ОБЗОРЫ

УДК: 543.51+681.3

ИНФОРМАЦИОННО-ЛОГИЧЕСКАЯ СИСТЕМА ХИМАРТ. ОТ МАСС-СПЕКТРА К СТРОЕНИЮ ОРГАНИЧЕСКОГО СОЕДИНЕНИЯ

Б.Г.Дерендяев, И.И.Строков, К.С.Лебедев*
Новосибирский институт органической химии им.Н.Н.Ворожцова СО РАН, 630090, Новосибирск, пр. Лаврентьева, 9
strokov@vmk.ru
**Новомосковский институт РХТУ им. Д.И.Менделеева*
Новомосковск Тульской обл., Дружбы, 8

Поступила в редакцию 19 ноября 2003 г., после исправления – 25 марта 2005 г.

Анализируется возможность установления строения органических соединений по масс-спектрам низкого разрешения с помощью информационно-логической системы ХимАрт. Обсуждается экспериментальный материал, полученный на основе решения более 100 задач.

***Дерендяев Борис Григорьевич** – руководитель отдела Научно-технический центр по химической информатике НИОХ СО РАН, заведующий лабораторией, доктор химических наук, профессор.*

***Область научных интересов** – физическая органическая химия, аналитическая химия, химическая информатика, молекулярная спектроскопия, компьютерные методы анализа физико-химических данных.*

Автор более 190 печатных работ.

***Строков Игорь Иванович** – научный сотрудник лаборатории программного обеспечения диалоговых систем в химии Новосибирского института органической химии им. Н.Н. Ворожцова Сибирского отделения РАН (НИОХ СО РАН), кандидат химических наук.*

***Область научных интересов** - компьютерные технологии, теория графов, методы установления строения органических соединений по спектральным данным, информационные технологии в химии.*

Автор 30 печатных работ.

***Лебедев Константин Сергеевич** – профессор кафедры аналитической химии Новомосковского института РХТУ им. Д.И. Менделеева, доктор химических наук.*

***Область научных интересов** – компьютерные методы установления строения органических соединений по спектральным данным (МС, ИК, ЯМР), поисковые и экспертные системы, информационные технологии в химии.*

Автор 80 печатных работ.

Установление строения органических соединений является одной из массовых задач химической практики. Ее решению способствует развитие инструментальной базы молекулярной спектроскопии и вычислительной техники. Однако во многих случаях анализ спектральной информации с целью извлечения сведений о строении соединения остается достаточно трудоемким и требует участия специалистов высокой квалификации. Особенно это характерно для масс-спектрометрии и хромато-масс-спектрометрии. В последнем случае при анализе многокомпонентных смесей в ходе одного эксперимента могут быть получены десятки-сотни спектров. Существующие средства идентификации соединений с помощью систем на основе баз данных (БД) эффективны лишь при анализе ранее изученных и представленных в БД своими спектрами веществ. Поэтому не ослабевает интерес к созданию компьютерных средств, оказывающих помощь при интерпретации спектральных данных вновь синтезированных или выделенных из природного сырья соединений.

Развитие компьютерных методов и систем идет по двум направлениям, связанным с «непрямым» и «прямым» [1] использованием накопленного спектрального материала. Непрямые методы (искусственный интеллект, распознавание образов, нейронные сети и т.п. [2, 3]) опираются или на известные корреляционные зависимости, характеризующие спектральное поведение определенных функциональных групп и классов химических соединений, вложенные в соответствующие базы знаний, или на математические при-

емы, позволяющие выявлять эти знания. Прямые методы используют информацию, содержащуюся в базах данных вида «спектр–структура соединения». Среди последних особый интерес представляет развитие комплексных (мультиспектральных, интегрированных) систем, оперирующих с экспериментальными данными по нескольким видам молекулярных спектров [4-13].

В работах [14-15] достаточно подробно изложена идеология разрабатываемой нами комплексной информационно-логической системы, названной ХимАрт (химическая артель). Опыт ее использования при анализе данных ЯМР спектроскопии и масс-спектрометрии показал, что с помощью ХимАрт можно устанавливать строение достаточно сложных органических соединений. При этом масс-спектрометрическая компонента системы выполняла лишь вспомогательную функцию: отбор из БД соединений, обладающих масс-спектрами, подобными спектру исследуемого соединения. Моделирование спектров ^{13}C-ЯМР для отобранных соединений и сопоставление модельных с экспериментальным спектром ^{13}C-ЯМР, позволяло выявлять структурные фрагменты неизвестного соединения, согласующиеся с двумя видами экспериментальных данных.

Вместе с тем хорошо известно, что на основе компьютерного анализа результатов поиска в масс-спектрометрической БД можно решать и более сложные задачи. Так, например, в ряде работ продемонстрирована возможность определения молекулярной массы и молекулярной формулы [17-19], выявление структурных особенностей изучаемых соединений [19-21], включая крупные связные фрагменты, генерирование структурных гипотез и предсказание спектров [22-24]. Поэтому расширение функциональных возможностей системы ХимАрт путем включения в состав масс-спектрометрической компоненты процедур (приложений), позволяющих решать подобные перечисленным выше задачи, способствовало бы росту ее потенциала в практике спектрального анализа.

В данной работе рассматривается масс-спектрометрическая компонента информационно-логической системы ХимАрт и приведены результаты ее использования при анализе масс-спектров более 100 различных органических соединений с целью выявления их структурных особенностей.

Методы хранения, поиска и обработки информации

База данных системы ХимАрт содержит ~52000 масс-спектров низкого разрешения, полученных при ионизации молекул пучком электронов. Исходно спектры представлены в виде набора пиков, для каждого из которых указаны интенсивность относительно максимального пика (1000 ед.) и целочисленное значение величины отношения массы к заряду (m/z). В базе данных системы запись спектра состоит из трех частей. Первая описывает только интенсивные (1000–64) пики с точностью до 64 единиц и представляет собой последовательность байт, в каждом из которых старшие 4 бита задают интенсивность пика, округленную на 64 единицы, а младшие 4 бита — расстояние от предыдущего пика в шкале m/z. Вторая часть записи содержит «остаток» спектра за вычетом интенсивных пиков, записанных в первой части. В каждом байте этой записи 6 битов отведено для интенсивности, а 2 — для расстояния между пиками исходного спектра. Третья часть записи содержит обобщенные данные о спектре: массу молекулярного иона и общий ионный ток. Такой способ описания спектра требует более чем в три раза меньший объем памяти по сравнению с исходным представлением. Заметим также, что кодирование интенсивных пиков позволяет быстро проводить оценочные расчеты похожести спектров БД на спектр запроса и проверку граничных условий поиска без восстановления его исходного вида. Это значительно экономит время поиска в БД спектров, максимально похожих на спектр изучаемого соединения.

Для каждого соединения в БД, кроме масс-спектра, хранится его структурная формула, молекулярная формула (брутто-формула) и название. Информация разного типа (спектр, структура, текст) разнесена по отдельным таблицам, записи в которых связаны уникальным для каждого соединения идентификатором (ключом).

Структурная формула соединения (молекулярный граф) представлена кодом, особенности которого достаточно подробно изложены в работах [23, 25]. Здесь отметим только, что код соответствующего молекулярного графа почти полностью образуется из меток (типов) его вершин, перечисляемых при обходе графа в глубину или в ширину. Поскольку типы химических связей в большинстве случаев определяются типами связных атомов, то такое кодирование графов оказывается компактным. Канонические коды структур позволяют реализовать эффективный структурный и подструктурный (по части структурной формулы) поиск. (Здесь и далее, для краткости, термин «структура» — синоним структурной формулы соединения, «фрагмент» — часть структурной формулы).

Особенности используемой файловой системы JoKey [26] и «классификатора» структур (индексный файл «широких» кодов структур) [25] позволяют наряду с подструктурным поиском реализовать поиск структурных аналогов соединения заданного строения. Эта функция системы важна для получения справочной информации о наличии в БД записей о соединениях с заданными структурными параметрами и особенностях поведения их спектров. Поисковый запрос в данных случаях может включать структурную формулу или фрагмент структуры, в том числе с указанием альтернативных атомов или возможных заместителей у вершин фрагментов. Отбор соединений по заданным структурным параметрам выполняется в системе как самостоятельная или как вспомогательная поисковая задача в составе других более сложных процедур (см. ниже).

Поиск в БД по спектральным параметрам основан на хорошо зарекомендовавших себя алгоритмах [17]. Он позволяет: сопоставлять пики ионов в шкале массовых чисел m/z, (поиск А); в шкале формальных первичных потерь $\Delta m = M^+ - m/z$ (поиск В); в обеих шкалах (поиск АВ); учитывать обязательные и желательные пики; включает прямой (X→R), обратный (R→X) и комбинированный (X→R, R→X) режимы сравнения спектров. В процедуре поиска предусмотрена возможность «накопления» результатов для формирования из нескольких видов поиска одного поискового ответа, задание ограничений на минимальное значение фактора спектрального подобия, числа соединений, отбираемых в поисковый ответ и т. п. (рис. 1). Поисковый ответ (ПО) представляет собой ранжированный по мере убывания фактора спектрального подобия (максимальное значение = 100 %) список ключей соединений БД, по которым пользователь может просмотреть структурные формулы и масс-спектры соответствующих соединений (рис. 1). Отметим, что при поиске спектров, подобных заданному, не используется техника индексации (инверсированные файлы), тем не менее, время поиска в полной БД не превышает 1 секунды для компьютеров с частотой процессора 166 Мгц и выше.

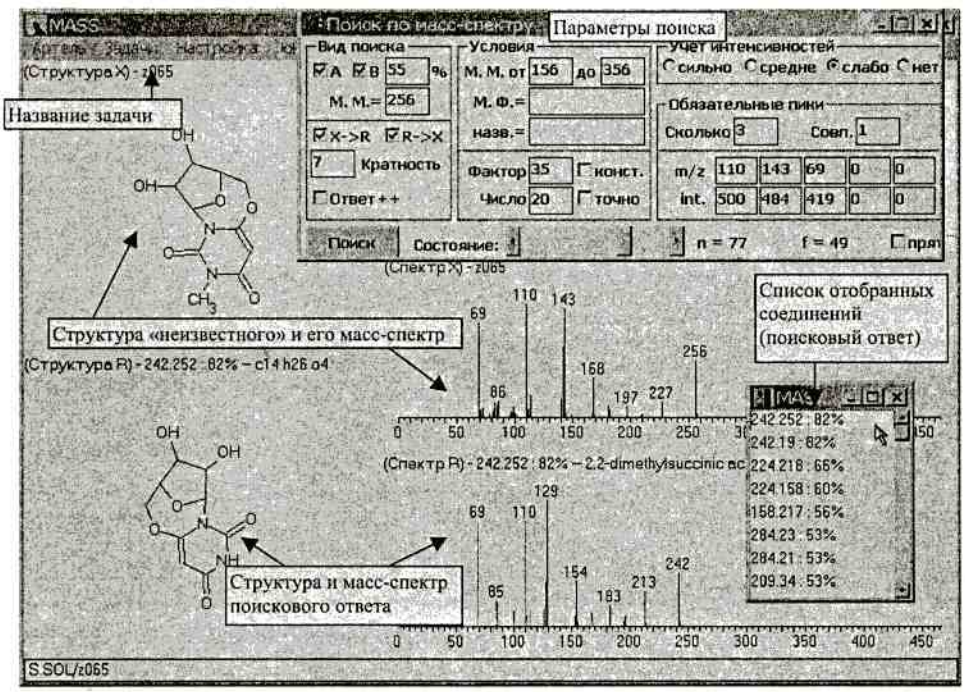


Рис. 1. Внешний вид ХимАрт при работе с процедурой «Спектральный поиск»

Наряду с указанными поисковыми процедурами в состав рассматриваемой компоненты ХимАрт входят также и процедуры обработки результатов поиска с целью решения более сложных задач. Опишем их кратко.

«Построение среднего спектра». Эта процедура вычисляет среднее арифметическое значение интенсивностей пиков ионов и потерь в анализируемом множестве (выборке) спектров. При этом рассматриваются только пики с интенсивностью не ниже некоторого порогового значения и с достаточно большой частотой встречаемости (например, частота 50 % означает, что пик или потеря должны встретиться не менее чем в половине анализируемых спектров). Средние спектры, построенные в шкалах m/z и Δm, могут быть выведены как самостоятельные или объединены в общий для абсолютной и относительной

шкал масс-спектр. Очевидно, что эта процедура, например, в сочетании с процедурой подструктурного поиска позволяет выявлять спектро-структурные корреляции, если таковые проявляются для заданного пользователем фрагмента. Она же с небольшими изменениями используется при попытках моделирования масс-спектра соединения по его структурной формуле. В этом случае множество усредняемых спектров формируется путем отбора из БД структурных аналогов заданного соединения [25].

«Пересечение структур». Эта процедура позволяет выделять все максимально общие неизоморфные фрагменты, присутствующие в структурах соединений поискового ответа путем попарного их пересечения.

Предполагается, что появление крупных фрагментов, общих для двух и более структур ПО, не случайно, и такие фрагменты, вероятно, присутствуют в структуре неизвестного. Для ранжирования общих фрагментов можно использовать размер фрагмента, число его повторений в структурах соединений ПО, произведение этих величин (чем больше фрагмент и чем чаще встречается в ответе, тем больше ему доверия [27]). В данной версии системы наряду с указанными параметрами используется фактор соответствия «спектрального отклика» фрагмента спектру неизвестного соединения. В качестве «отклика» выступает средний спектр, получаемый для соединений ПО, содержащих данный фрагмент. Далее этот «отклик» сравнивается со спектром неизвестного соединения. Рассчитываемый при этом фактор спектрального подобия используется для ранжирования списка фрагментов наряду с их размерами и частотой (см. рис.2, переключатели в поле «Сортировка»). Детали алгоритма приведены в работах [24, 26].

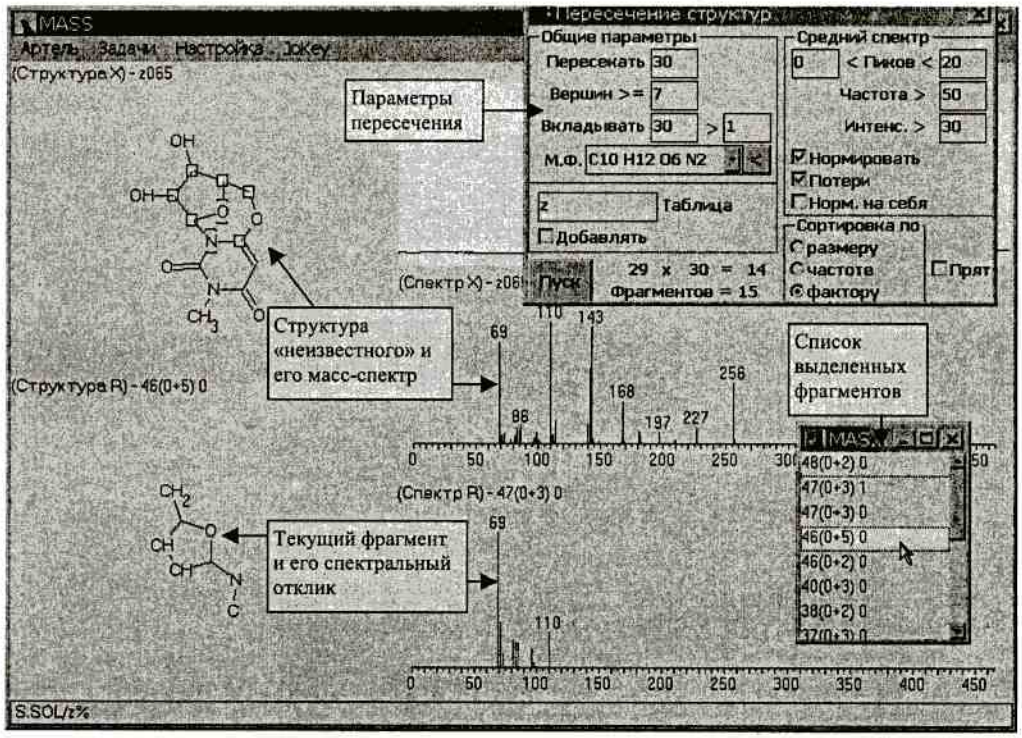


Рис. 2. Внешний вид ХимАрт при работе с процедурой «Пересечение структур». Параметры процедуры: число анализируемых структур ПО – 30, минимальный размер выделяемых фрагментов – 7 связных вершин, фрагмент общий не менее чем для 2-х структур. В структуре «неизвестного» выделены вершины, отвечающие вложению в нее выбранного фрагмента. В спектральном отклике фрагмента помечены пики, имеющиеся в спектре «неизвестного»

«Генерация структур» — пользовательский вариант ранее опубликованной программы [28]. Предоставляет удобные средства для задания исходных данных, сохранения и просмотра результатов генерации в графическом виде. Эта компонента получает данные из других процедур системы ХимАрт и передает в них свои результаты. Она используется автономно или как составная часть более сложных процедур (примеры см. ниже).

«Фрагментация структур». Процедура выполняет выделение из структур ПО фрагментов, формально подтвержденных простой моделью схемы первичной фрагментации, и построение на основе выделенных фрагментов возможных структурных формул «неизвестного».

На первом этапе (по команде «осколки») для каждой структуры ПО рассматриваются все способы ее декомпозиции («распада») на два «осколка» путем разрыва одной или двух связей, причем некоторые виды связей могут быть запрещены для разрыва. Каждый из осколков структуры рассматривается далее, если в масс-спектре данного соединения из ПО присутствует пик иона или первичной потери с массой, близкой к массе «осколка». Допустимое отличие масс изменяется параметрически от 0 (точное совпадение) до ±2, что симулирует возможную миграцию одного или двух атомов водорода при распаде молекулярного иона. Пик иона или потери считается «объясненным» всеми осколками, которые соответствуют ему по массе с наименьшим различием. Например, осколок, масса которого отличается от m/z пика на 1, не объясняет этот пик, если есть другой осколок с точным совпадением массы. В итоге, спектральный отклик осколка образуют все «объясненные» им пики в сумме с откликами осколков, вкладывающихся в данный. Эта операция повторяется для всех структур и спектров ПО. В результате формируется ранжированный по мере убывания фактора совпадения спектрального отклика со спектром неизвестного соединения (максимальное значение = 1000 ед.) список фрагментов, который выводится на экран вместе с их спектральными откликами (рис. 3).

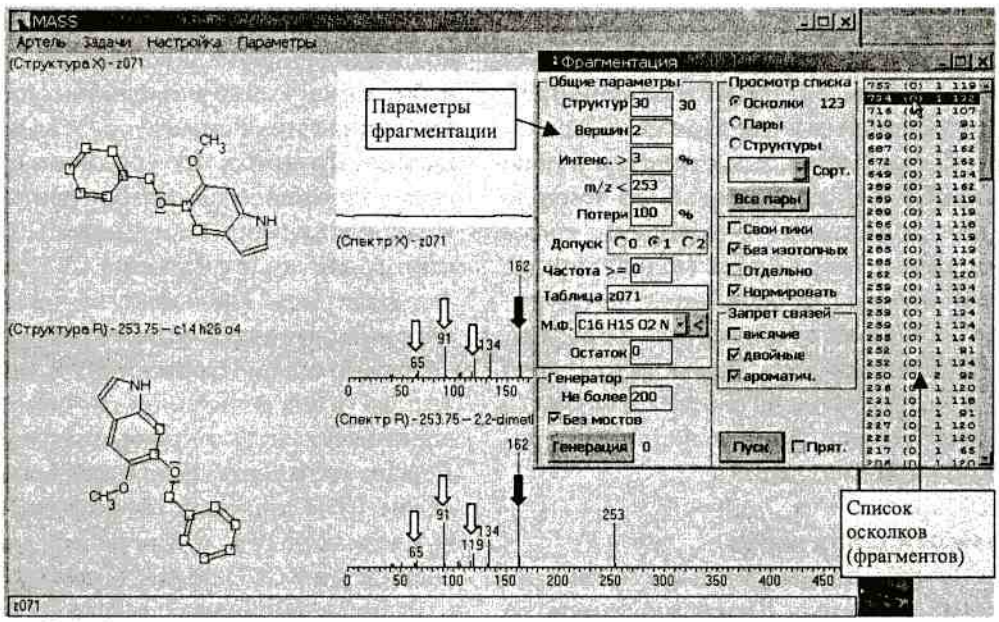


Рис. 3. Внешний вид ХимАрт при работе с процедурой «Фрагментация структур» (этап выделения фрагментов). Стрелками помечаются совпадающие сигналы в заданном и отобранном спектрах (пики ионов — белыми, потери — черными). Фрагменты, формально отвечающие за данные пики, помечены на структурных формулах. В левой части окна «Фрагментация» заданы параметры процедуры. В правой - приведен список полученных фрагментов; на следующих этапах анализа он заменяется на список пар фрагментов и генерированных структур

На втором этапе (команда «пары») из полученного списка фрагментов формируется ранжированный список пар фрагментов, удовлетворяющих точно или частично (параметр «остаток» = 0, 1, 2 и т.д.) заданной брутто-формуле соединения. Эта информация используется на третьем этапе (команда – «структуры) для генерации и ранжирования вероятных структурных формул изучаемого соединения. При ранжировании используются факторы спектрального подобия, участвующие в процессе генерирования фрагментов.

В отличие от процедуры «пересечения структур», осколки (фрагменты), полученные при моделировании распада молекул, имеют всего одну или две свободные связи, что упрощает задачу генерирования структурных гипотез. Кроме этого, оправдано одновременное задание на генерацию нескольких осколков из-за малой вероятности их пересечения. Последнее объясняется как способом их получения (деление структур на две части), так и предварительным вложением мелких осколков в более крупные. Более подробное описание этой процедуры см. в работах [24, 26].

Окна системы. На рис. 1-3 демонстрируется внешний вид системы при использовании трех ее основных процедур: «Спектральный поиск», «Пересечение структур» и «Фрагментация структур». Во всех случаях спектр неизвестного соеди-

нения и предполагаемая структурная формула располагаются *в верхней части* основного окна процедур. Информация, содержащаяся в окнах параметров, индивидуальна для каждой процедуры. Так, например, при спектральном поиске (см. рис.1) в окне «Параметры поиска» указано, что в данном случае необходимо провести комбинированный поиск в режиме АВ среди спектров соединений с молекулярными массами от 150 до 350 а.е.м., причем в отбираемых спектрах должен присутствовать по крайней мере один пик из трех наиболее интенсивных в спектре запроса.

Результаты работы процедур размещаются *в нижней части* основного окна. На рис. 1 (случай спектрального поиска) отображены структурная формула и соответствующий масс-спектр соединения из поискового ответа.

В поле результатов процедуры пересечения структур помещаются фрагмент и полученный для него спектральный отклик. При использовании процедуры фрагментации структур в поле результатов выводятся выявленные фрагменты со своими спектральными откликами, пары фрагментов, удовлетворяющие молекулярной формуле анализируемого соединения, и генерированные на их основе структурные формулы. Текущий результат (осколки, пары, структуры) выбирается пользователем из соответствующего ранжированного списка, расположенного в правой части окна «Фрагментация».

Экспериментальная часть

Эффективность использования масс-спектрометрической компоненты ХимАрт оценена на примерах выявления структурных особенностей около 120 различных органических соединений. Тестовая выборка содержала соединения с молекулярными массами (ММ) от 160 до 260 а.е.м. и обобщенной молекулярной формулой вида:

$$C_{2\text{-}17}\ H_{2\text{-}34}\ O_{0\text{-}6}\ N_{0\text{-}5}\ S_{0\text{-}4}\ Si_{0\text{-}2}\ F_{0\text{-}7}\ Cl_{0\text{-}4}\ Br_{0\text{-}1}\ J_{0\text{-}1}.$$

Выборку формировали из записей БД путем случайного выбора соединений в диапазоне масс 160–260, имеющих «степень структурного представительства (ССП)» не менее 50 %. ССП соединения является количественной мерой того, насколько хорошо оно представлено в базе данных своими структурными аналогами. Она оценена следующим образом.

Для каждой вершины текущего молекулярного графа БД строился канонический широкий код и с помощью классификатора структур [25] отбирались 4 наиболее близких (по числу совпавших начальных символов, *s*) кода из классификатора. Принято, что среднее *s* для 4-х ближайших кодов классификатора составляет ССП для данной вершины. Усреднение ССП для всех вершин графа дает значение ССП для рассматриваемого графа. Поскольку каждый символ широкого кода отвечает одному ребру графа, пороговое значение ССП, равное 50 %, можно интерпретировать как возможность описать не менее половины структурной формулы соединения с помощью его аналогов из базы данных. Распределение ССП для всех соединений базы данных приведено на рис. 4. Как видно, для большинства соединений ССП лежит в диапазоне 60–80 %, а ССП менее 50 % имеют только ~17 % соединений.

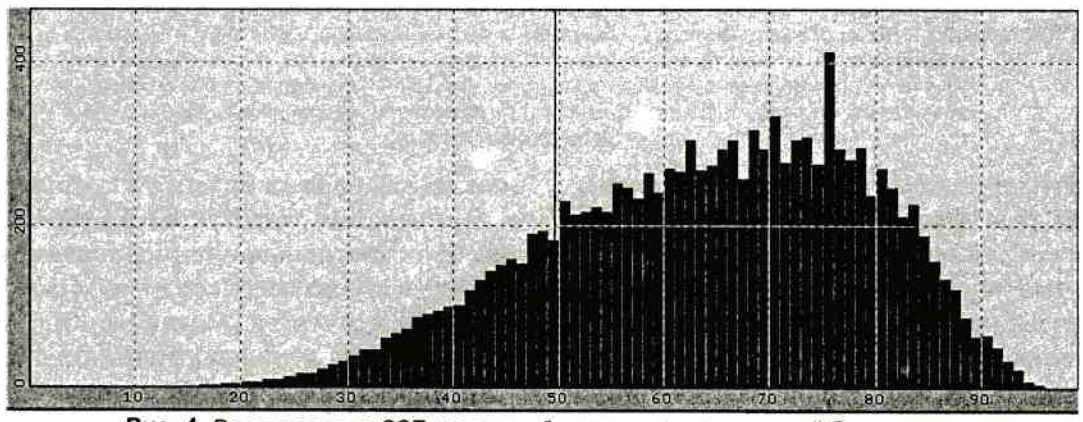


Рис. 4. Распределение ССП соединений масс-спектрометрической базы данных

На первом этапе экспериментов для каждого соединения выборки проводили поиск по его масс-спектру. Запись, характеризующую заданное соединение, удаляли из поискового ответа, моделируя случай его отсутствия в базе данных. Далее проводили анализ оставшихся в ПО записей. Результативность процедуры «пересечения структур» оценивали по наличию корректных (то есть вкладывающихся в структуру анализируемого «неизвестного») фрагментов среди первых пяти в ранжированном списке фрагментов.

На этапе выполнения процедуры «фрагмента-

ция структур» оценивали, удалось ли на стадии генерации сформировать список молекулярных графов, присутствует ли в ранжированном списке искомый граф, и если присутствует, то на каком месте.

Структурную формулу «неизвестного» использовали как эталон для оценки достигаемого результата. Этому же способствовали средства визуализации вложения фрагмента или генерируемой структуры в структуру «неизвестного».

Для всех объектов тестовой выборки условия спектрального поиска и анализа его результатов были унифицированы (табл. 1).

Таблица 1

Основные параметры процедуры спектрального поиска и анализа ее результатов

Поиск по спектру	Пересечение структур	Фрагментация структур
Вид поиска: по ионам и потерям (поиск АВ), с вложением сравниваемых спектров друг в друга (комбинированный режим сравнения), учет интенсивностей сигналов средний или слабый. **Ограничения:** диапазон ММ = ММ «неизвестного» ±100, фактор совпадения ≥ 35 % при обязательном совпадении одного из трех самых интенсивных пиков (рис. 1).	Попарно пересекать первые 15 структур ПО, а для построения среднего спектра использовать первые 20 структур. Число вершин во фрагментах — не менее половины от числа вершин в «неизвестном». **Средний спектр**: включать не более 20 пиков и потерь с интенсивностью не менее 3 % от основного пика и встречающихся не менее чем в половине анализируемых спектров.	Анализировать первые 15 структур ПО. Размер фрагмента (число связных вершин) ≥ 2. Запрет на разрыв двойных связей и связей в ароматическом кольце вдали от заместителя. **Генератор:** строить не более 200 структур из 1-го набора данных, отсеивать химически неправдоподобные структуры.

Результаты и обсуждение

Как и другие информационные системы по масс-спектрометрии, ХимАрт решает задачу поиска в БД спектров, тождественных заданному в запросе. На этой основе возможна идентификация соединения, если оно представлено в базе данных своим спектром.

При отсутствии в базе спектра искомого соединения с помощью системы можно отыскать спектры соединений, подобные заданному, и провести автоматизированный анализ результата поиска с помощью описанных процедур «пересечение» или «фрагментация». Это позволяет во многих случаях выносить суждения об особенностях строения изучаемого соединения.

На примере анализа поискового ответа, полученного по масс-спектру соединения I (код задачи z065), рассмотрим результат использования процедуры «пересечение». Анализ проведем в условии удаления из ПО сведений о «неизвестном» и параметрах процедуры, представленных на рис. 2.

В этих условиях выявлено 16 различных, содержащих не менее 7 связных вершин, общих фрагментов. Из них 6 фрагментов корректны, т.е. изоморфно вкладываются в структурную формулу «неизвестного» I. Причем четыре из них находятся среди первых пяти ранжированного списка фрагментов.

В качестве примера в табл.2 приведены примеры выявленных фрагментов и полученные для них спектральные отклики. Видно, что спектральные отклики фрагментов различаются друг от друга. Как правило, чем лучше спектральный отклик соответствует спектру анализируемого вещества, тем больше доверия соответствующему фрагменту.

В условиях пересечения структур, представленных в табл. 1, выявлено 3 фрагмента, содержащих не менее 9 связных вершин, один из которых корректен.

В целом при граничных условиях процедуры «пересечения», приведенных в табл. 1, для объектов тестовой выборки получена следующая картина распознавания фрагментов.

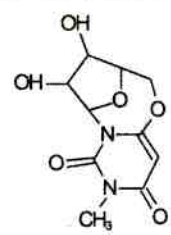
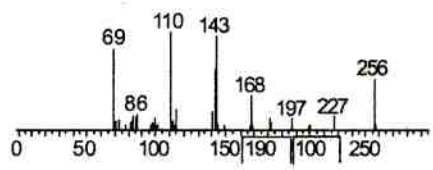
При ранжировании фрагментов в соответствии с их спектральными откликами, учитывающими массы осколочных ионов и формальных потерь, найдено, что в 73 % случаях корректный фрагмент (один или несколько) находится в списке первых пяти фрагментов, в 46 % случаев — на первом месте этого списка. Ранжирование фрагментов с учетом только их частот встречаемости в структурах соединений ПО менее предпочтительно. В этом случае на первых местах списка чаще оказываются общие фрагменты, соответствующие заданному порогу по числу связных вершин во фрагментах пересекаемых структур.

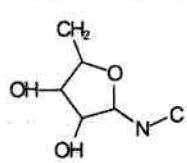
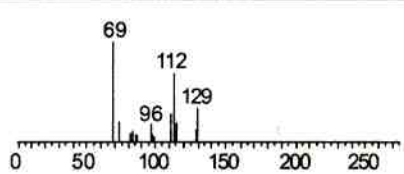
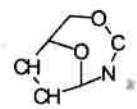
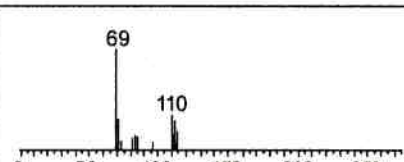
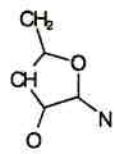
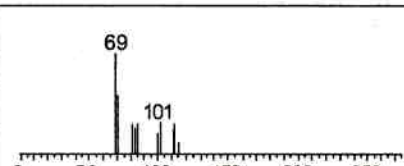
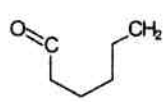
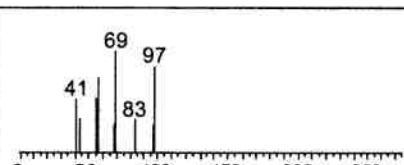
Эксперименты, выполненные на тестовой выборке, показывают, что в общем случае не избежать появления в ответе этой процедуры и лож-

ных фрагментов. Выделим две основные причины этого. Первая: в ответ спектрального поиска отбираются соединения не только требуемого химического класса, но и соединения других классов, имеющие общие спектральные признаки с анализируемым веществом. Характеризующие соединения иных классов общие части структур порождают ложные фрагменты. Вторая причина: принятый нами формализованный подход к оценке корректности фрагмента. Корректными считали только фрагменты, изоморфно вложимые в структуру «неизвестного».

Таблица 2

Результаты процедуры «пересечение структур» при анализе поискового ответа, полученного по масс-спектру «неизвестного» соединения I (код задачи z065)

№	Структура	Спектр
z065		

№ (фактор)	Фрагмент	Отклик (средний спектр)
1 (48)		
2 (47)		
3 (47)		
• • •		
15 (24)		

Напомним (см. табл. 1), что в ответ процедуры «пересечения» включали общие фрагменты, содержащие не менее половины связных вершин молекулярного графа «неизвестного». Очевидно, что выявляемые в таких условиях корректные фрагменты могут характеризовать химический класс

или другие значимые для описания особенностей строения изучаемого соединения признаки.

В реальной практике полезно опираться не только на приведенные выше данные, но и на число связных вершин (размер) выявленного фрагмента. Замечено, что во многих случаях самые крупные фрагменты могут содержать наиболее полную информацию о возможном строении неизвестного. Они часто представляют его большую часть, иногда с точностью до структурной изомерии (например, положения заместителя).

Полученные с помощью процедуры «пересечение» результаты согласуются с представленными в работе [21] данными. Наиболее существенное различие заключается в методе ранжирования компонент множества неизоморфных общих подграфов. В случае системы ХимАрт для этого используются спектральные отклики, что способствует получению спектрально обоснованных оценок достоверности выявляемых фрагментов. Этот прием апробирован в работе [24]. Несомненное достоинство рассматриваемого программного продукта состоит и в том, что в рамках полной системы ХимАрт выявленные по масс-спектру фрагменты можно проверить на соответствие другим спектральным данным, например ^{13}C-ЯМР или ИК спектроскопии. В этом случае корректные фрагменты могут быть дополнительно подтверждены, а ложные скорректированы или отвергнуты.

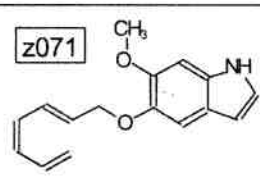
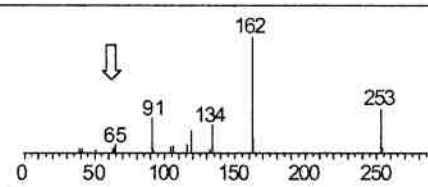
Типичное время выявления общих подграфов для семейства, содержащего 15–30 графов, составляет, в зависимости от размера графов, от 2-х секунд до минуты на компьютере с тактовой частотой не ниже 300МГц.

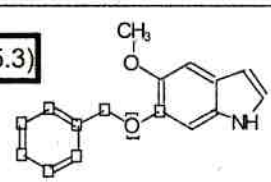
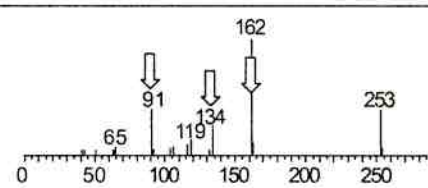
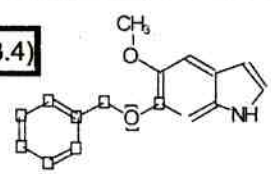
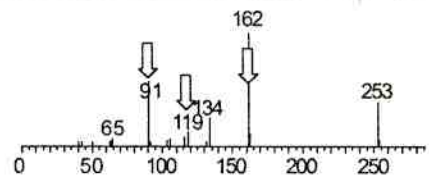
Легко заметить, что получаемых в результате выявления общих подграфов сведений, как правило, недостаточно для определения полной структурной формулы соединения. Между тем именно эта информация представляет наибольший интерес. Реализованная в системе ХимАрт процедура «фрагментация» со встроенным генератором структур способна в ряде случаев оказать исследователям помощь при решении этой задачи.

Рассмотрим характер получаемых при использовании этой процедуры сведений на примере анализа «неизвестного» II (z071). Структурная формула соединения II и масс-спектр представлены в табл. 3, а условия анализа ПО — на рис. 3.

Таблица 3

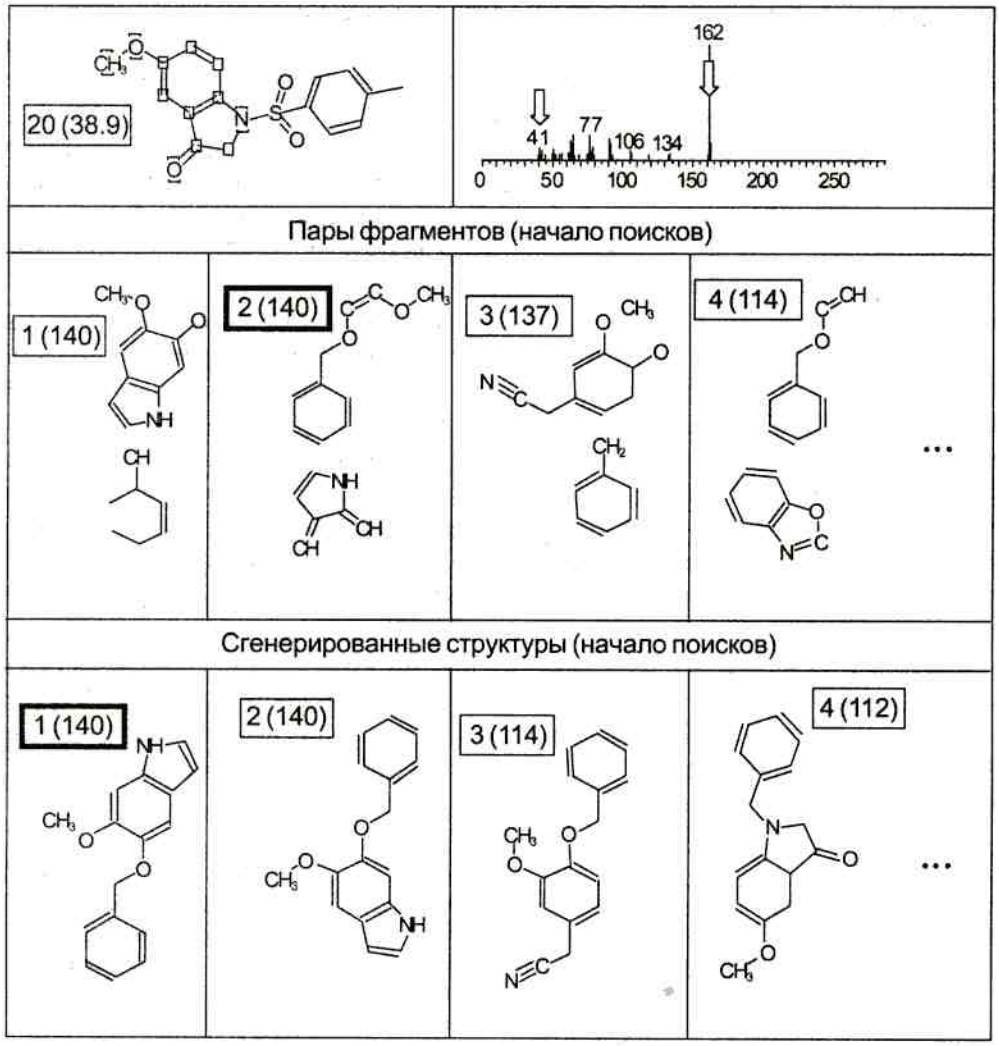
Результаты процедуры «фрагментация» при анализе масс-спектра «неизвестного» II (стрелками помечены «объясненные» пики ионов). В рамке приводятся порядковый номер и фактор ранжирования (в скобках). Жирная рамка отвечает корректным фрагментам и структурам

Структура	Спектр
z071 	

Фрагмент	Спектральный отклик
1 (75.3) 	
2 (73.4) 	

...

Окончание табл. 3



В этом случае при декомпозиции 30 молекулярных графов поискового ответа выявлено 45 различных осколков (фрагментов). Первые три фрагмента из ранжированного списка, и их спектральные отклики приведены в качестве примера в начале табл. 3. Спектральный фактор (в скобках после номера фрагмента) получен так же, как в процедуре «пересечения».

Далее, на основе полного списка фрагментов формируются все возможные пары фрагментов, согласующиеся с молекулярной формулой изучаемого соединения. Ранжирование пар проводится по сумме параметров ранжирования фрагментов, образующих пары. В рассматриваемом случае сформировано 12 пар фрагментов, примеры двух из которых приведены в таблице. На заключительной стадии анализа программное обеспечение позволяет генерировать исчерпывающий список структурных формул, которые образуют соответствующие пары фрагментов. Для рассматриваемого случая генерировано 9 структур (см. табл. 3). Два изомера, среди которых искомая структура, имеют равные параметры ранжирования (140 ед.). Параметр ранжирования последующих структур находится в диапазоне от 127 до 82 ед.

В условиях анализа ПО, приведенных в табл. 1, в рассматриваемом случае генерируются только две первых из представленных в табл. 3 структуры.

Более полную картину дает анализ результатов процедуры «фрагментация структур» для всех объектов тестовой выборки. Он выполнялся при различных условиях спектрального поиска (с целью оценки их влияния на результат), но тождественных параметрах последующей обработки поисковых ответов (см. табл. 1).

Как и следовало ожидать, результативность распознавания структуры «неизвестного» по его масс-спектру в основном определяется содержимым БД и результатом спектрального поиска. Отсутствие в поисковом ответе структурных аналогов «неизвестного» влечет за собой появление пустых списков пар фрагментов и генерирован-

ных структур или, что значительно хуже, генерирование списка структур, не содержащего требуемую структурную формулу «неизвестного».

Будем считать результат анализа поискового ответа положительным в двух случаях:

а) сформированный список вероятных структур содержит структуру «неизвестного»;

б) в результате анализа не генерирована ни одна структура.

Результат генерирования списка структур, который не содержит «неизвестное», будем считать отрицательным.

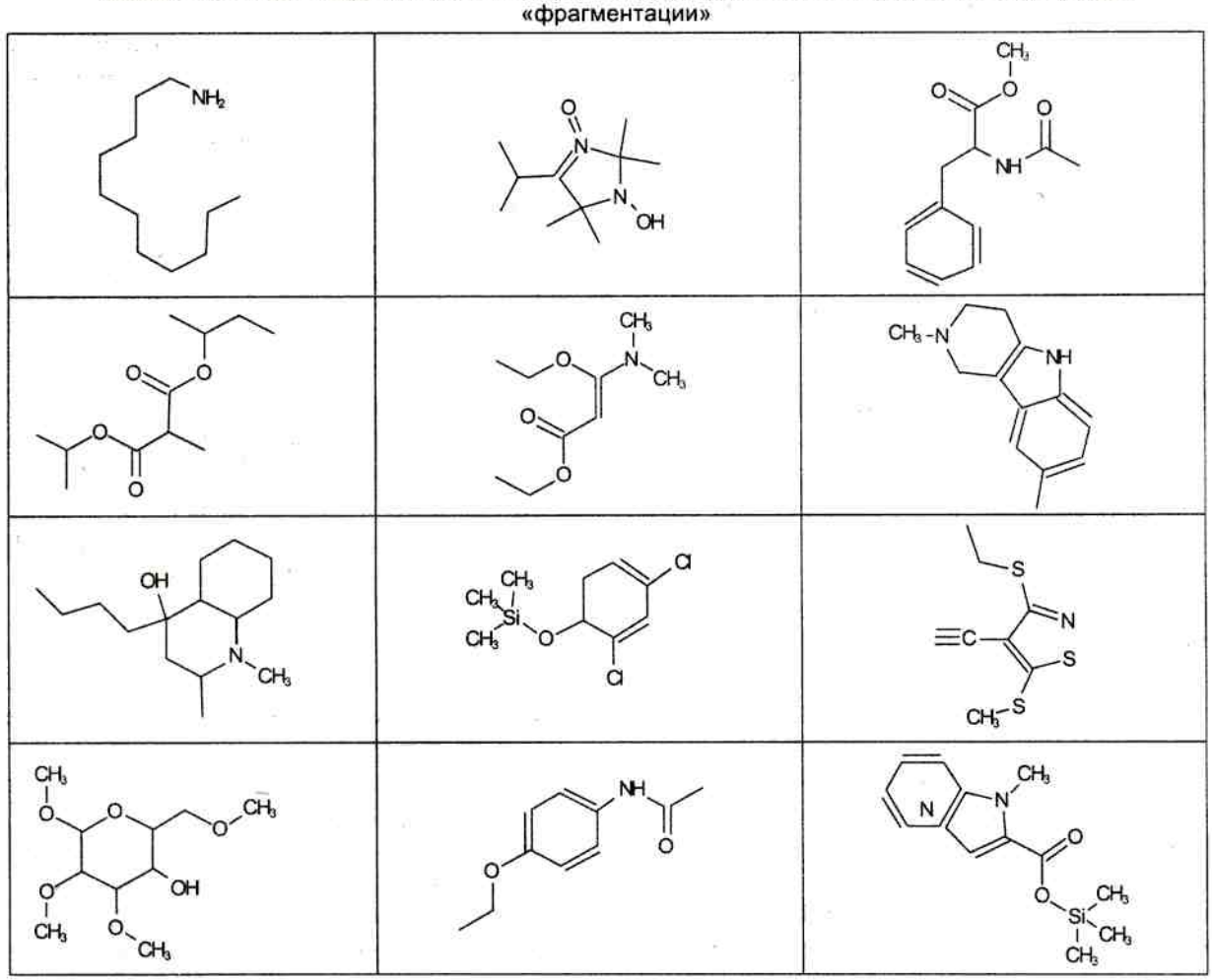
В этих условиях выявлена следующая картина. Как в случае сопоставления предъявленных спектров со спектрами БД в абсолютной шкале массовых чисел (поиск А), так и в случаях сопоставления спектров одновременно в абсолютной и относительных шкалах масс (поиск АВ), результат анализа поисковых ответов оказался сопоставим. Доля положительно решенных задач составляет ~77–78 %.

Выявлен заметный рост вероятности появления структуры «неизвестного» среди генерируемых, если при формировании списков ионов и потерь (из которых формируются пары) учитывать слабоинтенсивные пики масс-спектров. Так, например, найдено, что вовлечение в анализ пиков интенсивностью 30 ед. приводит к ~65 %, а 3 ед. — к ~75 % случаев генерирования требуемой структуры среди общего числа положительных решений. Влияние структур соединений, отобранных в ответ спектрального поиска однозначно. Отсутствие в нем соединений, подобных по строению исследуемому, влечет за собой пустой список генерированных структур или список, не содержащий структуру «неизвестного». В условиях проведенных экспериментов не удалось избежать отрицательных результатов анализа.

В табл.4 приведены примеры структурных формул соединений из тестовой выборки, для которых целевая структура была генерирована в процедуре «фрагментация». Как правило, если структурная формула «неизвестного» генерирована, то в подавляющем большинстве случаев она оказывается на первых местах ранжированного списка.

Таблица 4

Примеры тестовых задач, для которых целевая структура была сгенерирована в процедуре «фрагментации»



В заключение сделаем несколько общих замечаний. Еще раз подчеркнем, что при оценке потенциала использования разработанного программного обеспечения, мы исходили из сугубо формальной схемы «вложения» выявленных структурных особенностей (подграфов или графов) в эталон. Так, например, при оценке способности системы генерировать структуру «неизвестного», даже при появлении в списке генерированных структур изомера того же самого химического класса, что и «неизвестное», но в отсутствии последнего, результат анализа рассматривался как отрицательный. Опытный масс-спектрометрист вряд ли будет настолько формально относиться к получаемым данным, принимая во внимание тождественное поведение спектров ряда изомеров, процессы миграции и перегруппировок, характерные для масс-спектрометрии, или другие сведения о природе изучаемого соединения. Наряду с этим не следует забывать о специфических условиях формирования тестовой выборки. В реальной практике первое обстоятельство способно повлиять на рост выявленных на тестовой выборке статистических данных, а второе — наоборот — на их снижение.

Заметим также, что в экспериментах на тестовой выборке были использованы формализованные условия поиска и анализа поисковых ответов, тождественные для всех соединений выборки. Опытный пользователь, как правило, применяет индивидуальный подход при решении конкретной структурной задачи, выбирая, в частности, режим поиска с учетом особенностей регистрации спектра, спектрального поведения соединений предполагаемого химического класса, предыстории анализируемого образца. Влияние на конечный результат режима поиска и его граничных условий, вероятно, специфично для каждой задачи. Так, например, на тестовой выборке в ряде случаев (6-8 %) обнаружено, что структура «неизвестного» генерирована на основе ПО, полученного в режиме поиск А, но не генерирована при поиске АВ и наоборот. Вероятно, как отмечалось ранее [19], результаты поисков А, В и АВ могут дополнять друг друга.

Несомненно влияние на конечный результат параметров обработки отобранных при спектральном поиске данных (см. примеры анализа поисковых ответов, полученных по масс-спектрам соединений I и II). При увеличении числа анализируемых записей ПО несколько увеличивается вероятность появления корректного фрагмента в ответе процедуры «пересечение структур», но значительно возрастает число выявляемых фрагментов. Аналогичный эффект наблюдается при генерировании структур в процедуре «фрагментация». Попытки увеличения числа генерируемых структур путем расширения требований на «остаток» (до двух или трех атомов) приводят к значительному росту числа генерируемых структур, но в общем случае это сопровождается снижением числа положительно решенных задач. Очевидно, что конечный результат определяется наличием в БД и, соответственно, в поисковом ответе соединений, подобных исследуемому как в спектральном, так и структурном отношениях. Поэтому дальнейшее увеличение объема базы данных будет сопровождаться ростом результативности использования описанной выше системы.

ЛИТЕРАТУРА

1. Warr W.A. Computer-assisted Structure Elucidation. Part 1. Library Search and Spectral Data Collections *and* Part 2. Indirect Database Approaches and Established Systems // Anal. Chem. 1993. V.65. P.1045A-1050A, 1087A-1095A.

2. Computing applications in molecular spectroscopy / Ed. George W.O., Steele D. Cambridge: Royal Society of Chemistry. 1995. 236 p.

3. Эляшберг М. Е. Экспертные системы для установления структуры органических молекул спектральными методами // Успехи химии. 1999. Т.68. С.579–604.

4. Дробышев Ю.П. Комплексная машинная система для решения структурных задач методами молекулярной спектроскопии / Ю.П.Дробышев, В.А.Коптюг // Автометрия. 1972. Т.4. С.118-123.

5. Коптюг В.А. Использование ЭВМ при решении структурных задач органической химии методами молекулярной спектроскопии / В.С.Бочкарев, Б.Г.Дерендяев, С.А.Нехорошев, В.Н.Пиоттух-Пелецкий и др. // Журн. структ. химии. 1977. Т.18, №3. С.440-459.

6. Bremser W. SpecInfo — A Multidimensional Spectroscopic Interpretation System / W.Bremser, M.Grzonka // Microchim. Acta. 1991. V.11. P.483-491.

7. Лебедев К. С. Компьютерные методы решения структурно-аналитических задач с помощью банков данных по молекулярной спектроскопии (МС, ИК, ЯМР) / К.С.Лебедев, Б.Г.Дерендяев // Химия в интересах устойч. развития. 1995. Т.3. С.269–285.

8. Gray N. A. B. Computer-assisted structure elucidation. N.Y.: Wiley & Sons, 1986. 536 p.

9. Computer-supported spectroscopy databases / Ed. Zupan J. Chichester: Ellis Horwood, 1986. 344 p.

10. Barth A.SpecInfo: An Integrated Spectroscopic Information System //J.Chem. Inf. Comput. Sci. 1993. V.33. P.52-58.

12. Лебедев К.С. Использование баз данных по ИК и масс-спектрам для установления строения органических соединений. // Ж. аналит. химии. 1993. Т.48. С.851-863.
13. Debska B.J. The methodology of knowledge acquisition from the collection of IR and UV spectra / B.J.Debska, B.Guzowska-Swider // Fresenius J. Anal. Chem. 1998. V.361. P.235-238.
14. Strokov I.I. A New Modular Architecture for Structure Elucidation Systems / I.I. Strokov, K.S. Lebedev // J. Chem. Inf. Comput. Sci. 1996. V.36. P.741-745.
15. Strokov I.I. Computer Aided Method for Chemical Structure Elucidation Using Spectral Databases and ^{13}C NMR Correlation Tables / I.I.Strokov, K.S.Lebedev // J. Chem. Inf. Comput. Sci. 1999. V.39. P.659-665.
16. McLafferty F.W. Retrieval and interpretative computer programs for mass spectrometry. F.W.McLafferty, D.B.Stauffer // J.Chem. Inf. Comput. Sci. 1985. V.25. P.245-252.
17. Киршанский С.П. Система «КОМПАС-МС», база данных и принципы организации / С.П. Киршанский, К.С.Лебедев, Б.Г.Дерендяев // Ж. аналит. химии. 1987. Т.42, №6. С.1092-1097.
18. Дерендяев Б.Г. Информационный поиск — средство предсказания брутто-формулы соединения по его масс-спектру / Б.Г. Дерендяев, С.А. Нехорошев, С.П. Киршанский, К.С. Лебедев // Ж. аналит. химии. 1987. Т.42, №.7. С.1312-1319.
19. Киршанский С.П. Аналитические возможности системы «КОМПАС-МС» / С.П. Киршанский, К.С. Лебедев, С.А. Нехорошев, Б.Г. Дерендяев // Ж. аналит. химии. 1987. Т.42, №.7. С.1320-1329.
20. Varmuza K. Mass spectral classifiers for supporting systematic structure elucidation / K. Varmuza, W. Werther // J. Chem. Inf. Comput. Sci. 1996. V.36. P.323-333.
21. Дерендяев Б.Г. Использование базы данных «масс-спектр - фрагментный состав соединения» при установлении строения органических соединений / Б.Г.Дерендяев, В.Н.Пиоттух-Пелецкий, К.С.Чмутина, С.А.Нехорошев // Журн. аналит. химии. 2002, Т.57, №11. С.1176-1185.
22. Киршанский С.П. Извлечение структурной информации из масс-спектров с помощью ЭВМ. XII. Генерирование структурных гипотез и их ранжирование на основе результатов работы ИПС / С.П. Киршанский, С.Г. Молодцов, К.С. Лебедев // Изв. СО АН СССР, сер. хим. наук. 1989. Вып.5. С.3-9.
23. Strokov I. A Compact Code for Chemical Structure Storage and Retrieval // J. Chem. Inf. Comput. Sci. 1995. V.35. P.939-944.
24. Lebedev K.S. New Computer-Aided Methods for Revealing Structural Features of Unknown Compounds Using Low Resolution Mass Spectra / K.S. Lebedev, D. Cabrol-Bass // J. Chem. Inf. Comput. Sci. 1998. V.38. P.410-419.
25. Строков И.И. Представление структурной информации и поиск структурных аналогов в базах данных по молекулярной спектроскопии / И.И.Строков, К.С.Лебедев, Б.Г. Дерендяев // Ж. структ. химии. 1996. Т.37. С.1128-1138.
26. Строков И.И. ХимАрт — информационно-логическая система по молекулярной спектроскопии. Масс-спектрометрическая компонента / И.И.Строков, К.С.Лебедев, Б.Г.Дерендяев // НТИ. Сер.2. 2003. №2. С.18-26.
27. Лебедев К.С. Опознание крупных структурных фрагментов неизвестного соединения с помощью поисковой системы по ИК-спектрам / К.С.Лебедев, О.Н.Шарапова, И.К.Коробейничева, В.А.Кохов // Изв. СО РАН (Сиб. хим. журн.). 1993. №1. С.50-56.
28. Molodtsov S.G. Generation of molecular graphs with a given set of nonoverlapping fragments. // Math. Chem. (MATCH) 1994. 30. P.203-224.

* * * * *

SEARCH AND RETRIEVAL SYSTEM CHEMART. OBTAINING A STRUCTURAL FORMULA FROM A MASS SPECTRUM

B. G.Derendjaev, I. I.Strokov, K. S.Lebedev

The papers illustrates a possibility of organic structure elucidation starting from low resolution mass spectra in the frame of search and retrieval system ChemArt. The discussed experimental material includes solving of more than 100 practice structural problems.