ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ
Том 40, № 4 *Июль – август* *1999*

УДК: 543.51+543.42:681.32

Б.Г. ДЕРЕНДЯЕВ, В.Н. ПИОТТУХ-ПЕЛЕЦКИЙ, С.А. НЕХОРОШЕВ, С.П. КИРШАНСКИЙ, Т.Ф. БОГДАНОВА, Л.И. МАКАРОВ

О ПРИМЕНИМОСТИ БАЗЫ ДАННЫХ ТИПА МАСС-СПЕКТР — ФРАГМЕНТНЫЙ СОСТАВ ДЛЯ ВЫЯВЛЕНИЯ ОСОБЕННОСТЕЙ СТРОЕНИЯ ИССЛЕДУЕМОГО СОЕДИНЕНИЯ

Представление структур соединений в базе данных по масс-спектрометрии в виде исчерпывающего набора неизоморфных *k*-вершинных подграфов (фрагментов) открывает новые возможности при определении фрагментного состава соединений, исследуемых по масс-спектрам. На примерах анализа результатов поисков по масс-спектрам 300 соединений известного строения проанализированы зависимости распознавания корректных фрагментов от порога по неслучайности их появления в поисковом ответе.

Ранее мы показали [1], что представление структурных формул соединений базы данных (БД) по ИК спектроскопии в виде ″исчерпывающего набора фрагментных составов″ [2] оказывается весьма продуктивным для выявления информации о фрагментах исследуемого соединения. Представляет интерес распространить разработанный прием на случай масс-спектрометрии, которая также широко используется в исследовательской и прикладной химической практике. Этот метод чрезвычайно информативен, требует малого количества вещества, масс-спектрометры стыкуются с хроматографами, что дополнительно расширяет сферу использования метода. Компьютерные приемы извлечения информации из этого вида спектров весьма разнообразны [3—7]. Выявляемые особенности строения соединения охватывают молекулярный вес, элементный состав, крупные структурные фрагменты и более мелкие функциональные группы. Крупные фрагменты, как правило, представляют собой связанные блоки атомов, общие для структур отобранных в результате поиска соединений, имеющих спектры, наиболее похожие на предъявленный [8,9]. Более мелкие — список заранее заданных фрагментов, определение которых возможно с помощью соответствующего программного обеспечения [10].

Характерная особенность представления структурных формул (далее структур, молекулярных графов) в виде фрагментных составов заключается в их описании исчерпывающим набором всех входящих в структуру *k*-вершинных неизоморфных связных фрагментов ($k = 2 \div 7$). Это позволяет анализировать структуры отбираемых из БД соединений на все имеющиеся в их составе заранее не заданные фрагменты. Иными словами, открывается возможность распознавания практически любых фрагментов, входящих в структуры соединений БД. Используя этот прием на примере ИК спектроскопии, мы установили, что могут распознаваться тысячи [11] самых разнообразных скелетообразующих фрагментов. В настоящей работе поставлена задача: оценить перспективы применения аналогичного приема при распознава-

нии фрагментов исследуемого соединения в случае анализа с помощью соответствующей БД его масс-спектра низкого разрешения.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Описываемые ниже эксперименты проведены с использованием коллекции масс-спектрометрических данных NIST/EPA. При этом из основной коллекции сформирована экспериментальная БД, содержащая записи, в которых каждое соединение представлено только одной структурой и соответственно одним спектром. Далее декомпозицией каждого молекулярного графа средствами, описанными в [2], были построены полные наборы содержащихся в них неизоморфных связных k-вершинных фрагментов ($k = 2 \div 7$). Использование регистрационной системы [1] позволило связать записи, относящиеся к одному соединению: спектр, структура и фрагментный состав — и получить информацию о частотах встречаемости неизоморфных фрагментов в структурах экспериментальной БД.

Суть поставленных экспериментов сводится к следующему. Из базы данных с помощью датчика случайных чисел сформирована тестовая выборка, включающая ~300 записей типа спектр—структура. Далее, каждый спектр этой выборки последовательно предъявляли поисковой системе по масс-спектрометрии с целью отбора из БД спектров, наиболее похожих на предъявленный. При сопоставлении спектров использовали критерий спектрального подобия

$$MF1 = 2W_c /(W_x + W_r),$$

где W_c — параметр, характеризующий суммарную значимость совпавших спектральных признаков (массовые числа и интенсивности) сравниваемых спектров; W_x и W_r — параметры, характеризующие все спектральные признаки исследуемого спектра и спектра БД соответственно. (Подробнее об алгоритме поиска см. в работе [12].)

В поисковый ответ включали 11 первых ($MF1 \geq 35$) отобранных из БД спектров и структур, в том числе структуру и спектр ″неизвестного″. На последующих этапах анализа информацию о ″неизвестном″ удаляли из поискового ответа, т.е. моделировали условия новизны изучаемого соединения. Сведения о структуре и фрагментном составе ″неизвестного″ выступали далее как эталонные для сопоставления с соответствующими данными, характеризующими десять других соединений, отобранных в поисковый ответ. В частности, в каждом конкретном эксперименте оценивали соответствие структуре-эталону фрагментов этих структур.

Все эксперименты выполнены на ПК типа IBM PC 486 и Pentium-166.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Проиллюстрируем на примере характер информации о фрагментах исследуемого соединения, получаемой в результате анализа фрагментных составов структур соединений, отобранных из БД в поисковый ответ по предъявленному масс-спектру.

На рис. 1 приведен масс-спектр и структура эталона, а ниже в таблице — примеры k-вершинных фрагментов ($k = 3, 5, 7$), присутствующих во фрагментных составах 10 структур поискового ответа. Для каждого фрагмента в скобках приведена частота его встречаемости в отобранных структурах и через разделитель (косая черта) в структурах всех соединений БД. Отметим, что используемый способ представления структурных данных в виде k-вершинных связных фрагментов игнорирует

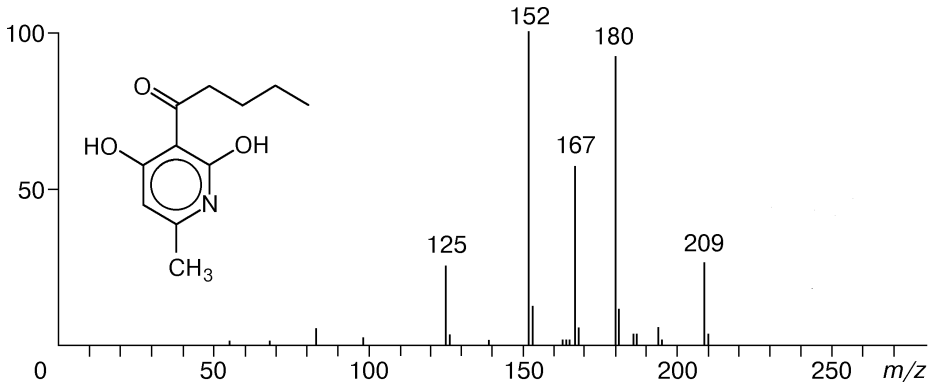
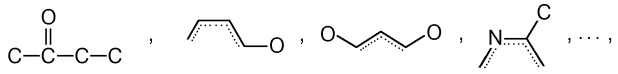
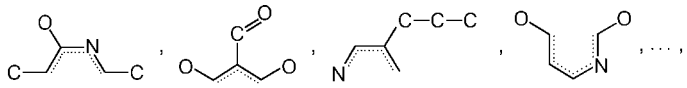
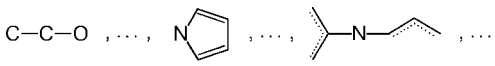


Рис. 1. Структура и масс-спектр исследуемого соединения

k	Корректные фрагменты
3	C—C=O (5/12199), O—C=C (3/4582), C—C=C (8/11830), C=N=C (3/2381), ...
5	 (2/2785), (3/220), (3/598), (3/273), ...
7	 (3/14), (3/83), (2/90), (3/70), ...
	Ошибочные фрагменты
3, 5, 7	 (3/1042), (3/68), (4/205), ...

П р и м е ч а н и е. Символом ⎯····· отмечены ароматические связи.

атомы водорода при описании фрагментов (ср. [1, 2]). В списках выявляемых фрагментов наряду с корректными (вложимыми в структуру-эталон) фрагментами присутствуют и некорректные фрагменты, т.е. отсутствующие в структуре исследуемого по спектру соединения. В данном случае число корректных фрагментов, выявленных в 10 структурах соединений поискового ответа, превышает число ошибочных приблизительно в три раза. Можно заметить, что корректные фрагменты (например, при k = 7), неоднократно перекрываясь своими компонентами, могут достаточно полно описывать структурные особенности скелета соответствующего эталона.

При обработке результатов поиска по масс-спектрам всех объектов тестовой выборки нас интересовало следующее. Во-первых — насколько полно выявляемые фрагменты описывают фрагментный состав изучаемого соединения; во-вторых — какова степень зашумленности списка выявленных фрагментов фрагментами, отсутствующими в соответствующих структурах-эталонах; в-третьих — зависимость этих

параметров от "неслучайности" [13] появления того или иного фрагмента в структурах соединений поискового ответа.

Оценим полноту представления в поисковом ответе фрагментов структуры-эталона как отношение числа корректных фрагментов (n_c) заданного размера, выявленных при анализе конкретного поискового ответа, к общему числу (n_e) фрагментов данного размера структуры-эталона и проследим по всем поискам (N = 300) как изменяется усредненная величина параметра D

$$D = 1/N_r \sum (n_c / n_e)$$

в зависимости от "неслучайности" появления соответствующих фрагментов в каждом конкретном эксперименте. Здесь N_r — число результативных поисков ($N_r \le N$). Поиск считается результативным, если в поисковом ответе присутствует хотя бы один фрагмент заданного размера, удовлетворяющий заданному пороговому значению параметра "неслучайности". В свою очередь "неслучайность" (NR) появления некоторого фрагмента среди фрагментов 10 структур поискового ответа определим по аналогии с работами [11, 13]:

$$NR = 1 - P(m)/P(10x),$$
$$P(z) = 10!(x)z(1 - x)(10 - z)/G(z + 1)G(10 - z + 1).$$

Здесь $z = m$ или $10x$; G — гамма-функция; $P(m)$ — вероятность того, что при случайном выборе структур из БД в выборке размером 10 окажется m структур, содержащих фрагмент с относительной частотой встречаемости в структурах БД, равной x. Очевидно, что параметр NR учитывает индивидуальные частоты встречаемости каждого фрагмента в БД и в поисковом ответе.

Величина параметра неслучайности позволяет на качественном уровне охарактеризовать достоверность решения об идентификации того или иного фрагмента, принимаемого по результату поиска. В первую очередь это относится к фрагментам, достаточно часто представленным в структурах соединений БД: их появление в структурах поискового ответа может носить случайный характер. С другой стороны, требование излишне высокой пороговой величины параметра NR, которой должны удовлетворять фрагменты, рассматриваемые как идентифицированные, может сопровождаться [11] значительным уменьшением числа сохраняемых в списке фрагментов.

На рис. 2 приведены зависимости параметра D от величины заданного порогового значения параметра NR для всех типов (k = 2÷7) рассматриваемых фрагментов. При этом анализе не контролировалось согласование выявляемых по поисковому ответу фрагментов с брутто-формулой соответствующего эталона.

Как видно, при пороговом значении параметра NR = 0,0 (фрагмент считается всегда "идентифицированным", если он появился хотя бы один раз в структурах соединений поискового ответа) для значений $k \le 6$ распознается свыше 50 % фрагментов структуры-эталона. Однако значительная часть распознаваемых в этих условиях фрагментов может носить случайный характер. Вероятно, более достоверными следует признавать результаты, достигаемые при значениях $NR > 0{,}8$.

Иная картина (рис. 3) характеризует долю корректных фрагментов (C) среди всех выявляемых в поисковом ответе. Этот параметр оценивался следующим образом:

$$C = 1/N_r \sum (n_c / n),$$

где n — общее число фрагментов данного размера с неслучайностью выше заданного порога.

Легко видеть из рис. 3, что списки выявляемых k-вершинных фрагментов при $k = 5 \div 7$ довольно сильно зашумлены. С ростом параметра неслучайности наблюдается лишь незначительное увеличение доли корректных фрагментов (ср. с аналогичной зависимостью для ИК спектроскопии [11]). Это может быть обусловлено как природой самого метода масс-спектрометрии, так и выбранным алгоритмом поиска ближайших спектральных аналогов.

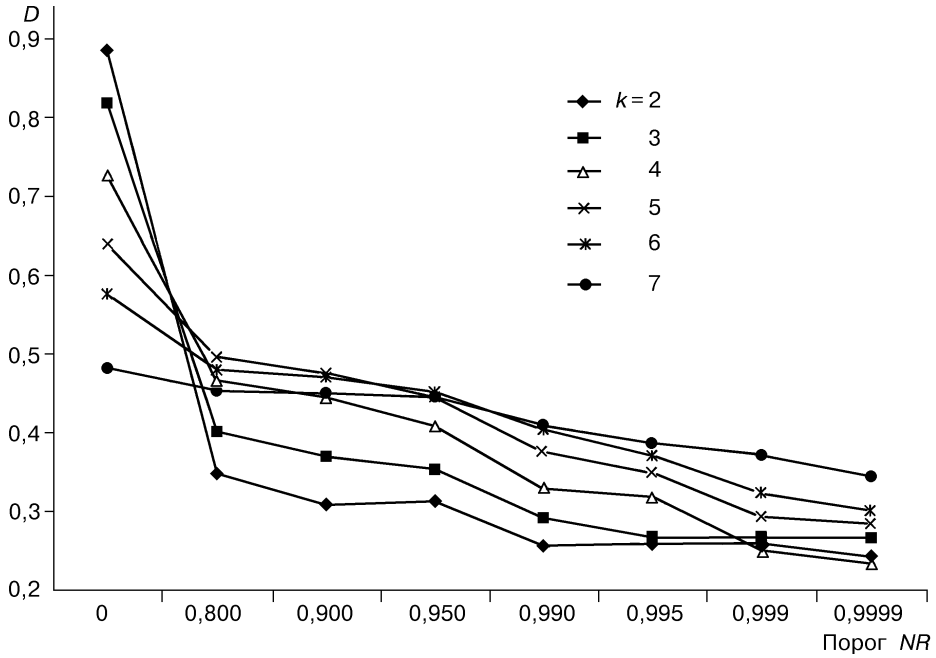


Рис. 2. Доля фрагментов, распознаваемых в структуре-эталоне в зависимости от порога неслучайности

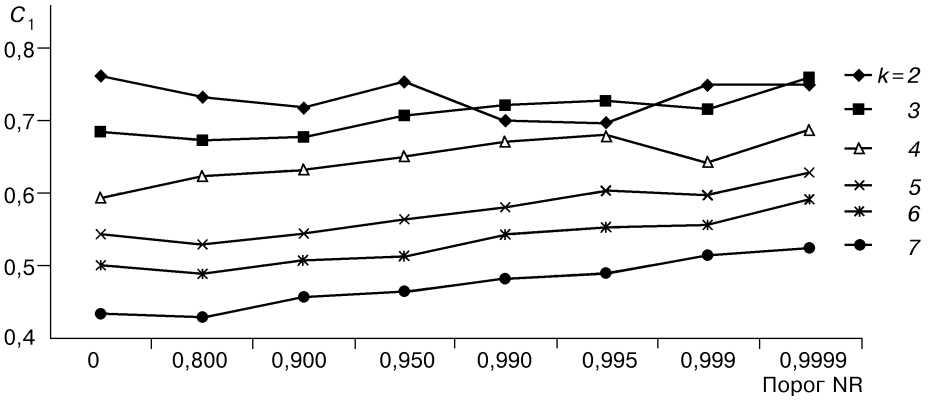


Рис.3. Доля корректных фрагментов в поисковом ответе в зависимости от порога неслучайности

Поведение зависимостей (см. рис. 2 и 3) практически не изменяется в условиях, когда в каждом конкретном эксперименте в списке выявленных фрагментов сохра-

няются только те, которые не противоречат брутто-формуле анализируемого соединения. Параметры *D* и *C* в этих случаях возрастают на 0,02—0,03 ед., что, вероятно, отражает выявленную ранее селективность масс-спектрального поиска по отношению к элементному составу исследуемого соединения [14].

Анализ результатов поисков по 300 масс-спектрам "отсутствующих" в БД соединений свидетельствует о возможности использования рассматриваемого подхода при определении фрагментов изучаемого соединения. Список распознаваемых в его рамках фрагментов заметно зашумлен, но, наряду с этим, в нем может содержаться до 50 % неслучайно выявленных фрагментов исследуемого соединения. Поэтому выяснение типов и вероятностей распознавания различных фрагментов, а также сопоставление этих данных с аналогичными для случая ИК спектроскопии представляется необходимым в дальнейших исследованиях. Как известно, только на основе непротиворечивой совокупности получаемых из различных видов спектроскопии сведений можно формулировать аргументированное заключение о строении исследуемого соединения.

Авторы благодарят С. Стейна (Национальный институт стандартов и технологии, США) за предоставленную возможность использования базы масс-спектрометрических данных, Российский фонд фундаментальных исследований и Сибирское отделение РАН за поддержку данной работы, а также К.С. Чмутину за содействие ее выполнению.

СПИСОК ЛИТЕРАТУРЫ

1. *Пиоттух-Пелецкий В.Н., Дерендяеев Б.Г., Молодцов С.Г., Богданова Т.Ф.* // Журн. структур. химии. – 1997. – **38**, № 4. – С. 785 – 794.
2. *Пиоттух-Пелецкий В.Н., Смирнов В.И., Румянцев А.К., Дерендяев Б.Г.* // Сиб. хим. журн. – 1993. – **3**. – С. 65 – 73.
3. *Gray N.A.B.* Computer-Assisted Structure Elucidation. – N.Y.: Wiley, 1986. – 536 P.
4. *Warr W.A.* // Anal. Chem. – 1993. – **65**. – P. 1045a – 1047a, 1087a – 1095a.
5. *Haraki K.S., Venkataraghavan R., McLafferty F.W.* // Ibid. – 1981. – **53**. – P. 386 – 392.
6. *Martinsen D.P., Song B.H.* // Mass Spectrom. Review. – 1985. – **4**. – P. 461 – 490.
7. *Лебедев К.С., Дерендяев Б.Г.* // Химия в интересах устойч. развития. – 1995. – **3**. – С. 269 – 285.
8. *Лебедев К.С., Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Коптюг В.А.* // Изв. СО АН СССР. Сер. хим. наук. – 1982. – **1**. – С. 105 – 114.
9. *Dayringer H.E., Pesyna G.M., Venkataraghavan R., McLafferty F.W.* // Org. Mass Spectrom. – 1976. – **11**. – P. 529 – 542.
10. *Stein S.F.* // J. Amer. Soc. Mass. Spectrom. – 1995. – **6**. – P. 644 – 655.
11. *Пиоттух-Пелецкий В.Н., Дерендяев Б.Г., Богданова Т.Ф.* // Журн. структур. химии. – 1997. – **38**. – С. 369 – 379.
12. *Лебедев К.С., Киршанский С.П., Нехорошев С.А., Дерендяев Б.Г.* // Журн. аналит. химии. – 1987. – **42**. – С. 1320 – 1331.
13. *Вентцель Е.G.* Теория вероятностей. – М.: Физматгиз, 1963. – С. 58.
14. *Нехорошев С.А., Киршанский С.П., Дерендяев Б.Г.* // Изв. СО АН СССР. Сер. хим. наук. – 1986. – **6**. – С. 113 – 122.

Новосибирский институт органической химии им. Н.Н. Ворожцова СО РАН
Институт математики им. С. Л. Соболева СО РАН
Новосибирск

Статья поступила 11 июня 1998 г.

E-mail: bogd@nioch.nsc.su