ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

2008. Том 49, № 2 *Март – апрель* *С. 238 – 248*

УДК 543.51+543.42:681.32

ОЦЕНКА ЭФФЕКТИВНОСТИ СОВМЕСТНОГО ИСПОЛЬЗОВАНИЯ БАЗ ДАННЫХ ПО ИК И МАСС-СПЕКТРОМЕТРИИ ДЛЯ УСТАНОВЛЕНИЯ СТРОЕНИЯ НЕИЗВЕСТНОГО СОЕДИНЕНИЯ

© 2008 **Т.А. Корнакова*, Т.Ф. Богданова, В.Н. Пиоттух-Пелецкий**

Новосибирский институт органической химии им. Н.Н. Ворожцова

Статья поступила 12 июля 2007 г.

Проведено исследование применимости структурной аналогии для установления строения органического соединения при совместном поиске в двух БД по различным видам молекулярных спектров. Использование структурной аналогии основывается на представлении структур соединений БД в виде наборов структурных фрагментов. Наибольший интерес представляют структурные фрагменты, которые представлены в поисковом ответе как по ИК, так по масс-спектрометрии. Приведены статистически обоснованные оценки эффективности совместного поиска в зависимости от спектрального подобия.

Ключевые слова:

В практике исследований, проводимых химиками, спектроскопистами, экологами, в криминалистических службах, в аналитических лабораториях, при анализе компонентов химических производств требуется определение структурной формулы неизвестного соединения. Основным, а иногда единственным источником информации о строении неизвестного соединения являются данные, полученные по спектрам различной физической природы, таким как ИК, масс ЯМР, УФ и т.д. Совместное использование нескольких спектральных методов для установления строения имеет давнюю историю [1—9]. Самый простой вариант совместного использования — идентификация соединения по двум видам спектров, т.е. нахождение в базах данных (БД) спектров, совпадающих с искомыми. Результат идентификации обычно бывает гораздо более надежным и убедительным, чем идентификация по одному виду спектров, так как для каждого вида спектроскопии существуют примеры различающихся соединений, имеющих одинаковые спектры, а вероятность совпадения спектров сразу в двух БД по различным видам спектроскопии крайне мала. Однако возможности точной идентификации по двум видам спектров ограничены: соединения, присутствующие одновременно в обеих спектральных БД, составляют лишь небольшую долю, от их объема [10]. Например, используемые в данной работе БД по ИК и масс-спектрам имеют примерно 10 % общих соединений [11].

Второй подход, используемый при установлении строения, основан на анализе спектральных аналогов и соответствующих им структур, отбираемых при поиске в БД по спектру исследуемого соединения [12—16]. Результатом его применения является список структурных фрагментов, распознанных в отобранных по спектру структурах и предположительно присутствующих в исследуемом соединении. Эти структурные фрагменты впоследствии могут быть использованы генератором структур для построения гипотетических структурных формул исследуемого соединения. Применение такого подхода оказывается гораздо более перспективным, поскольку его распознающая способность определяется не только числом спектров и структур в используемой БД, но и количеством потенциальных аналогов для структур БД со

* E-mail:

степенью структурной аналогии выше определенного порога. Можно ожидать, что число соединений, строение которых можно определить с помощью описанного метода, как минимум, на порядок превышает объем БД.

Каждый вид молекулярных спектров (ИК, масс- и др.) позволяет выявлять специфические для данного вида спектроскопии структурные особенности исследуемого соединения. В связи с тем, что данные, полученные из анализа спектров различной физической природы, могут быть взаимодополняющими, их совместное использование в информационных системах уже давно представляет интерес для целей установления строения неизвестных соединений. Использование подхода, базирующегося на анализе ближайших спектральных и структурных аналогов, позволяет решать задачи установления строения соединений, спектры которых отсутствуют в БД. При обработке результатов поиска в двух спектральных БД основное внимание уделяется отбору соединений, спектры которых наиболее близки исследуемому. Для них можно ожидать также и структурной близости к исследуемому соединению. В последнее время растет интерес к применению концепции ″структурного подобия″ для прогнозирования химических свойств соединений [17], в том числе и для решения задач взаимосвязи спектр-структура [13, 15, 16, 18]. Ранее для используемых в данной работе баз данных по ИК и масс спектрам было установлено [11], что большинство соединений из одной БД имеет хотя бы один структурный аналог из другой БД. Использование структурной аналогии в предыдущих наших работах и в данной работе основывается на описании структур соединений БД полными наборами фрагментов заданного размера.

Целью данного исследования является оценка эффективности описанного метода при совместном поиске в двух БД по ИК и масс-спектрам для установления строения органических соединений.

ОБЩАЯ СХЕМА УСТАНОВЛЕНИЯ СТРУКТУРЫ НЕИЗВЕСТНОГО СОЕДИНЕНИЯ

Установление структуры неизвестного соединения с помощью БД предусматривает прохождение двух основных этапов: отбор в БД соединений, спектры которых наиболее близки предъявленному, и анализ результатов поиска с целью построения гипотез о строении исследуемого соединения (рис. 1).

При поиске спектральных аналогов исследуемого соединения его молекулярный спектр (ИК или масс-) сопоставляют с каждым спектром из БД, вычисляют степень совпадения сравниваемых спектров и по окончании поиска результаты, имеющие степень совпадения не ниже заданного порога, ранжируют и записывают в поисковый ответ (ПО).

Традиционно информацию, извлекаемую из ИК и масс-спектров, представляют в виде небольшого набора структурных фрагментов, характеризующих химический класс или особенности строения соединения. При установлении строения исследуемого соединения используют манипулирование структурами в топологическом описании или представленными полными фрагментными составами структур. Последний метод менее трудоемкий и более адекватный природе взаимосвязи ИК и масс-спектров со структурой. Характерной особенностью нашего подхода является однократное разбиение структур всех соединений БД на неизоморфные связные k-вершинные фрагменты молекулярных графов структур ($k = 2 \div 7$) в процессе формирования базы данных [19]. В качестве вершин фрагментов выступают все атомы, кроме атомов водорода. Такое описание позволяет представить структурные формулы соединений в виде числового вектора, каждая компонента которого соответствует присутствию или отсутствию в структуре определенного фрагмента из исчерпывающего набора (2÷7)-вершинных фрагментов. Сопоставление этих векторов для двух структур дает возможность оценивать структурное подобие соединений

$$W = 2F_{12} / (F_1 + F_2), \qquad (1)$$

где F_{12} — число совпавших k-вершинных фрагментов; F_1 и F_2 — количество k-вершинных фрагментов в сравниваемых структурах.

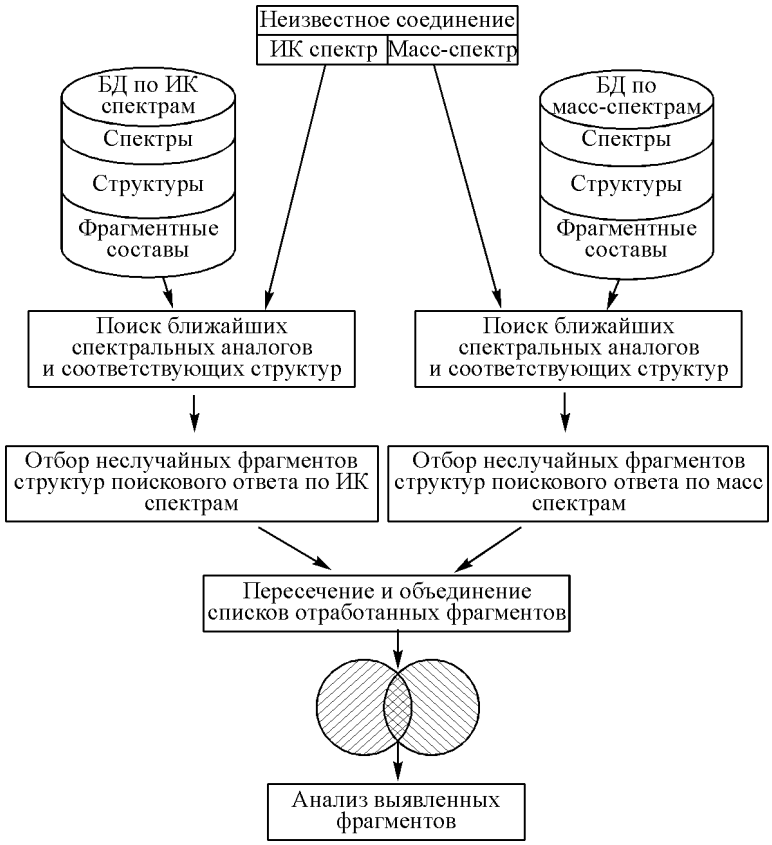

Рис. 1. Схема установления строения неизвестного соединения с помощью использования БД по ИК и масс-спектрам

Описание структурных формул соединений в виде полных наборов связных структурных фрагментов позволяет формировать списки фрагментов всех структур ПО. Среди фрагментов структур ПО встречаются как тривиальные, т.е. присутствующие практически во всех структурах БД, так и фрагменты, обусловленные спектральной близостью отобранных спектров к спектру исследуемого соединения. Для различения тривиальных и неслучайных фрагментов можно использовать их статистические характеристики. Анализ частот встречаемости k-вершинных фрагментов во фрагментных составах структур ПО наряду с известными данными о частотах их встречаемости в структурах всех соединений БД (хранящихся в БД) позволяет вычислить неслучайность появления фрагмента j в структурах N соединений поискового ответа [20]:

$$NR = 1 - P(n) / P(Nx),\ P(z) = N!(x)z(1–x)(N–z) / G(z+1)\ G(N–z+1). \qquad (2)$$

Здесь $z = n$ или Nx; G — гамма функция; $P(n)$ — вероятность того, что при случайном выборе структур из БД в выборке размером N окажется n структур, содержащих фрагмент с относительной частотой встречаемости в структурах БД, равной x.

Параметр NR учитывает индивидуальные частоты встречаемости каждого фрагмента в БД и в поисковом ответе и выступает в качестве оценки достоверности распознания соответствующего фрагмента. Представляют интерес те структурные фрагменты, которые имеют неслучайность выше определенного порога.

Таким образом, при анализе поискового ответа формируется список наиболее вероятных k-вершинных связных структурных фрагментов, предположительно присутствующих в исследуемом соединении, называемый далее фрагментным составом поискового ответа. На основе выявленных фрагментов могут строиться гипотезы о строении исследуемого соединения (вручную или с помощью специальной программы-генератора структур). Как было показано ранее,

описанный метод оказывается весьма продуктивным для установления строения исследуемого соединения на примере ИК спектроскопии [20] и масс-спектрометрии [14].

При совместном использовании БД по ИК спектроскопии и масс-спектрометрии формируется список фрагментов, объединяющий фрагментные составы поисковых ответов по ИК и масс-спектрам.

ОЦЕНКА ПОИСКОВОГО ОТВЕТА

В списках выявляемых фрагментов с неслучайностью выше определенного порога наряду с правильно распознанными (корректными) фрагментами могут присутствовать фрагменты, отсутствующие в структуре исследуемого по спектру соединения (некорректные, ложно распознанные). Появление некорректных фрагментов можно объяснить неоднозначностью спектроструктурных зависимостей, когда одному и тому же положению полосы поглощения в ИК спектре или массе осколка в масс-спектре могут соответствовать различные элементы структуры. Другой причиной появления некорректных фрагментов может быть слабая представленность в БД химических классов, характеризующих особенности строения исследуемого соединения. В каждом частном случае результат конкретного поиска определяется объемом и составом БД, представлением спектральных и структурных данных, размером ПО и другими параметрами, которые могут также привести к получению некорректной информации. Для отсева некорректных фрагментов можно использовать имеющуюся дополнительную информацию о соединении, например, сведения о брутто-формуле, о предыстории образца, данные других физических методов и т.д.

В наших работах, например [20], при исследовании эффективности процедуры выявления фрагментов по результатам спектрального поиска используются статистические оценки. Для этой цели на примере статистически значимой выборки соединений с заранее известными структурными формулами (далее соединения-эталоны) анализируется эффективность выявления корректных фрагментов этих соединений по двум параметрам. Параметр D отражает, насколько полно выявленные фрагменты описывают фрагментный состав исследуемого соединения (степень покрытия корректными фрагментами структуры соединения-эталона). Параметр C показывает, какова доля корректных фрагментов среди всех фрагментов поискового ответа:

$$D = 1 / N\Sigma(n_s / n_e), \tag{3}$$

$$C = 1 / N\Sigma\ (n_s / n_f), \tag{4}$$

где n_s — число корректных фрагментов, выявленных при поиске; n_e — общее число фрагментов в структуре соединения-эталона; n_f — общее число фрагментов, выявленных при поиске; N — количество структур, для которых поиск привел к выявлению хотя бы одного фрагмента при заданном пороге неслучайности.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Из БД по ИК спектроскопии (содержащей 31989 соединений с усредненной брутто-формулой $C_{14,7}H_{16,1}N_{1,5}O_{2,3}S_{0,3}Cl_{0,3}F_{0,2}Br_{0,2}$) и БД по масс-спектрометрии (содержащей 53912 соединений $C_{14}H_{19}N_1O_2S_{0,2}Cl_{0,2}F_{0,3}Si_{0,2}$) было выбрано в качестве соединений-эталонов 3600 соединений, присутствующих как в БД по ИК спектроскопии, так и в БД по масс-спектрометрии с количеством остовных атомов в структурах, больше или равным семи. Усредненная брутто-формула этой выборки — $C_{11,4}H_{15}N_{0,6}O_{1,7}S_{0,1}Cl_{0,2}F_{0,1}Br_{0,1}$. Количество 7-вершинных фрагментов в БД по ИК спектроскопии и по масс-спектрометрии составляет 76514 и 83280 соответственно.

Для каждого соединения из выборки был произведен поиск десяти ближайших спектральных аналогов с критерием подобия спектров для ИК спектров ≥30, для масс-спектров ≥40. При сопоставлении спектров в базе данных по ИК спектроскопии использовали параметр спектрального подобия MF [13]:

$$MF = (A + M + P)/3, \tag{5}$$

где A, M, P — компоненты, описывающие степень совпадения спектров: P — только по положениям полос поглощения; A — только по положениям полос с близкими значениями интенсивностей (в наших экспериментах — $T \pm 30$ %); M — по положениям и интенсивностям всех сравниваемых полос, значения параметра MF варьируются от 0 до 100.

При отборе ближайших спектральных аналогов из БД по масс-спектрометрии использовался критерий спектрального подобия *MS*, вычисляемый по формуле:

$$MS = 100 \cdot (1 - (\Sigma x_i + \Sigma y_j - \Sigma (x_k + y_k) + \Sigma |x_k - y_k|) / (\Sigma x_i + \Sigma y_j)), \quad (6)$$

где x_i — интенсивность пика в исследуемом спектре; y_j — интенсивность пика в спектре из БД, x_k, y_k — интенсивности совпавших пиков исследуемого спектра и спектра из БД, значения параметра *MS* варьируются от 0 до 100.

По методу, описанному выше, были сформированы списки фрагментов структур ПО со значениями неслучайности их появления ≥0,95 (вычисляемыми в соответствии с (2)) и частотой встречаемости в поисковом ответе не менее двух. Было выявлено 3186 ПО со списками фрагментов, удовлетворяющих этим условиям. Информацию о соединении-эталоне не включали в поисковый ответ, таким образом моделировали условия ″новизны″ изучаемого соединения. В списки включали, как наиболее информативные, только семивершинные фрагменты.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для анализа эффективности совместного использования баз данных по ИК спектроскопии и масс-спектрометрии были введены понятия **объединения** и **пересечения** фрагментных составов поисковых ответов. Объединение фрагментных составов поисковых ответов (упоминаемое далее как ″объединение ФС″) представляет собой полный список 7-вершинных фрагментов, каждый из которых встретился либо в поисковом ответе по ИК спектрам, либо по масс-спектрам, либо в обоих поисковых ответах. Пересечение фрагментных составов поисковых ответов (упоминаемое далее как ″пересечение ФС″) состоит из 7-вершинных фрагментов, которые одновременно присутствуют в результатах поиска как по ИК спектроскопии, так и по масс-спектрометрии.

На рис. 2 отображена взаимосвязь объединения и пересечения фрагментных составов на конкретном примере, где в качестве тестовой поисковой структуры выбран 1-амино-4-гидрокси-2-метокси-9,10-антрахинон. В состав структуры соединения входят 144 семивершинных фрагмента, из них 15 не были выявлены в результате поиска. Все 35 структурных фрагментов пересечения ФС оказались корректными. Среди 225 структурных фрагментов объединения ФС 129 — корректные и 96 — некорректные.

Оценки результативности распознавания структурных фрагментов при поиске в БД по ИК спектроскопии и масс-спектрометрии для 2091 ПО представлены в таблицы (из 3186 оставлены только поисковые ответы с непустым пересечением ФС). Для оценки результативности пересечения фрагментных составов были использованы параметры D_p и C_p, вычисляемые по формулам (3) и (4) соответственно, аналогично для оценки объединения фрагментных составов — параметры D_v и C_v. Среднее количество 7-вершинных фрагментов в структурах-эталонах выборки — 27.

Обращает на себя внимание высокая доля корректных фрагментов (81 %) в пересечении ФС. Такое достаточно высокое значение параметра C_p показывает, что фрагменты, выявленные одновременно как при поиске по ИК спектрам, так и по масс-спектрам, могут оказаться, с

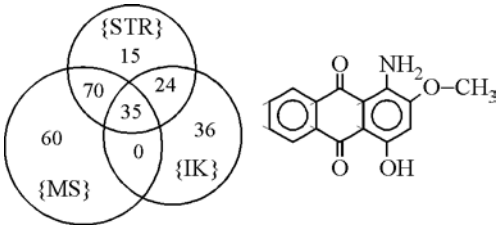

Рис. 2. Взаимосвязь множеств {STR}, {IK} и {MS} на примере поиска 1-амино-4-гидрокси-2-метокси-9,10-антрахинона при пороге ″неслучайности″ появления фрагментов 0,95. {STR} — фрагментный состав исследуемого соединения, {IK} — фрагментный состав поискового ответа по ИК спектру, {MS} — фрагментный состав поискового ответа по масс-спектру

Таблица 1

Эффективность выявления фрагментов по результатам поиска в каждой из БД *по* ИК- *и масс-спектрам, а также в совместном поисковом ответе*

Сопоставление результатов поиска	*F*	*S*	*D*·100(%)	*C*·100(%)
Фрагментные составы ПО в БД по ИК-спектрам	47	15	62	40
Фрагментные составы ПО в БД по масс-спектрам	37	14	56	45
Пересечение фрагментных составов (D_p, C_p)	11	9	42	81
Объединение фрагментных составов (D_v, C_v)	73	20	78	33

Примечания. *F* — средний размер фрагментного состава ПО (для 7-вершинных фрагментов), *S* — среднее количество корректных 7-вершинных фрагментов в ПО, *D* — степень покрытия структуры соединения-эталона фрагментами поискового ответа. *C* — доля фрагментов соединения-эталона (″корректных″) среди фрагментов поискового ответа.

большой степенью вероятности, фрагментами исследуемой структуры. В то же время покрытие структуры исследуемого соединения фрагментами пересечения ФС составляет в среднем около 40 %, т.е. ниже, чем при поиске только по ИК спектрам или только по масс-спектрам. Полученные результаты обусловлены различной физической природой спектров, поскольку при поиске по различным видам молекулярных спектров в поисковый ответ отбираются соединения, в состав которых входят фрагменты, отражающие разные особенности их строения. Поэтому вероятность появления случайных фрагментов в пересечении фрагментных составов поисковых ответов по ИК и масс-спектрам уменьшается.

Совместное использование двух БД (по ИК и по масс-спектрам) расширяет список фрагментов, предположительно входящих в структуру исследуемого соединения. Поэтому, как показывает значение D_v (78 %), покрытие структуры исследуемого соединения фрагментами объединения ФС должно быть высоким. В то же время, при увеличении количества корректных фрагментов резко увеличивается зашумленность некорректными фрагментами, на что указывает относительно невысокое значение C_v (33 %).

Во многих случаях анализ структурных фрагментов пересечения и объединения ФС позволяет подсказать детали строения не представленного в БД соединения, даже если отобранные из базы спектры заметно отличаются от искомого. Для иллюстрации рассмотрим пример условно ″неизвестного″ соединения (взятого из БД), ИК и масс-спектры которого представлены на рис. 3. При анализе результатов поиска сведения, относящиеся к нему, были исключены из

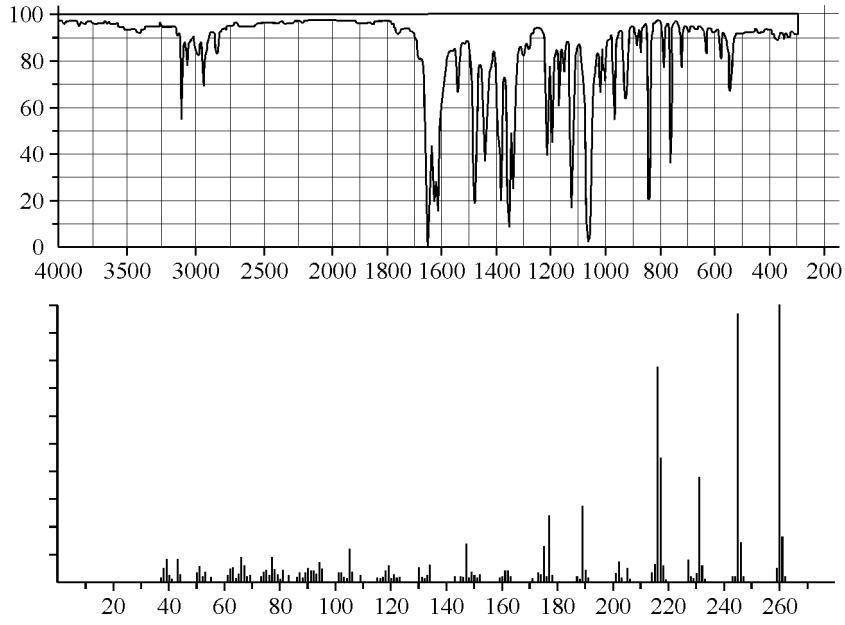

Рис. 3. ИК- и масс-спектры ″неизвестного″ соединения

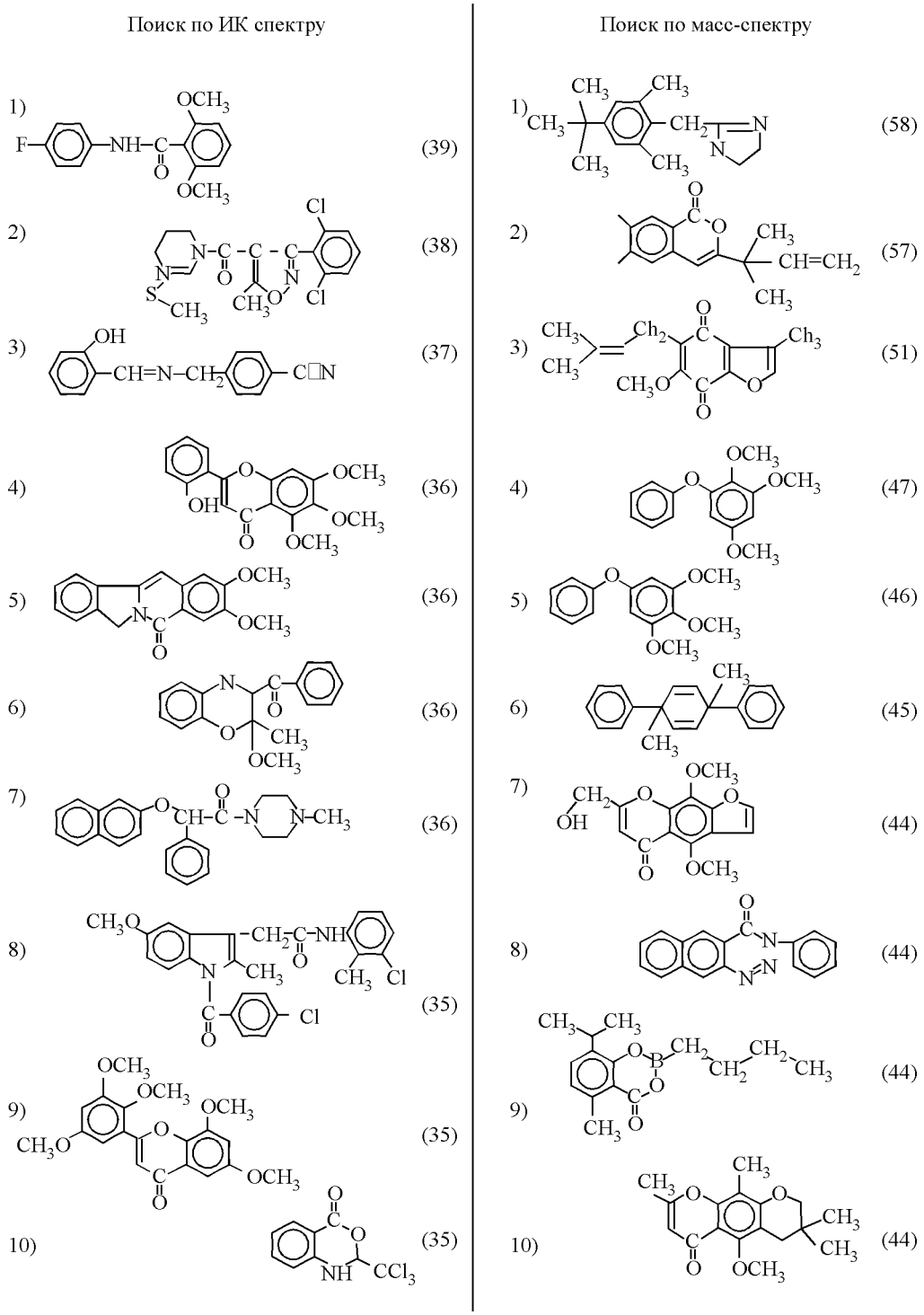

Рис. 4. Структуры первых десяти спектральных аналогов, полученных при поиске по ИК и масс-спектрам "неизвестного" соединения. В скобках указаны факторы совпадения спектров отобранных соединений со спектром исследуемого соединения

ПО. Из визуального анализа структур первых десяти спектральных аналогов достаточно сложно сформулировать предположение о строении исследуемого соединения (рис. 4).

Анализ фрагментных составов структур отобранных соединений позволил сформировать пересечение и объединение фрагментных составов поисковых ответов по ИК и масс-спектрам.

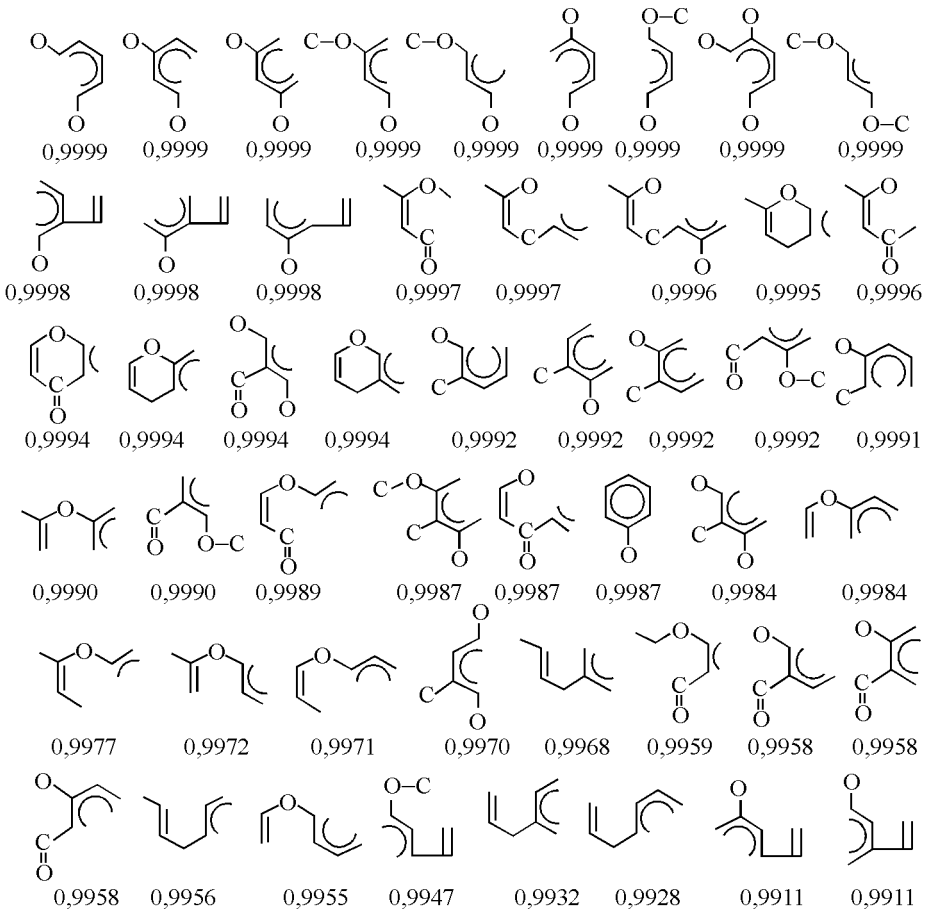

Рис. 5. Структурные фрагменты пересечения ФС поисковых ответов со значениями неслучайности появления ≥0,99

Пример пересечения ФС при пороге неслучайности появления структурных фрагментов 0,99 представлен на рис. 5.

Анализ структурных фрагментов позволяет достаточно уверенно высказать предположение о наличии крупного фрагмента (1), изображенного на рис. 6.

Для каждой структуры в поисковых ответах по ИК и масс-спектрам определяли количество фрагментов, совпадающих с пересечением ФС. Максимальное количество таких совпадающих фрагментов наблюдали для структуры под номером 7 при поиске по масс-спектру (см. рис. 4). Используя данную структуру как модельную, а также учитывая дополнительные знания, например, молекулярный вес (МВ=260), можно уверенно предположить, что исследуемое соединение является амикардином (структурная формула которого представлена на рис. 6), используемым в медицине как лекарственное средство.

Структурная формула спектрального аналога, имеющая максимальное количество совпадающих фрагментов с пересечением ФС, во многих случаях может использоваться как своего рода модель, полезная для построения структуры исследуемого соединения. Так, для 45 % соединений из выборки 2091 соединения, значение структурного подобия с исследуемым соединением было ≥0,8. Значение структурного подобия ≥0,6 наблюдалось для 68 % соединений. В то же время для спектральных аналогов в БД по ИК спектроскопии, имеющих наибольшее спектральное подобие с исследуемым соединением, струк-

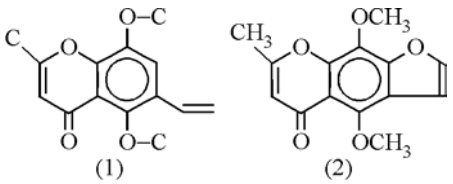

Рис. 6. 1 — Структурный фрагмент полученный при анализе пересечения и объединения фрагментных составов поисковых ответов по ИК и масс-спектрам — *2*

турное подобие ≥0,8 наблюдалось для 31 % соединений, для 50 % соединений со структурным подобием ≥0,6. Для первого спектрального аналога в БД по масс-спектрометрии доля таких соединений составляла 32 и 54 % соответственно.

При переходе от статистических оценок метода выявления фрагментов *D* и *C* к анализу каждого конкретного поискового ответа требуется знание зависимости значений этих параметров от факторов, влияющих на результативность поиска. Одним из основных факторов, от которых зависит результативность поиска, является наличие структурных аналогов исследуемого соединения в БД. С одной стороны, кажется очевидным, что чем больше структурных аналогов в базе данных, тем вероятнее их появление в ПО, т.е. спектральные аналоги окажутся структурными аналогами. Как было показано по результатам поиска в БД по ИК спектроскопии [21] и масс-спектрометрии [22], действительно, вероятность появления структурных аналогов исследуемого соединения в поисковом ответе достаточно высока. С другой стороны, существует много причин, по которым в поисковый ответ в ряде случаев не отбираются даже ближайшие структурные аналоги.

С целью определения зависимости эффективности поиска от количества структурных аналогов в базах данных, для каждого из 2091 соединения при пороге ″неслучайности″ появления фрагментов в ПО 0,95 были проведены поиски всех структурных аналогов в БД по ИК спектроскопии и масс-спектрометрии. Было установлено, что для получения среднестатистических оценок *D* и *C*, указанных в табл. 1, в базах данных по ИК спектроскопии и масс-спектрометрии количество структурных аналогов со степенью структурного подобия ≥0,6 должно быть примерно 50 (для структурного подобия ≥0,8 — около пяти структур). Было также установлено, что в объединении ФС поисковых ответов достаточно присутствия четырех структурных аналогов со степенью структурного подобия 0,6 (2 структурных аналога для структурного подобия ≥0,8), чтобы оценки эффективности поиска D_v и C_v имели среднестатистические значения. Таким образом, при установлении строения наиболее критическим является, по-видимому, первый этап, в ходе которого в поисковый ответ отбираются спектральные аналоги, в ряде случаев не являющиеся структурными аналогами. При анализе же поискового ответа, который включает четыре или более структурных аналога, успешность установления строения исследуемого соединения почти гарантирована. Сто́ит отметить, что значительное увеличение частоты встречаемости структурного фрагмента в БД также снижает возможность его распознавания описанными методами, поскольку при этом сильно падает неслучайность появления фрагмента в структурах поискового ответа. Эти рекомендации могут быть полезными как при формировании БД, так и при предварительной оценке возможности анализа отдельных химических классов структур.

В рассмотренном выше примере установления структуры амикардина среди спектральных аналогов, полученных при поиске по ИК спектру, для структуры 4 (см. рис. 4) структурное подобие со структурой исследуемого соединения равно 0,81, для структуры 9 — 0,76. Для поиска по масс-спектру для структур 7 и 10 структурное подобие равно 0,97 и 0,72 соответственно, т.е. при поиске в двух базах данных нашлось 4 структурных аналога, структурные фрагменты которых внесли основной вклад в формирование результативной гипотезы. Анализ фрагментного состава поисковых ответов, а также наличие одной очень похожей структуры дало возможность определить строение соединения.

При поиске по спектру исследуемого соединения, как правило, не известно, присутствует ли искомое соединение или его структурные аналоги в используемой БД. В поисковый ответ соединения отбираются исключительно по спектральным характеристикам — факторам совпадения спектра исследуемого соединения и спектров, отобранных из БД. Использование значений этих факторов для предварительной оценки результативности установления строения может представлять практическую ценность.

С целью установления зависимости эффективности выявления фрагментов исследуемого соединения от степени спектрального подобия отобранных спектров (″эффективности поиска″), для каждого из 2091 соединения были определены величины параметров *D* и *C* и усредненные значения спектрального подобия по первым десяти спектральным аналогам, графики которых

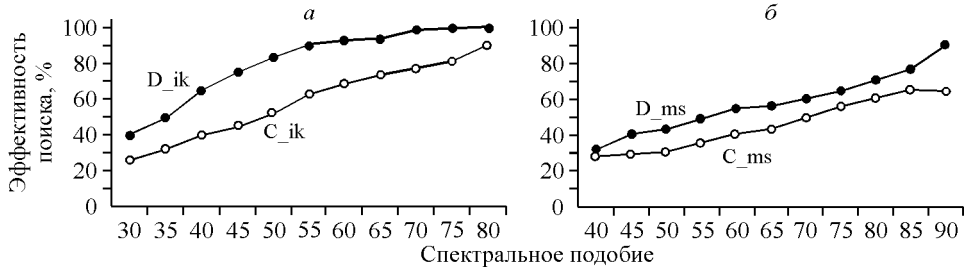

Рис. 7. Эффективность выявления фрагментов в результатах поиска в зависимости от спектрального подобия для ИК спектроскопии и масс-спектрометрии исследуемое соединение — амикардин

приведены на рис. 7. Средние значения спектрального подобия по всей выборке составляют 43,3 для ИК и 65,8 для масс-спектров.

Как видно из графика, среднестатистические значения *D* и *C* (см. таблицу) получаются при значениях спектрального подобия 40 для поисков по ИК спектрам и 65 по масс-спектрам. При более низких значениях спектрального подобия следует ожидать, что структурные формулы спектральных аналогов и их фрагменты будут малоинформативными. Тем не менее, как показано на рассмотренном выше примере (см. рис. 4), даже при значениях среднего спектрального подобия около 36 (по 10 первым ИК спектрам) и 48 (по 10 первым масс-спектрам) совместное использование ИК спектроскопии и масс-спектрометрии может быть успешным для целей установления структуры.

Другим важным параметром, на который следует обращать внимание при анализе поискового ответа, является значение неслучайности появления фрагментов. Как было исследовано [14, 20], повышение порога неслучайности появления фрагментов в структурах ПО может существенно улучшать результативность анализа. Так, на примере поиска амикардина, при повышении порога неслучайности появления фрагментов с 0,95 до 0,99 (см. рис. 5) удалось полностью избежать появления некорректных фрагментов в пересечении ФС.

ЗАКЛЮЧЕНИЕ

Таким образом, приведенные статистически обоснованные оценки и примеры демонстрируют эффективность предложенного метода установления строения, базирующегося на совместном использовании БД по ИК спектроскопии и масс-спектрометрии, а также на представлении структур соединений в виде полных фрагментных составов. В среднем в 2 раза увеличилась доля корректно распознанных фрагментов в пересечении фрагментных составов ПО в сравнении с результатами поиска по отдельным видам спектров. В среднем в 1,5 раза увеличилась доля корректно распознанных фрагментов в исследуемом соединении при использовании объединения фрагментных составов ПО в сравнении с результатами поиска по отдельным видам спектров. Ближайшие спектральные аналоги исследуемого соединения, соответствующие им структуры и, главным образом, структурные фрагменты, выявленные в результате поиска по различным видам молекулярных спектров, в большинстве случаев дают очень полезную информацию для установления строения исследуемого соединения.

Наибольший интерес представляют структурные фрагменты, встречающиеся одновременно в поисковых ответах по ИК и масс-спектрам, поскольку вероятность выявления случайных фрагментов в этом случае существенно уменьшается. Ценность представляют и фрагменты, присутствующие в объединении ФС, но не вошедшие в пересечение ФС. Такие фрагменты могут либо служить источником дополнительной информации для формирования гипотезы о строении соединения, либо подтверждать уже предложенную гипотезу. Структура спектрального аналога, имеющая максимальное число фрагментов, совпавших с фрагментами пересечения ФС, может использоваться как возможный близкий аналог исследуемого соединения, полезный для установления его строения.

Использование количественных оценок эффективности поиска в зависимости от значений спектрального подобия позволяет ограничивать размер поискового ответа, включая в него со-

единения, имеющие степень совпадения спектров не ниже заданного порога. По значениям неслучайности появления фрагментов в ПО можно предположить, какие из них могут быть более полезны для целей установления строения исследуемого соединения.

В отличие от большинства ранних работ, описывавших достоинства совместного использования нескольких спектральных методов для установления строения органических соединений на отдельных удачных примерах, в данной работе приводится объективное статистически обоснованное исследование, демонстрирующее количественные оценки результативности данного подхода.

СПИСОК ЛИТЕРАТУРЫ

1. *Gray N.A.B*. Computer-Assisted Structure Elucidation. – N. Y.: Wiley&Sons, 1986.
2. *Zupan J*. Computer-Supported Spectroscopic Databases. – Chichester: Ellis Horwood, 1986.
3. *Advanced* Chemistry Development/Labs (http://www.acdlabs.com/publish/reviews.html).
4. *Elyashberg M.E., Martirosian E.R., Karasev Yu.Z., Thiele H., Somberg H.* // Anal. Chim. Acta. – 1997. – **337**. – P. 265 – 286.
5. *Funatsu K., Sasaki S.* // J. Chem. Inf. Comput. Sci. – 1996. – **36**. – P. 190 – 204.
6. *Munk M.E.* // Ibid. – 1998. – **38**. – P. 997 – 1009.
7. *Hu C.Y., Xu L.* // Anal. Chim. Acta. – 1994. – **295**. – P. 127 – 134.
8. *Cadish D., Pretsch E.* // Ibid. – 1993. – **277**. – P. 297 – 304.
9. *Лебедев К.С*. // Журн. аналит. химии. – 1993. – **48**. – С. 851 – 863.
10. *Heller S.R*. // J. Chem. Inf. Comput. Sci. – 1985. – **25**. – P. 224 – 231.
11. *Kornakova T.A., Bogdanova T.F., Derendyaev B.G., Piottukh-Peletsky V.N.* // Anal. Chim. Acta. – 2005. – **543**. – P. 177 – 180.
12. *Steinbeck C*. // Nat. Prod. Rep. – 2004. – **21**. – P. 512 – 518.
13. *Пиоттух-Пелецкий В.Н., Чмутина К.С., Королевич М.В*. // Журн. структур. химии. – 2003. – **44**, №5. – С. 835 – 842.
14. *Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Чмутина К.С., Нехорошев С.А*. // Журн. аналит. химии. – 2002. – **57**, № 11. – С. 1176 – 1185.
15. *Demuth W., Karlovits M., Varmuza K*. // Anal. Chim. Acta. – 2003. – **490**. – P. 313 – 324.
16. *Demuth W., Karlovits M., Varmuza K*. // Ibid. – 2004. – **516**. – P. 75 – 85.
17. *Макаров Л.И. Скворцова М.И., Станкевич И.В. и др*. // Успехи химии. – 2006. – **75**, № 11. – С. 1074 – 1093.
18. *Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Шарапова О.Н*. // Журн. структур. химии. – 2000. – **41**, № 2. – С. 379 – 390.
19. *Пиоттух-Пелецкий В.Н., Богданова Т.Ф., Дерендяев Б.Г*. // Там же. – 1996. – **37**, № 2. – С. 368 –378.
20. *Piottukh-Peletsky V.N., Korobeinicheva I.K., Bogdanova T.F. и др*. // Anal. Chim. Acta. – 2000. – **409**. – P. 181 – 195.
21. *Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Корнакова Т.А*. // Химия в интересах уст. развития. – 2001. – **9**, № 1. – С. 17 – 26.
22. *Дерендяев Б.Г., Пиоттух-Пелецкий В.Н., Нехорошев С.А. и др*. // Журн. структур. химии. – 1999. – **40**, № 4. – С. 728 – 733.