

УДК: 543.51+543.42:681.32

Анализ органических веществ с использованием базы данных "масс-спектр – фрагментный состав соединения"

Б. Г. ДЕРЕНДЯЕВ, В. Н. ПИОТТУХ-ПЕЛЕЦКИЙ, К. С. ЧМУТИНА, С. А. НЕХОРОШЕВ

Новосибирский институт органической химии имени Н. Н. Ворожцова Сибирского отделения РАН, проспект Академика Лаврентьева, 9, Новосибирск 630090 (Россия)

E-mail: der@nioch.nsc.ru

(Поступила 05.04.2001)

Аннотация

Показано, что спектральный поиск в базе данных "масс-спектр – полный фрагментный состав соединения" обеспечивает распознавание разнообразных фрагментов изучаемого соединения. На примерах анализа масс-спектров более 13 000 соединений выявлены типы распознаваемых фрагментов и вероятности их идентификации. Установлены зависимости, связывающие доли корректно и ошибочно распознаваемой структурной информации.

ВВЕДЕНИЕ

Современная масс- и хромато-масс-спектрометрия в сочетании с информационно-поисковыми системами (ИПС), содержащими крупные базы данных (БД), стала мощным средством идентификации соединений и компонентов сложных смесей природного или антропогенного происхождения. Базы данных используются для этой цели, если анализируемый спектр содержится в них. Этот метод стал рутинным в аналитической практике.

Решение более сложной задачи массового анализа – установления строения соединения, не представленного своим спектром в БД, – требует новых приемов и средств, несмотря на определенные успехи в этой области [1]. В случае идентификации соединения достаточно найти в БД спектр, совпадающий по массовым числам и интенсивностям с исследуемым. При использовании БД для установления строения нового соединения ситуация усложняется. В этом случае разрабатывают специализированные системы и алгорит-

мы поиска. Они учитывают поведение спектров не только в абсолютной, но и в относительной шкале масс, возможные сдвиги массовых чисел (m/z) пиков ионов, обусловленные влиянием заместителей или функциональных групп, информативность различных по величинам m/z и интенсивностям пиков и т.п. Примеры таких разработок – системы STIRS [2], SISCOM [3], КОМПАС-МС [4] и др., например [5]. Эти системы характеризуют оригинальные алгоритмы отбора спектров из БД в поисковые ответы и, что не менее важно, средства манипулирования структурными данными. Анализ структурных формул, отбираемых из БД (по предъявленному спектру), и определение их общих частей позволяют во многих случаях выносить суждения об особенностях строения исследуемого соединения. В ряде работ показана возможность опознания на основе этого анализа некоторых заданных групп атомов [6–8] или крупных связанных структурных единиц исследуемых соединений [9–12] путем выявления максимальных общих подграфов отбираемых из БД соединений.

Очевидно, что эффективность соответствующего программного обеспечения в значительной степени определяется методами представления структурных данных и манипулирования ими. Именно поэтому наряду с традиционным способом представления структурных формул соединений (далее для краткости – структур) в виде молекулярных графов используют и другие приемы описания структур или фрагментов молекул, например HOSE коды [13], древовидные коды [14], описание структур в виде заданного набора фрагментов [15, 16].

В работах [15–17] для случая описания структур соединений в виде исчерпывающего набора неизоморфных связанных фрагментов показана возможность распознавания с помощью БД по ИК-спектроскопии самых разнообразных фрагментов изучаемых соединений с числом вершин от двух до семи. Выказано предположение [18] о возможном распространении этого приема на анализ данных масс-спектрометрии. В то же время представленные в [19] результаты можно рассматривать как предварительные и, вероятно, излишне оптимистичные. Это обусловлено особенностью формирования и ограниченностью использованной для апробирования подхода выборки спектроструктурных данных. Более строгую и полную картину, характеризующую перспективы описания структур в виде исчерпывающего набора неизоморфных связанных

фрагментов в масс-спектрометрических информационных системах, могут дать эксперименты, выполненные на существенно большей по объему выборке исследуемых спектров соединений различных химических классов. Результаты таких экспериментов представлены в данной работе.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Общая схема поставленных экспериментов представлена на рис. 1. Опишем кратко ее основные элементы.

БД спектров. В экспериментах использована часть БД масс-спектров NIST/EPA [6], содержащая 24 000 спектров различных соединений, состоящих из атомов C, H, O, N, P, S, Si, F, Cl, Br, с молекулярными массами в диапазоне от 40 до 264 а.е.м. Отбор спектральных признаков для сопоставления спектров проводился аналогично [4].

БД структур содержит поатомные коды молекулярных графов всех соединений, спектры которых включены в БД спектров. Каждое соединение представлено в базах данных спектров и структур только один раз.

БД фрагментов. Эта база сформирована путем последовательной декомпозиции всех молекулярных графов БД структур на все входящие в их состав k -вершинные связанные фрагменты ($k = 2-7$) [20]. Под вершинами

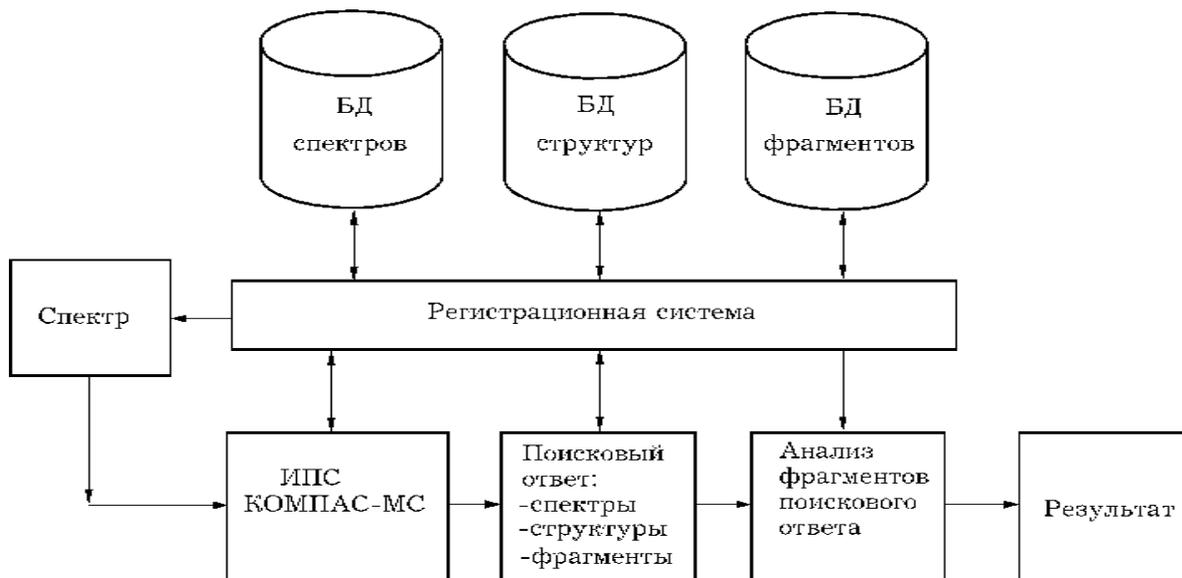


Рис. 1. Общая схема эксперимента.

ТАБЛИЦА 1

Число неизоморфных k -вершинных фрагментов в БД фрагментов

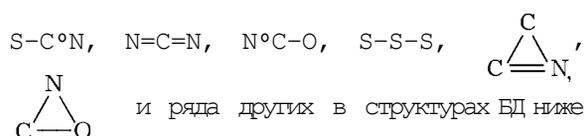
k	Общее число фрагментов	Частота встречаемости f^3					
		2	5	10	50	100	500
2	91	67	62	53	41	37	20
3	441	318	280	224	135	100	48
4	1719	1132	937	700	373	259	104
5	5898	3574	2874	2084	936	593	188
6	18485	10341	7977	5472	1913	1099	244
7	51054	25472	18331	11424	3120	1608	260

здесь и далее понимаются все атомы, кроме атомов водорода. Примеры полной картины формируемых при декомпозиции графов наборов k -вершинных фрагментов можно найти в работах [16, 18]. Для 24 000 структур зарегистрированы и включены в состав этой БД 77 688 различных фрагментов с числом вершин от двух до семи.

В табл. 1 для $k = 2, 3, \dots, 7$ приведены общее число неизоморфных k -вершинных фрагментов в составах всех структур БД, а также частота встречаемости фрагментов в структурах БД выше заданного ($f^3 2, 5, \dots, 500$) порога. Как видно, для всех значений k в БД представлено около половины фрагментов с $f^3 2$. Например, из 51 054 фрагментов, содержащих 7 связных вершин, 25 472 фрагмента встретились в различных структурах два раза и более, 11 424 – десять раз и более. Отметим, что в структурах соединений базы данных по ИК-спектроскопии выявлено в ~1.5 раза большее разнообразие фрагментов (ср. с [16]). Очевидно, это свидетельствует о меньшем разнообразии структур, представленных в масс-спектрометрической БД. Напомним, что спектры первичной базы масс-спектрометрических данных записаны в условиях ионизации молекул пучком электронов.

В данном случае 2- и 3-вершинные фрагменты вида $X \sim Y$ и $X \sim Y \sim Z$ содержат вершины с типичными для органических соединений атомами ($X, Y, Z = C, O, N, S, Hal, Si, P$). Эти фрагменты, как правило, представлены статистически значимым числом структур БД. Их комбинации определяют состав и более крупных фрагментов.

Знание списков редко (или, наоборот, часто) представленных в БД фрагментов важно с практической точки зрения. Оно позволяет на начальном этапе постановки задачи идентификации по анализируемому спектру того или иного фрагмента оценить потенциал используемой для ее решения БД. Например, частота встречаемости фрагментов вида



и ряда других в структурах БД ниже десяти, что, разумеется, влияет на возможность их распознавания с помощью базы данных. Соединения используемой части БД характеризует усредненная брутто-формула $C_{9.67}H_{13.45}O_{1.45}N_{0.79}S_{0.15}Cl_{0.15}F_{0.14}Br_{0.03}Si_{0.04}P_{0.01}$.

Регистрационная система – элемент схемы, позволяющий по индивидуальным регистрационным номерам спектров, структур и фрагментов быстро отыскать:

- спектр, соответствующий данной структуре;
- фрагменты, описывающие некоторую заданную структуру;
- структуры, обладающие определенным фрагментом или набором фрагментов;
- спектры соединений (структур), содержащих заданные фрагменты.

В этой же компоненте схемы хранится информация о частотах встречаемости зарегистрированных при декомпозиции структур фрагментов.

ИПС КОМПАС-МС – ранее описанная [4] информационная система по масс-спектрометрии молекул.

Поисковый ответ. В результате сопоставления предъявленного для анализа спектра со спектрами базы данных ИПС КОМПАС-МС отыскивает в БД спектры, наиболее похожие на заданный в запросе на поиск. Поисковый ответ (ПО) включает спектры, структуры отобранных соединений и список неизоморфных k -вершинных фрагментов, входящих в состав соответствующих структур. Каждый фрагмент этого списка сопровождается информацией о частоте его встречаемости во фрагментных составах структур ПО.

Тестовая выборка и условия экспериментов. Тестовую выборку составили ~13 тыс. записей спектр-структура из используемой базы данных. Структурные формулы соединений этой выборки содержат не менее восьми связанных вершин, молекулярные массы лежат в диапазоне от 126 до 200 а.е.м., а их усредненная брутто-формула имеет вид $C_{8.95}H_{13.02}O_{1.28}N_{0.69}S_{0.13}Cl_{0.12}F_{0.08}Br_{0.03}Si_{0.02}P_{0.01}$.

Учитывая, что объем используемой в экспериментах БД существенно меньше объема известных коммерческих БД, при формировании тестовой выборки мы включали в нее спектры соединений-эталонов с молекулярными массами, наиболее широко представленными в БД. Предполагается, что достигаемый в этом случае результат может быть экстраполирован на другие, более представительные базы данных.

Каждый включенный в выборку спектр предъявляли (см. рис. 1) в качестве запроса на поиск с целью отбора из БД одиннадцати спектров, наиболее похожих на заданный. Далее информацию о заданном соединении исключали из ПО и анализировали десять остающихся в нем записей. Этим моделировали условия "новизны" соединения и его спектра для используемой БД. При анализе результатов поиска структуру и фрагментный состав заданного соединения использовали как эталонные. Поиск спектров, подобных заданному, проводили в двух режимах, соответствующих отбору из БД спектров в абсолютной (поиск А) и относительной (шкала "первичных" потеря, поиск В) шкалах масс [4]. В поисковые ответы включали спектры, удовлетворяющие условию $MF^3 \leq 35$, где MF – соответствующая мера близости спектров [4].

По регистрационным номерам спектров, включаемых в поисковый ответ, из БД структур и БД фрагментов отбирали структуры и фрагментные составы соответствующих соединений. В ходе анализа ПО проводили контроль как за корректно распознаваемыми фрагментами эталонов, так и за всеми фрагментами, распознаваемыми в соответствующих экспериментах ошибочно.

Все эксперименты проведены на ЭВМ "Пен-тиум-266".

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Взаимосвязь структурного и спектрального подобия

Масс-спектры органических соединений настолько индивидуальны, что ИПС стали общепринятым инструментом идентификации веществ, спектры которых представлены в БД. Использование ИПС для анализа новых объектов основывают на предположении, что по предъявленному спектру из БД отбираются спектры соединений, подобных по строению изучаемому. Это лежит в основе всех известных методов компьютерного анализа структур соединений ПО с целью извлечения информации о строении молекул неизвестного вещества. Наряду с этим, однако, высказывалось мнение, что применение БД по масс-спектрам при анализе новых соединений ограничено [1, 10]. Поэтому, прежде чем детально анализировать результативность применения ИПС при решении поставленной задачи, необходимо убедиться, что в рамках выбранных приемов описания и сопоставления структур и масс-спектров предположение о подобии спектров похожих по строению соединений распространяется на самые разнообразные соединения.

Возможность отбора из БД соединений, подобных исследуемому, подтвердим, используя статистически значимую выборку спектров и структур "отсутствующих" в БД соединений (эталонов) в следующем эксперименте. По спектру каждого эталона отберем из БД десять наиболее похожих спектров соответствующих соединений. Сравним структуру-эталон с отобранными на основе подобию спектров структурами с целью опреде-

ления меры подобия. Наконец, по аналогии с [21] обобщим результаты проведенного анализа для всех сопоставляемых спектров и структур.

В качестве эталонов в данном случае выступают последовательно все 24 тыс. записей БД. Условие "новизны" эталона при каждом поиске моделируем простым изъятием его из соответствующего ПО.

Подобие отобранных в ПО спектров оценим параметром MF (в режиме поиск А, см. экспериментальную часть). Подобие структур-эталона отобранным структурам определим по близости фрагментных составов сравниваемых структур [18, 21]:

$$W = 100 \times 2S / (F_x + F) \quad (1)$$

где F_x и F - общее число неизоморфных k -вершинных фрагментов структуры эталона и сравниваемой с ней структуры соответственно ($k = 2, \dots, 7$); S - число совпавших k -вершинных фрагментов.

На рис. 2 приведена усредненная зависимость меры подобия отбираемых из БД структур от меры спектрального подобия для ~ 240 тыс. рассматриваемых в эксперименте

случаев (в каждом отдельном эксперименте эталон сравнивается с десятью первыми спектрами и структурами ПО). Характер выявленной зависимости показывает, что для выбранного алгоритма поиска спектров и метода сопоставления структур близость отбираемых из БД спектров к эталонному влечет за собой подобие соответствующих структур заданным эталонам. Хорошо видно, что в области MF 40-75 ед. в структурах отобранных соединений в среднем присутствуют от 20 до 70 % неизоморфных k -вершинных связанных фрагментов структуры анализируемого "нового" для БД соединения. При значениях $MF > 65$ поисковый ответ содержит структуры соединений с величинами $W > 0.5$ по отношению к эталону.

Насколько известно, подобный эксперимент за многолетнюю историю развития поисковых систем по масс-спектрометрии не описан. Это, возможно, и влияет на осторожность в оценках потенциала применения БД при анализе особенностей строения новых соединений. Представленные на рис. 2 данные позволяют утверждать, что для широкого набора органических соединений выявленная статистичес-

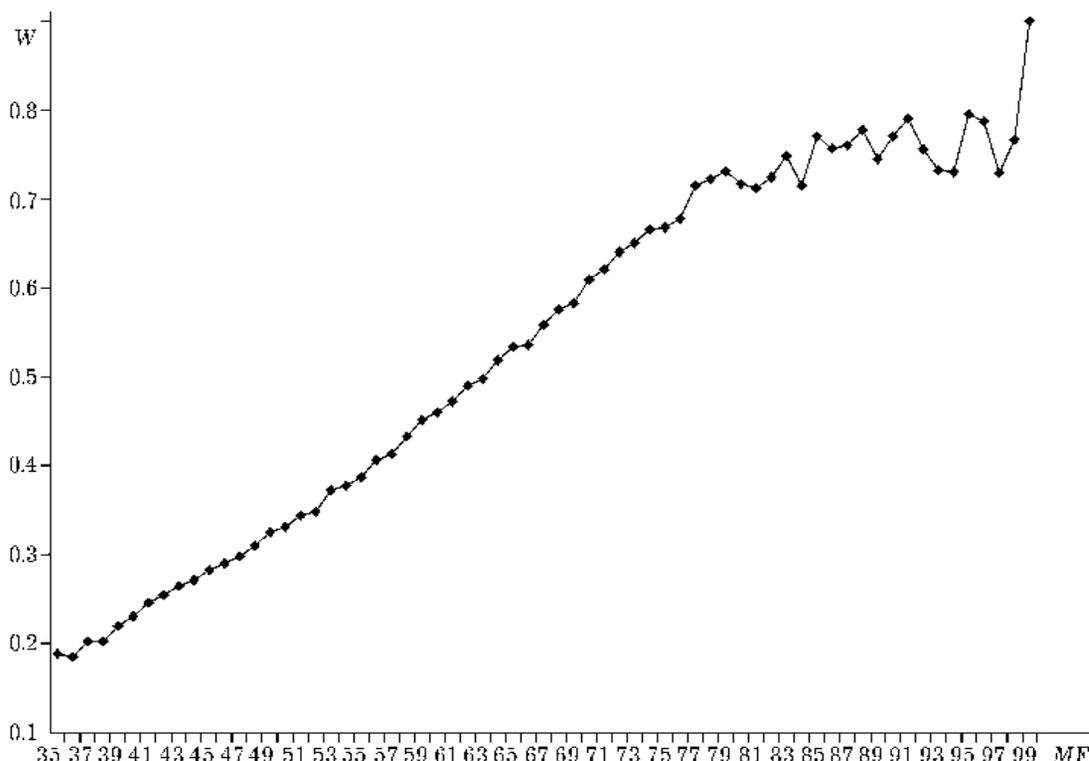


Рис. 2. Зависимость структурного подобия от спектрального подобия.

кая связь спектров и структур соединений закономерно выполняется. С другой стороны, наличие этой закономерности убеждает в возможности определения по масс-спектру k -вершинных фрагментов ($2 \leq k \leq 7$), входящих в состав структурной формулы действительно неизвестного соединения.

Попытаемся на примерах сопоставления обобщенных фрагментных составов десяти первых соединений ПО с составами соответствующих эталонов выяснить типы распознаваемых по масс-спектрам неизоморфных связанных фрагментов, вероятности и достоверности их опознания. Одновременно оценим долю распознаваемых фрагментов от их общего числа в соответствующих эталонах, а также соотношение между корректно и ошибочно распознаваемыми фрагментами. Соответствующие усредненные данные позволяют полностью охарактеризовать ожидаемый потенциал исследуемого приема представления структур в практике масс-спектрометрического анализа.

Расознаваемые фрагменты

Сразу же отметим, что в ходе анализа не рассматривались фрагменты структур эталонов, представленные в структурах тестовой выборки менее десяти раз. Это, с одной стороны, сокращает список анализируемых фрагментов, с другой – приближает к требованию статистической значимости получаемых результатов. Каждый фрагмент охарактеризован следующими параметрами.

1. Эффективность корректного распознавания фрагмента j (RC – recall [22]) по результату анализа поисковых ответов:

$$RC_j = N_j^+ / N_j \quad (2)$$

где N_j – число структур-эталонов, в составе которых присутствует фрагмент j ; N_j^+ – число корректных распознаваний фрагмента j . Фрагмент считается распознанным, если он появился не менее двух раз в списке соединений поискового ответа.

2. Достоверность (RL – reliability) опознания фрагмента j :

$$RL_j = RC_j / (RC_j + FP_j) \quad (3)$$

где $FP_j = N_j^- / (N - N_j)$ – возможность принятия ошибочного решения об идентификации фрагмента j (False Positive); N_j^- – число ошибочных опознаний фрагмента j ; N – общее число экспериментов (поисковых ответов, структур-эталонов). Из общего числа k -вершинных фрагментов ~ 40 тыс. ($2 \leq k \leq 7$) величины RC и RL рассчитывали лишь для ~ 8 тыс. фрагментов, встретившихся в структурах соединений выборки десять раз и более.

3. Неслучайность появления фрагмента j среди фрагментов 10 структур соединений поискового ответа:

$$NR = 1 - P(n) / P(10x) \quad (4)$$

$$P(z) = 10! (x)^z (1-x)^{10-z} / G(z+1) G(10-z+1)$$

Здесь $z = n$ или $10x$, G – гамма-функция, $P(n)$ – вероятность того, что при случайном выборе структур из БД в выборке размером 10 окажется n структур, содержащих фрагмент с относительной частотой встречаемости в структурах БД, равной x (ср. [16, 22, 33]). Параметр NR учитывает индивидуальную частоту встречаемости каждого фрагмента в БД и в поисковом ответе (ср. [24]). Его рассчитывали для всех фрагментов, присутствующих в поисковом ответе.

В табл. 2. приведено общее число корректно распознаваемых на объектах тестовой выборки фрагментов при условиях $RL \geq 0.95$ и $NR \geq 0.95$ для трех пороговых значений параметра RC . Эти данные получены в результате анализа ПО для двух различных режимов поиска системы КОМПАС-МС. Хорошо видно, что результат поиска в относительной шкале масс (поиск В) уступает таковому для случая анализа спектров в абсолютной шкале масс (поиск А). Эта общая тенденция наблюдается для всех типов фрагментов, поэтому в дальнейшем будем рассматривать данные, полученные в основном в режиме поиск А. В этом случае свыше 5.2 тыс. разнообразных фрагментов с числом связанных вершин от двух до семи распознаются с $RC \geq 0.5$, а более 3.5 тыс. фрагментов – с $RC \geq 0.75$. Например, из 4548 7-вершинных фрагментов, встретившихся в структурах соединений выборки десять раз и более, 1948 фрагментов имеют $RC \geq 0.75$, а 2544 – $RC \geq 0.5$.

ТАБЛИЦА 2

Число корректно распознаваемых *k*-вершинных фрагментов для трех пороговых значений параметра *RC*

<i>k</i>	<i>RC</i> ³ (поиск А)			<i>RC</i> ³ (поиск В)		
	0.3	0.5	0.75	0.3	0.5	0.75
2	33	28	12	30	22	7
3	112	93	50	92	74	34
4	327	287	164	266	209	116
5	840	737	474	710	566	365
6	1731	1558	1097	1414	1144	792
7	2745	2544	1938	2178	1880	1377

При пороговых значениях параметра *NR* > 0.95 заметно снижается общее число распознаваемых фрагментов. Одновременно увеличивается число поисковых ответов, в которых отсутствуют фрагменты, удовлетворяющие этому пороговому значению. Доля пустых поисковых ответов (поиск считается нерезультативным) особенно велика для 2-3-вершинных фрагментов и достигает 30 % даже при *NR*³ 0.95 (ср. с [16]). Для фрагментов большего размера вероятность нерезультативного поиска существенно ниже. Проиллюстрируем это на примере 7-вершинных фрагментов. В этом случае, если структура исследуемого соединения содержит 8 вершин и более, то при *NR*³ 0.95, 0.99 и 0.999 на объектах тестовой выборки выявлены ~ 3, 7 и 15 % нерезультативных поисков соответственно.

Анализ типов распознаваемых фрагментов и параметра *RC* позволяет в ряде случаев проследить согласованность найденных значений *RC* с устоявшимися взглядами на возможности метода масс-спектрометрии при их идентификации [25, 26]. Например, при *NR*³ 0.95 достаточно уверенно (*RC*³ 0.5) распознаются 2-вершинные фрагменты вида C-Hal, где Hal = Cl, Br, F, I, идентифицируемые в масс-спектрах или по характерным изотопным пи-

кам, или по потерям, или по соответствующим осколочным ионам. Прослеживается рост вероятности распознавания фрагментов в рядах C-C < C=C < C^oC; C-N < C=N < C^oN и т.п., что также согласуется с традиционными представлениями. Не противоречат им и низкие значения *RC*, типичные для этих фрагментов. Хорошо идентифицируются фрагменты N^oN, N=N, N=O, O=S (*RC*³ 0.7) и соответственно 3-вершинные вида N^oN=N, NO₂, SO₂ (*RC* = 1.0) и ряд других. Низкая вероятность распознавания ряда перечисленных выше связанных групп атомов, а также фрагментов вида C-O, C,C, C=O, C-N (знаком , обозначена связь в ароматическом цикле) и составленных из них трехвершинных фрагментов обусловлена широкой представленностью таких групп атомов в структурах соединений БД и завышенными в этом случае требованиями к пороговому значению параметра *NR*.

На достаточно простых примерах 7-вершинных фрагментов в ряде случаев также легко прослеживается аналогичная связь величин *RC* с традиционными представлениями о возможности опознания фрагментов по масс-спектрам, полученным при ионизации молекул электронным пучком. Например, *RC* возрастают в ряду фрагментов, показанных на схеме 1 (фрагмент, *RC*) (см. также табл. 3, стр. [2]).

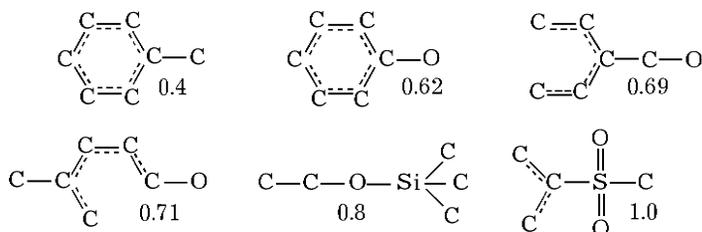
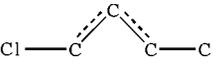
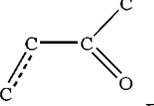
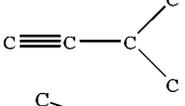
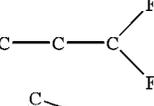
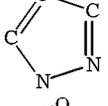
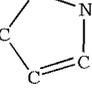
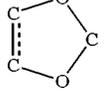
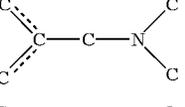
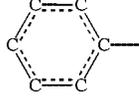
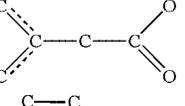
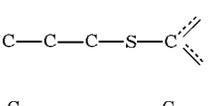
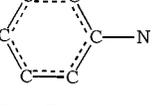
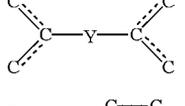
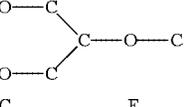
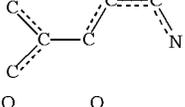
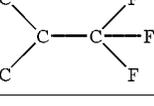
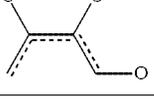


Схема 1.

ТАБЛИЦА 3

Примеры распознаваемых фрагментов

k	RC = 0.5-0.75			RC = 0.75-1.0		
	Фрагмент	RC	f	Фрагмент	RC	f
3	C, N, C	0.54	1080	C-Si-O	0.76	130
	N=C-O	0.56	103	N°C-C	0.79	353
	C°C-C	0.64	222	O=N-O	0.96	28
	N=N-C	0.65	58	N=C=O	1.0	35
	C, C-Br	0.74	88	N°N=N	1.0	24
5	C-O-P-O-C	0.51	33	C-C=C-S-C	0.76	94
		0.56	34		0.95	62
		0.60	351		1.0	33
		0.65	31		1.0	44
		0.71	99		1.0	17
7		0.52	48		0.81	268
		0.6	77		0.88	43
		0.62	1514		0.82-1.0	18-108
		0.7	47		1.0	19
		0.75	40		1.0	26

Примечание. Символом \circ и штриховой линией обозначены связи в ароматическом цикле; Y = C, O, N, S, C=O.

Многочисленность опознаваемых фрагментов (см. табл. 2) не позволяет перечислить даже ту их часть, которая идентифицируется при пороге $RC \geq 0.75$. Поэтому в табл. 3 приведены лишь по пять фрагментов ($k = 3, 5$ и 7) из групп, характеризующихся двумя различными интервалами значений параметра RC . Здесь же приводятся соответствующие значения RC и частоты встречаемости фрагментов

в структурах соединений БД и в использованной выборке.

Вероятность распознавания мелких фрагментов, изоморфно вкладывающихся в более крупные, как правило, меньше, чем таковая для крупных фрагментов. Это обусловлено двумя основными причинами: 1) взаимной коррелируемостью фрагментов и повышенной частотой встречаемости в БД мелких фрагментов

ментов; 2) относительно большей ролью крупных фрагментов, например 7-вершинных, в описании строения соединения и, следовательно, в отражении процессов распада ионов в соответствующих масс-спектрах. Вероятно, поэтому в реальной практике масс-спектрометрии интерес будут представлять не все формально определяемые по поисковому результату фрагменты, а лишь наиболее крупные из них и те из мелких, которые изоморфно не вкладываются в крупные или, вкладываясь, имеют большее значение величины неслучайности появления в ПО. Оценка показала, однако, что доля таких фрагментов в соответствующих списках не превышает 1.5 %.

Заметим, что наличие в ПО некоторых 7-вершинных фрагментов дает информацию о присутствии более крупных элементов структуры. Примеры таких фрагментов приведены на схеме 2. В сочетании со взаимной перекрываемостью частей выявляемых в ПО фрагментов это позволяет в ряде случаев получать достаточно полную информацию о скелете структуры изучаемого по масс-спектру соединения.

Общая характеристика метода

На рис. 3 даны несколько примеров, показывающих характер сведений о соединении, извлекаемых при анализе масс-спектра в рамках обсуждаемого подхода. Здесь приводятся масс-спектр, структурная формула соединения, примеры 7-вершинных корректных и ошибочных фрагментов, выявленных при анализе соответствующего спектра и поискового ответа. Под фрагментами указано их общее число в ПО для приведенного под изображением структуры порогового значения параметра NR (n_c - для корректных фрагментов, n_f - для ошибочных). Видно, что даже в тех случаях, когда набор выявляемых корректных фрагментов достаточно полно опи-

сывает особенность строения рассматриваемого "неизвестного", он сопровождается перечнем удовлетворяющих заданному пороговому значению параметра NR ложных (не вкладывающихся в эталон) фрагментов.

Набор выявляемых фрагментов полностью определяется заданным спектром, параметром NR и содержимым базы данных. Он индивидуален в каждом конкретном случае. Поэтому для более полной характеристики рассматриваемого приема представления структур соединений БД и анализа масс-спектров с целью распознавания особенностей строения соединения приведем ряд дополнительных данных. Среди них, вероятно, наибольший практический интерес представляют средняя доля корректных (т.е. вложимых в структуру-эталон) фрагментов среди всех фрагментов, определяемых по поисковому ответу, а также средняя доля корректно выявляемых фрагментов среди фрагментов анализируемых эталонов. Соответствующие данные для фрагментов различного размера и ряда пороговых значений параметра NR представлены на рис. 4. Они получены с помощью выборки эталонов, содержащей ~ 13 000 масс-спектров разнообразных органических соединений.

Как и следовало ожидать, с ростом порога неслучайности NR доля корректных фрагментов (C) заданного размера (k) среди фрагментов поисковых ответов, имеющих неслучайность выше заданного порога, растет. Она вычислялась следующим образом:

$$C = \frac{1}{N_{res}} \sum \frac{n_c}{n}$$

где N_{res} - число поисковых ответов, содержащих хотя бы один фрагмент заданного размера, n_c - число корректных фрагментов в поисковом ответе среди всех (n) фрагментов заданного размера, имеющих неслучайность выше установленного порога.

Доля корректно распознаваемых фрагментов заданного размера среди фрагментов со-

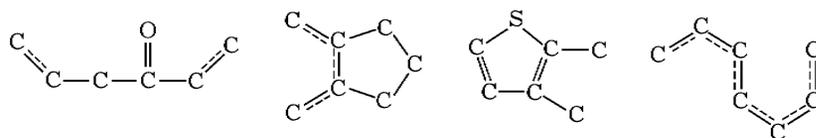


Схема 2.

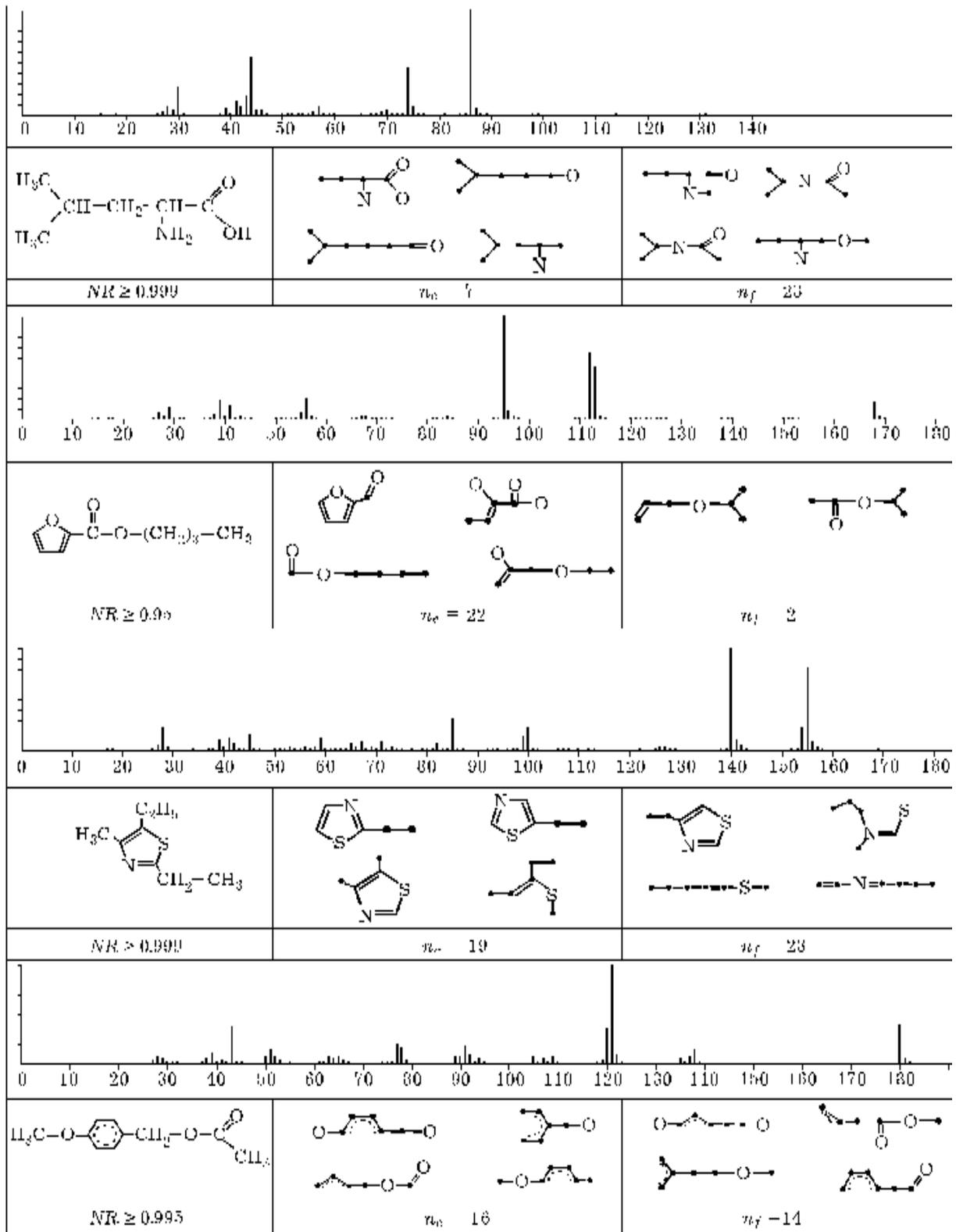


Рис. 3. Примеры масс-спектров, структур-эталонных и выявленных фрагментов (в ряде случаев углеродсодержащие вершины фрагментов помечены точкой).

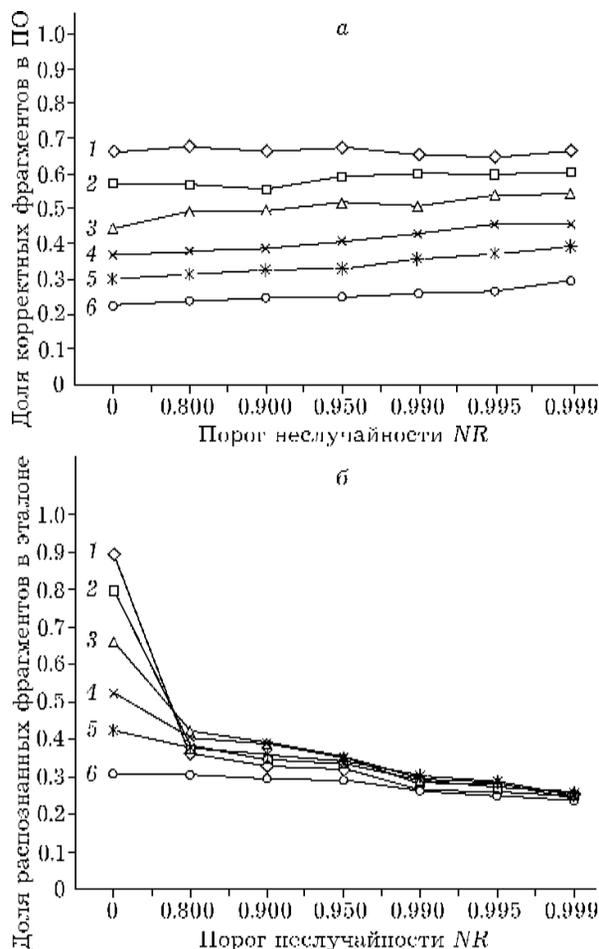


Рис. 4. Доля корректных фрагментов среди всех фрагментов ПО (а) и доля фрагментов, распознанных в структуре-эталоне (б). Размер фрагментов: 2 (1), 3 (2), 4 (3), 5 (4), 6 (5), 7 (6).

ответствующих структур-эталонов в зависимости от величины порога неслучайности NR рассчитывалась по следующему выражению:

$$D = \frac{1}{N_{res}} \sum \frac{n_c}{n_e}$$

где n_e - количество фрагментов в структуре-эталоне.

Как видно, при значениях $NR \geq 0.95$ число корректных фрагментов в среднем превышает число ошибочных для значений $k \leq 4$. Это согласуется с известными данными о возможности использования масс-спектрометрических поисковых систем для предсказания молекулярной формулы соединения. Для $k = 5-7$ доля корректных фрагментов в общем списке заметно ниже, чем соответствующая в случае ИК-спектроскопии [16]. Одна из возможных причин этого эффекта, по-видимому, обусловлена меньшей способностью масс-

спектрометрии различать структурные изомеры.

Максимальное среднее значение параметра D для всех значений k достигается при $NR < 0.8$. Нельзя, однако, при этом забывать, что в этом случае результат идентификации некоторых фрагментов будет носить случайный характер, поэтому для практических целей целесообразно выбирать условия, соответствующие $0.95 \leq NR \leq 0.999$.

Сопоставление полученных данных с аналогичными при анализе ИК-спектров [16] показывает более высокую результативность распознавания фрагментов в ИК-спектроскопии, что хорошо согласуется с большей устойчивостью спектроструктурных корреляционных связей в этом случае. Тем не менее представленные экспериментальные данные и тщательный анализ многочисленных примеров свидетельствуют о значительном потенциале метода поиска спектральных аналогов в масс-спектрометрической БД, содержащей информацию о полном наборе неизоморфных k -вершинных фрагментов структур соответствующих соединений.

Многочисленность и разнообразие распознаваемых фрагментов при сопоставимом содержании "шума" в ПО, дублирование различными фрагментами структурных особенностей, ответственных за спектр, близость достигаемых результатов к аналогичным в случае ИК-спектроскопии позволяют предположить, что на основе выявляемой информации возможно формирование наиболее вероятной структуры изучаемого соединения в рамках приема, описанного в работе [27]. В дальнейшем несомненный интерес представляют результаты, достигаемые при анализе двух типов (ИК- и масс-) спектров соединения, или использование получаемых при анализе масс-спектра данных в комплексных системах по различным видам спектроскопии молекул.

Особенность рассмотренного подхода заключается в том, что декомпозиция молекулярных графов представленных в БД соединений и регистрация содержащихся в них фрагментов реализуются один раз на начальном этапе формирования баз данных структур и фрагментов. Это приводит к тому, что временные затраты на анализ ПО становятся

ся минимальными, а сопровождающие фрагменты значения параметра NR рассчитываются с учетом реальной частоты встречаемости каждого фрагмента в структурах соединений БД.

Возможности данного подхода продемонстрированы на фрагментах, не содержащих атомы водорода в качестве вершин. Очевидно, что дальнейшая детализация описания фрагментов может позволить уточнить характер выявляемой информации о соединении, если другие причины, например типичные для масс-спектрометрии перегруппировки атомов водорода или статистически низкая представленность в БД соответствующих водородсодержащих фрагментов, не будут препятствовать этому.

Авторы благодарят Российский фонд фундаментальных исследований (грант 01-03-32357) за частичную поддержку данной работы, а также С. П. Киршанского за содействие в ее выполнении.

СПИСОК ЛИТЕРАТУРЫ

- 1 W. A. Warr, *Anal. Chem.*, 65 (1993) 1045A.
- 2 K. S. Haraiki, R. Venkataraghavan, F. W. McLafferty, *Ibid.*, 53 (1981) 386.
- 3 W. L. Domokoš, D. Henneberg, *Anal. Chim. Acta*, 150 (1983) 37.
- 4 С. П. Киршанский, К. С. Лебедев, Б. Г. Дерендяев, *Журн. аналит. химии*, 42 (1987) 1092.
- 5 Л. М. Покровский, Б. Г. Дерендяев, *Изв. СО АН СССР. Сер. хим. наук*, 4 (1989) 88.
- 6 S. E. Stein, *J. Am. Soc. Mass Spectrom.*, 6 (1995) 644.
- 7 R. Neudert, W. Bremser, H. Wagner, *Org. Mass Spectrometry*, 22 (1987) 321.
- 8 К. С. Лебедев, С. П. Киршанский, *Изв. СО АН СССР. Сер. хим. наук*, 4 (1989) 79.
- 9 M. M. Cone, R. Venkataraghavan, F. W. McLafferty, *J. Anal. Chem. Soc.*, 99 (1977) 7668.
- 10 K. S. Lebedev, V. M. Tormyshev, B. G. Derendyaev, V. A. Koptuyug, *Anal. Chim. Acta*, 133 (1981) 517.
- 11 H. Scsibrany, K. Varmuza, *Fresenius J. Anal. Chem.*, 344 (1992) 220.
- 12 K. S. Lebedev, D. Cabrol-Bass, *J. Chem. Inf. Comput. Sci.*, 38 (1998) 410.
- 13 W. Bremser, *Anal. Chim. Acta*, 103 (1978) 355.
- 14 И. И. Строков, К. С. Лебедев, Б. Г. Дерендяев, *Журн. структур. химии*, 37 (1996) 1128.
- 15 N. A. V. Gray, *Computer-Assisted Structure Elucidation*, Wiley & Sons, New York, 1986.
- 16 В. Н. Пиотух-Пелецкий, И. К. Коробейничева, Б. Г. Дерендяев, *Журн. аналит. химии*, 54 (1999) 1020.
- 17 V. N. Piottukh-Peletsky, I. K. Korobeinicheva, T. F. Bogdanova et al., *Anal. Chim. Acta*, 409 (2000) 181.
- 18 V. N. Piottukh-Peletsky, B. G. Derendyaev, *Ibid.*, 396 (2000) 99.
- 19 Б. Г. Дерендяев, В. Н. Пиотух-Пелецкий, С. А. Нехорошев, и др., *Журн. структур. химии*, 40 (1999) 728.
- 20 В. Н. Пиотух-Пелецкий, В. И. Смирнов, А. К. Румянцев, Б. Г. Дерендяев, *Сиб. хим. журн.*, 3 (1993) 65.
- 21 Б. Г. Дерендяев, В. Н. Пиотух-Пелецкий, *Журн. структур. химии*, 40 (1999) 198.
- 22 H. E. Dayringer, G. M. Pesyna, R. Venkataraghavan, F. W. McLafferty, *Org. Mass Spectrometry*, 11 (1976) 529.
- 23 Е. С. Вентцель, *Теория вероятностей*, Физматгиз, Москва, 1963, с. 58.
- 24 K. Varmuza, P. N. Penchev, H. Scsibrany, *J. Chem. Inf. Comput. Sci.*, 38 (1998) 420.
- 25 Г. Будзикович, К. Джерасси, Д. Уильямс, *Интерпретация масс-спектров органических соединений*, Мир, Москва, 1966, с. 323.
- 26 И. Г. Зенкевич, Б. В. Иоффе, *Интерпретация масс-спектров органических соединений*, Химия, Ленинград, 1986, с. 175.
- 27 В. Н. Пиотух-Пелецкий, Б. Г. Дерендяев, С. Г. Молодцов, Т. Ф. Богданова, *Журн. структур. химии*, 38 (1997) 791.