilon^{\mathrm{I}}_{\lambda_i} + \varepsilon^{\mathrm{II}}_{\lambda_i} + \ldots\right)$.

Указанные особенности ЭСП позволяют при наличии каких-то предположений о строении молекулы исследуемого соединения проверить эти предположения путем сопоставления его ЭСП со спектрами более простых модельных соединений.

Работая над общим алгоритмом решения задачи установления строения вещества, мы пришли к выводу о целесообразности привлечения электронной спектроскопии на более ранней стадии исследования на стадии формирования предположений о возможной структуре исследуемого вещества. Это может быть в принципе сделано следующим образом: из имеющихся коллекций ЭСП (см., например, [4]), охватывающих около 40 тыс. соединений, отбираются спектры, наиболее близкие спектру исследуемого вещества, и хромофорные системы соединений, соответствующих отобранным спектрам, рассматриваются в качестве возможных структурных блоков молекулы исследуемого соединения. Чтобы сделать подобную выборку реально осуществимой, можно было бы пойти по пути составления указателей [5], в которых ЭСП охарактеризованы положением и интенсивностью полос поглощения и определенным образом упорядочены. Однако такие указатели имеют ряд серьезных недостатков, наиболее важным из которых является невозможность их дополнения непрерывно поступающими новыми данными. Сейчас уже довольно очевидно, что кардинальное решение вопроса возможно лишь на базе использования ЭЦВМ.

Выбранный нами общий принцип построения специализированных систем для различных видов спектроскопии молекул заключается в следующем. Основным информационным фондом системы служит коллекция соответствующих спектров (атласы, каталоги и картотеки). Наиболее существенная часть спектральной информации о веществе, а также некоторые сведения о его строении и физических свойствах находятся в закодированном виде в долговременном запоминающем устройстве ЭЦВМ и образуют машинный каталог. При введении запроса ЭЦВМ отбирает в машинном каталоге соединения, близкие исследуемому по заданному набору признаков и удовлетворяющие выбранному критерию близости.

В машинный каталог рассматриваемой системы были введены следующие сведения: название соединения, его номер по каталогу основного фонда, брутто-формула, молекулярный вес, температуры плавления и кипения (если они указаны в основном каталоге), код хромофорной группы и описание ЭСП. Все данные, относящиеся к одному соединению, записаны подряд на магнитную ленту и образуют своеобразный реферат,

характеризующий соединение. В машинном каталоге рефераты располагаются подряд по мере поступления материала. В качестве элемента записи принято шестисимвольное машинное слово. Поскольку количество слов в реферате не является постоянным, отдельные характеристики соединения могут оказаться в реферате на разных местах. Поэтому для их выделения были использованы специальные метки (разделители).

Можно предложить различные способы описания ЭСП. Одним из эффективных приемов является, в частности, представление спектра в виде суммы некоторых специальных функций, например, гауссиан [6]. Однако при ручном кодировании большого числа спектральных кривых этот способ не пригоден из-за своей трудоемкости. Кодирование спектра равномерными отсчетами (указание $\lg \varepsilon$ при различных λ, изменяемых с постоянным шагом $\Delta\lambda$) при ограниченном числе отсчетов сопряжено с риском потери экстремальных точек. Выбор же большого числа отсчетов ведет к чрезмерному увеличению объема реферата. На наш взгляд, достаточно простым и в то же время весьма эффективным способом кодирования ЭСП является представление спектральной кривой узловыми точками линейно ломаной линии, аппроксимирующей спектральную кривую с выбранной степенью точности[1], причем ломаная линия строится таким образом, чтобы все характерные точки спектральной кривой (максимумы, минимумы и перегибы) оказывались узловыми точками (рис. 2). При

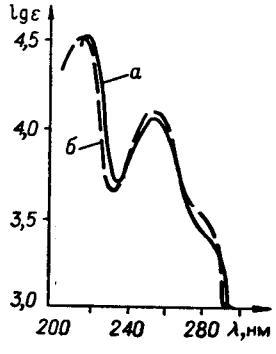


Рис. 1. Электронные спектры поглощения [3]: *а* — аспидоспермина (*I*); *б* — N-ацетил-8-метоксигексагидрокарбазола (*II*).

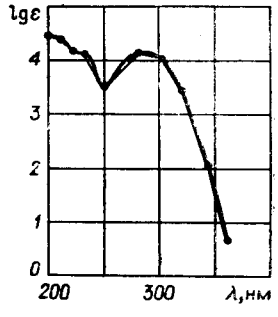


Рис. 2. Аппроксимация электронного спектра поглощения ломаной линией.

этом способе кодирования каждый спектр в машинном каталоге представляется таблицей чисел, характеризующей координаты выбранных узловых точек. Восстановленный по этой таблице спектр описывается выражениями:

$$y = y_i + k_i (\lambda - \lambda_i); \tag{1}$$

$$k_i = \frac{y_{i+1} - y_i}{\lambda_{i+1} - \lambda_i}, \tag{2}$$

где $y \equiv \lg \varepsilon$ (при длине волны λ, причем $\lambda_i \leqslant \lambda \leqslant \lambda_{i+1}$); i — номер узла.

Поскольку при выбранном способе кодирования спектры в машинном каталоге представлены неравномерными отсчетами (шаг $\Delta\lambda = \lambda_{i+1} - \lambda_i$ не является определенным), сравнение спектра исследуемого соединения с данными машинного каталога требует обязательного вос-

[1] Было принято, что отклонение по λ не должно превышать 2 нм, а по $\lg \varepsilon$ — 0,05.

становления закодированных спектров. Для этого, согласно (1) и (2), необходимо выполнить операции умножения и деления, которые для цифровых ЭВМ с плавающей запятой не являются элементарными. В описываемой системе для сокращения времени поиска осуществлена замена арифметических операций логическими. Это достигнуто путем введения в реферат величин k_i, вычисленных при первоначальном вводе информации в машину. Поскольку для некоторых алгоритмов обработки информации по ЭСП необходимы величины ε_i, они также были вычислены при вводе данных в ЭВМ и включены в реферат.

Таким образом, каждый отсчет спектральной кривой в реферате охарактеризован набором величин λ_i, y_i, ε_i и k_i. Кроме того, отсчет наиболее интенсивного максимума поглощения повторяется еще раз в определенном месте реферата. Выбранный вариант кодирования и записи информации позволяет хранить на одной магнитной ленте ЦВМ БЭСМ-6 сведения приблизительно о 7 тыс. соединений.

Программа поиска построена по блочному принципу и представляет собой набор отдельных подпрограмм, каждая из которых осуществляет отбор соединений, по одному из возможных типов признаков. Имеющийся набор подпрограмм позволяет вести поиск по следующим признакам:

А) номерам соединений в каталогах основного фонда;

Б) положению максимума и величине $\lg \varepsilon$ для наиболее интенсивной полосы поглощения (указывается координатный прямоугольник, в который должен попадать этот максимум);

В) положению максимумов и минимумов спектральной кривой;

Г) спектральной кривой, задаваемой набором координатных прямоугольников, через которые эта кривая должна проходить;

Д) коду хромофорной группы;

Е) наличию или отсутствию в брутто-формуле атомов определенных элементов;

Ж) молекулярному весу в заданном интервале $M \pm \Delta M$;

З) температурам плавления и кипения[1] в заданном интервале $T \pm \Delta T$.

Подпрограмма В позволяет проводить отбор соединений по наличию в ЭСП максимумов и минимумов в указанных интервалах длин волн без наложения ограничений на величины $\lg \varepsilon$, что может оказаться полезным в тех случаях, когда молекулярный вес исследуемого соединения неизвестен и величины ε не могут быть вычислены. При использовании подпрограммы Г отбираются соединения, спектральная кривая которых проходит через заданные координатные квадраты, число, положение и размеры которых выбираются таким образом, чтобы обеспечить необходимую близость отобранных спектральных кривых спектру исследуемого соединения (примеры см. ниже).

При обращении к системе абонент составляет запрос (рис. 3), в котором указывается номер запроса, шифр требуемой подпрограммы, критерии для отбора с помощью этой подпрограммы, разделитель (OOOEND), шифр следующей подпрограммы и т. д. В конце каждого запроса ставится символ «КОНЕЦ*».

Введенный в машину запрос поступает в блок анализатора, который формирует программу по-

```
Б NO1398
Б МАКСИМ
Б 225240
Б 410430
Б OOOEND
Б ЭКСТРА
Б 330345
Б OOOEND
Б ЭКСТРУ
Б 265275
Б OOOEND
Б ТЕМПЛА
Б 00+080
Б 00+090
Б OOOEND
Б КОНЕЦ*
```

Рис. 3. Запрос машине для поиска вещества по его электронному спектру поглощения.

[1] Если этот признак в каталоге не указан, то соединение считается удовлетворяющим заданным требованиям.

иска в соответствии с указанными в запросе подпрограммами. После проведения поиска абоненту выдаются номера соединений (по каталогу ЭСП основного фонда), удовлетворяющих заданным признакам, код их хромофорных групп и названия в латинской транскрипции (в пределах 36 символов). Предусмотрена также возможность указания номеров отобранных соединений по каталогу ИК-спектров. Кроме того, в машинной выдаче приводится некоторая служебная информация — в начале выдачи воспроизводится содержание запроса, а в конце указывается общее число просмотренных рефератов и количество соединений, отобранных каждой из использованных подпрограмм. Пример машинной выдачи приведен на рис. 4.

ЗАПРОС: 000549

ОТСЧЕТ	200210	400422	213226
440468	228242	368391	244256
375400	265280	215240	325340
345380	375390	190215	OOOEND
КОНЕЦ*			

КОД ХР. ГРУППЫ	N ПО УФ-КАТАЛОГУ	N ПО ИК-КАТАЛОГУ	НАИМЕНОВАНИЕ
0000D9	00L549	000000	ANTHRANILIC ACID
0000D9	00L550	000000	ETHYL ANTHRANILATE
000H22	0L1049	000000	N — METHYLISATOIC ANHYDRIDE

ВСЕГО СПЕКТРОВ 004742
СОДЕРЖИМОЕ СЧЕТЧИКОВ СП:

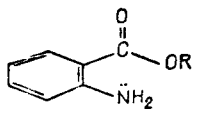
1*	2*	3*	4*	5*	6*	7*
000000	000000	000000	000000	000000	000000	000003
8*	9*	10*	11*	12*	13*	14*
000000	000000	000000	000000	000000	000000	000000

15*	16*	17*	18*	19*	20*
000000	000000	000000	000000	000000	000000

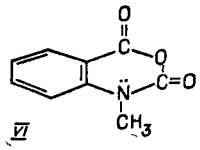
Рис. 4. Ответ машины на запрос, составленный по спектру атраниловой кислоты (*IV*).

Машинный каталог описываемой информационно-поисковой системы для электронной спектроскопии содержит в настоящее время информацию приблизительно о 5 тыс. соединений, взятую из атласов [4а] и [4б]. Для проверки эффективности работы системы и нахождения оптимальных вариантов составления запросов было обработано около 120 запросов по ЭСП соединений, которые заведомо присутствуют в указанных атласах. Полученные результаты продемонстрировали правильность выбора принципов построения машинной системы для обработки больших массивов информации по электронной спектроскопии. В качестве иллюстрации этого приведем несколько примеров использования подпрограмм Б, В и Г для решения вопроса о возможном строении хромофорной системы исследуемого соединения.

В одном из примеров в качестве исследуемого соединения была выбрана антраниловая кислота (*IV*). Набор координатных прямоугольников, которыми был задан ее ЭСП при обращении к машине, показан на рис. 5. Среди отобранных машиной соединений (см. рис. 4), кроме кислоты (*IV*), присутствовал ее этиловый эфир (*V*) и соединение (*VI*), имеющее родственную хромофорную систему (их ЭСП приведены на рис. 5):

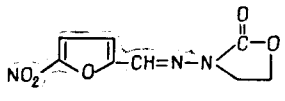


IV R = H;
V R = C_2H_5

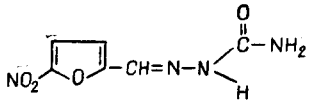


VI

В другом примере при поиске по спектру N- (5-нитрофурфурилиден-2)-3-аминооксазолидона-2 (*VII*) машиной, кроме этого соединения, было выдано соединение (*VIII*), имеющее очень близкую хромофорную систему и практически тождественный спектр поглощения (ср. каталог [4б] № 1757 и № 599):

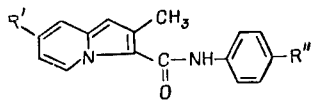


VII



VIII

Аналогично при использовании в качестве неизвестного соединения 4-метоксианилида 2-метилиндолизин-3-карбоновой кислоты (*IX*) были найдены родственные структуры (*X*) и (*XI*):



IX R′ = H, R″ = OCH_3
X R′ = CH_3, R″ = OCH_3
XI R′ = H, R″ = Cl

При поиске по спектру 3-метил-5,5-дифенилдитиогидантоина (*XII*), заданному четырьмя координатными прямоугольниками и интервалом

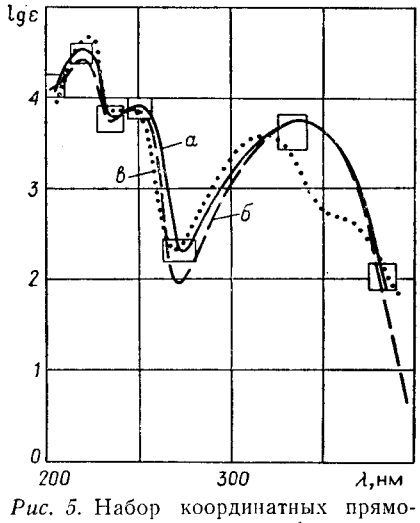


Рис. 5. Набор координатных прямоугольников, которыми был задан спектр антраниловой кислоты при обращении к машине, и спектры, отобранные машиной:
а — антраниловой кислоты (*IV*); *б* — этилового эфира антраниловой кислоты (*V*), *в* — соединения (*VI*).

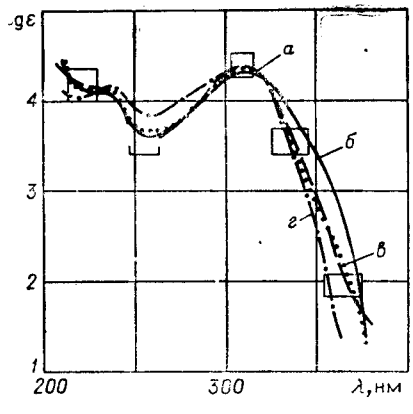
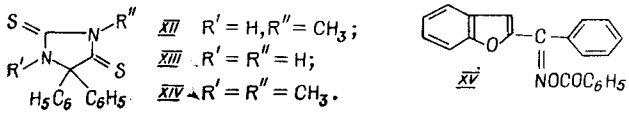


Рис. 6. Электронные спектры поглощения, отобранные машиной при поиске по спектру 3-метил-5,5-дифенилтиогидантоина (*XII*):
а — спектр соединения *XII*; *б* — спектр соединения *XIII*; *в* — спектр соединения *XIV*; *г* — спектр соединения *XV*. Отмеченные интервалы длин волн и lg ε соответствуют указанным в поисковом запросе.

положения одного из минимумов (рис. 6), кроме соединения (*XII*), машиной были отобраны два его близких аналога (*XIII*) и (*XIV*), а также соединение (*XV*), имеющее иной тип хромофорной системы, но близкий к ЭСП (см. рис. 6):



В этом случае исследователю необходимо было бы сделать выбор между двумя хромофорными системами (например, по наличию или отсутствию серы в изучаемом соединении).

Приведенные примеры показывают, что описываемая информационно-поисковая система для электронной спектроскопии может оказать существенную помощь при структурных исследованиях, поскольку она позволяет идентифицировать большие структурные блоки молекулы. Система позволяет вести поиск одновременно по восьми запросам. Программа поиска хранится на магнитной ленте и вводится специальной программой вызова. Время от ввода запроса до выдачи ответа на печать не превышает 5 мин. Время же работы центрального процессора составляет приблизительно 15 с в расчете на один запрос. Кроме основных, разработаны вспомогательные программы предварительной проверки вводимого материала, стирания и печати, хранящихся на магнитной ленте рефератов и др.

ЛИТЕРАТУРА

1. Ю. П. Дробышев, Р. С. Нигматуллин, В. И. Лобанов, И. К. Коробейничева, В. С. Бочкарев, В. А. Коптюг. Использование ЭВМ для опознания химических соединений по их спектральным характеристикам.— Вестник АН СССР, 1970, № 8.
2. Ю. П. Дробышев, Р. С. Нигматуллин, В. И. Лобанов, И. К. Коробейничева, В. С. Бочкарев, В. А. Коптюг. Машинная система поиска ИК-спектров по каталогу Садтлера.— Изв. СО АН СССР, серия хим., 1972, № 2.
3. J. R. Chalmers, H. T. Openshaw, G. F. Smith. The Constitution of Aspidospermine. Part II. Ultraviolet Absorption the Bz — Methoxy — tetra — and — hexa — hydrocarbasoles.— J. Chem. Soc., 1957, 1115.
4. а) The UV Atlas of Organic Compounds (DMS), v. I — V. Verlag Chemie, Weinheim; Butterworths, London;
   б) Absorption Spectra in the Ultraviolet and Visible Region, ed. L. Lang, v. I—XV, Budapest;
   в) Sadtler Standard Ultraviolet Spectra, The Sadtler Research Laboratories, Philadelphia;
   г) Catalog of Selected Ultraviolet Spectral Data, American Petroleum Institute Research Project 44.
5. P. Sadtler. UV Locator. The Paper presented at the Pittsburgh Analytical Meetings, 1966.
6. Спектроскопические методы в химии комплексных соединений. Под ред. В. М. Вдовенко. М., «Химия», 1964.

Поступила в редакцию
11 января 1972 г.