

Экспертные системы для установления структуры органических молекул спектральными методами

М.Е.Эляшберг

*Всероссийский научно-исследовательский институт органического синтеза
107005 Москва, ул. Радио, 12, факс (095)261–0777*

Проанализировано состояние исследований, направленных на создание экспертных систем для установления структуры органических молекул по ИК-спектрам и спектрам ЯМР на ядрах ^1H и ^{13}C . Рассмотрены компьютерные методы выявления молекулярных фрагментов, генерации структур и прогнозирования их спектров. Обсуждены принципы построения современных экспертных систем и общая стратегия решения структурных задач. Библиография — 174 ссылки.

Оглавление

I. Введение	579
II. Методы выявления структурных фрагментов	580
III. Генерация структурных формул изомеров	586
IV. Прогнозирование спектров вероятных структур	587
V. Направления развития экспертных систем	593
VI. Стратегия установления структуры молекул	598
VII. Заключение	602

I. Введение

Со времени опубликования работ^{1–5}, в которых впервые рассматривался вопрос об использовании математических методов и ЭВМ для установления структуры органических молекул по их спектрам, прошло более 30 лет. Уже в этих статьях было показано, что структурно-групповой анализ (СГА) может быть формализован с помощью методов математической логики и выполнен на ЭВМ,^{1,2} а применение теории графов и комбинаторики позволяет автоматически генерировать (математически синтезировать) структурные формулы из наборов фрагментов.³ За прошедший период в пограничной области многих наук, объединенных на основе математических алгоритмов в виде компьютерных программ, возникло новое научное направление — применение искусственного интеллекта (ИИ) для выявления молекулярных структур по спектральным данным.

В послевоенный период спектральные методы получили широкое распространение в химии. Не будет преувеличением сказать, что методы ИК-, ЯМР-спектроскопии и масс-спектрометрии произвели революцию в методологии химического исследования и «химического» мышления и стали для химика-органика одним из основных источников структурной информации. Неоспоримы достоинства спектральных методов — универсальность, экспрессность, легкость сопря-

жения с компьютером и использование хорошо разработанных теорий (особенно это касается ИК- и ЯМР-спектроскопии). Именно это обстоятельство, а также накопление обширных сведений об эмпирических корреляциях спектр–структура стимулировали исследования по применению ИИ в химии.

Для формализации процесса установления структуры молекул потребовалось найти математический аппарат, адекватный задаче исследования. Поскольку молекула по природе дискретна и обладает дискретными уровнями энергии, то в качестве такого аппарата была выбрана математическая логика, теория графов и комбинаторика. Именно дискретная математика, описывающая модели типа «черного ящика», стала одним из основных инструментов математической химии, интенсивно развивающейся в последние десятилетия.

Математический синтез структур из атомов и фрагментов на первых порах показался всего лишь одним из этапов в процессе идентификации органических соединений, однако вскоре стало ясно, что его значение выходит далеко за рамки рассматриваемой задачи. В самом деле, неисчислимо множество органических молекул, конструируемых генератором структур, можно интерпретировать как перечисление логических следствий из системы аксиом, лежащих в основе теории строения молекул. В то же время одно только знание структурной формулы соединения дает возможность предсказывать целый ряд его важных свойств. Это позволило поставить вопрос о разработке математических методов прогнозирования структур соединений с желаемыми свойствами и предсказания свойств соединений по заданной структуре (работы школы Н.С.Зефирова в этой области получили мировое признание). Богатый опыт, накопленный химией, показывает, что большинство прогнозов классической теории строения органических веществ о возможности существования в природе и синтетического получения не известных пока соединений было реализовано, причем процент неверных прогнозов относительно невелик.⁶

М.Е.Эляшберг. Доктор химических наук, кандидат физико-математических наук, профессор, заведующий лабораторией молекулярной спектроскопии ВНИИОС. Телефон: (095)261–5161, e-mail: elyas@molspec.msk.ru
Область научных интересов: ИК- и ЯМР-спектроскопия, экспертные системы для установления структур молекул по спектрам, компьютерные методы в химии.

Дата поступления 25 февраля 1999 г.

Важным результатом работ по применению ИИ в молекулярной спектроскопии стала формализация процесса химического исследования на базе интеграции многих наук.

Формирование концепции и выработка методологического подхода в этой новой области в целом завершились к началу 1980-х годов, что нашло свое отражение в ряде монографий^{7–10} и обзоров.^{11–15} В результате поисков исследователей разных стран были найдены общие принципы построения систем ИИ (в настоящее время их называют экспертными системами (ЭС)) и разработаны соответствующие алгоритмы, проверенные на большом числе задач.

В литературе описано около двадцати ЭС для молекулярной спектроскопии, которые обладают разными возможностями, используют разные алгоритмы и инструментальные средства, ориентированы на применение различных методов спектрального анализа. Однако общим для этих систем является то, что они имитируют рассуждения эксперта-спектроскописта при решении задачи идентификации. Этот мыслительный процесс складывается из следующих основных этапов:

- 1) структурно-групповой анализ неизвестного вещества;
- 2) построение возможных структурных формул из обнаруженных наборов фрагментов;
- 3) проверка соответствия структурных формул экспериментальным спектрам и дополнительной информации с использованием прогнозирования спектров предполагаемых структур.

Каждый из этапов моделируется соответствующим блоком экспертной системы. Перечисленные блоки являются относительно независимыми, а выполняемые ими функции обычно реализуются на базе разных математических методов. В настоящее время методы выявления фрагментов, генерации структур и предсказания их спектров стали практически самостоятельными разделами компьютерной спектроскопии и математической химии.

Чтобы подойти как можно ближе к решению реальных задач с помощью ЭС, потребовалось глубже изучить природу задачи установления структуры. Авторы работы¹⁶ рассмотрели особенности этой задачи как пример обратной задачи.¹⁷ Обратные задачи, как известно, относятся к классу так называемых некорректно поставленных задач: без наложения априорных ограничений они не имеют единственного решения.¹⁸ Был сделан важный вывод о принципиальной невозможности исключения химиков из процесса установления структуры и необходимости решать задачи в интерактивном режиме. В этот период также было осознано, что информация о взаимозависимости между спектрами и структурой обычно является «нечеткой» (неоднозначной). С учетом этого обстоятельства в работе¹⁹ для формализации СГА и приближенного прогнозирования спектров было предложено использовать методы нечеткой математики.²⁰ Результаты, полученные разными исследовательскими группами к началу 1990-х годов, отражены в обзорах^{21–23}. Примерно в это же время возник большой интерес исследователей к алгоритмам искусственных нейронных сетей (ИНС),^{24–27} применение которых в различных областях компьютерной химии представляется весьма перспективным. Число работ в этом направлении быстро растет. Оно оказалось настолько новым для химии и хемометрики, что, например, в монографии «Хемометрика»,²⁸ опубликованной за рубежом в 1986 г., отсутствует даже упоминание об ИНС.

В это же время стало ясно, что ЭС имеют шанс стать инструментом органического анализа только в том случае, если они «научатся» работать с крупными неаддитивными структурами (например, со структурами природных соединений) и будут обладать высокой степенью дружелюбия. На работе в этом направлении были сконцентрированы усилия нескольких исследовательских групп, что привело к созданию новых ЭС (см., например,^{29–32}).

В данной работе обобщен опыт, накопленный главным образом за последнее десятилетие, по созданию ЭС, которые ориентированы на установление структуры молекул по ИК-спектрам и спектрам ЯМР на ядрах ¹H и ¹³C. Вначале рассматриваются методы выявления наборов фрагментов, генерации структур и предсказания их спектров, а затем на примере наиболее совершенных ЭС проанализированы стратегии установления структуры. Особое внимание уделено обсуждению трудностей, с которыми встречаются системы, и анализу возможных путей их преодоления.

II. Методы выявления структурных фрагментов

Для выявления структурных фрагментов в большинстве ЭС используются характеристические признаки функциональных групп и других элементов структуры в молекулярных спектрах. Базу знаний ЭС обычно составляют молекулярные фрагменты и интервалы изменения их характеристических признаков (частот в ИК-, химических сдвигов в ЯМР-спектрах), а также правила, связывающие спектры со структурой. В ряде систем наряду с молекулярными фрагментами в базу знаний включены также библиотеки структур и соответствующих им спектров. В зависимости от типа базы знаний используются разные методы извлечения структурной информации.

1. Формально-логический метод

В работах^{1,33} показано, что схему связей между элементами структуры молекулы и их спектральными признаками можно описать дискретной структурной моделью, поэтому применение математической логики при исследовании взаимосвязи спектр – строение молекулы естественно вытекает из самой природы изучаемого объекта. Предполагается, что молекула является аддитивной или частично аддитивной, а каждый фрагмент A_i из множества фрагментов A , имеющих отношение к данной задаче, характеризуется одним или несколькими интервалами изменения характеристических признаков $\omega_i \in \Omega$ (Ω — множество спектральных признаков).

В качестве аппарата теории структурно-группового анализа была использована булева алгебра. Это позволило получить в общем виде логические соотношения, связывающие структурные элементы с их характеристическими спектральными признаками, и построить модель рассуждений при выявлении структурных фрагментов.

Булева алгебра имеет дело с простыми высказываниями и их комбинациями, которые называются булевыми функциями. Комбинации высказываний строятся при помощи союзов «и» (конъюнкция), «или» (дизъюнкция), операции отрицания истинности высказывания и операции импликации. Последняя имеет смысл сложного высказывания «если ..., то ...» и обозначается стрелкой ($\rightarrow$). Например, высказывание «если молекула содержит фрагмент A_i , то в ее спектре наблюдается признак ω » представляется импликацией $A_i \rightarrow \omega$. Элементы булевой алгебры и теория СГА изложены в монографии⁷.

Общий ход рассуждений специалиста при выполнении СГА представляется следующим образом: если булева функция $T(A, \Omega)$ правильно отражает взаимосвязь спектра со строением молекул, то наличие в спектре комбинации спектральных признаков, описываемой булевой функцией $R(\Omega)$, свидетельствует о присутствии в исследуемой молекуле хотя бы одной из комбинаций структурных элементов, определяемых булевой функцией $f(A)$, т.е.

$$T(A, \Omega) \rightarrow [R(\Omega) \rightarrow f(A)].$$

Это выражение представляет собой общую формулировку задачи качественного спектрального анализа на языке булевой алгебры, оно является логическим уравнением отно-

сительно функции $f(A)$. Решение логического уравнения сводится к вычислению функции $f(A)$, алгоритм вычисления которой на основе концепции изображающих чисел описан в монографии⁷. Функция $f(A)$ содержит наборы фрагментов, каждый из которых может объяснить характеристические признаки в экспериментальных спектрах.

Описанный подход положен в основу блока СГА в нескольких поколениях системы РАСТР,^{34–36} а также используется в системах X-PERT,²⁹ CHEMICS,³⁷ EXPIRS^{38,39} и др. Он позволяет успешно отбирать вероятные фрагменты для создания из них логически непротиворечивых наборов, пригодных для последующей генерации структур. Однако недостатком такого подхода является его слишком жесткая логическая детерминированность. Например, незначительный выход характеристического признака фрагмента в экспериментальном спектре за пределы указанного интервала ведет к потере соответствующего фрагмента. Кроме того, база знаний, с которой работает алгоритм, остается в значительной степени подобной корреляционным таблицам, следовательно, обширная информация, изложенная естественным языком в многочисленных монографиях по молекулярной спектроскопии (закономерности, эмпирические правила и т.д.), не используется. В связи с этим были предприняты исследования по созданию такой ЭС, в которой взаимосвязь структура–спектр описывается более детально.^{19,40}

2. Методы логики предикатов

Для построения ЭС, оперирующей знаниями, представленными на профессионально ограниченном естественном языке, необходимо найти подходящий для этой цели математический аппарат. Требования, предъявляемые к такому аппарату, обусловлены характером знаний, сформулированных на естественном языке. С одной стороны, описательным (декларативным) знаниям, в частности в области ИК-спектроскопии (см., например,⁴¹), свойственна некоторая нечеткость, размытость, неоднозначность. Примером может служить следующее утверждение:⁴¹ «...валентные колебания C=C-связи обычно наблюдаются в ИК-спектре около 1650 см⁻¹, но иногда смещаются в область 1685 см⁻¹...». С другой стороны, логические умозаключения в процессе решения качественных задач производятся также на естественном языке с присущими ему неоднозначностями и скрытыми оценками достоверности (например, «часто», «редко», «иногда» и т.д.). Человек мыслит не числами, а нечеткими понятиями. Поэтому теоретический аппарат, достаточно адекватно моделирующий наши знания и рассуждения, должен удовлетворять требованию, образно сформулированному создателем нечеткой математики Л.Заде:⁴² «Теория о природе должна отражать то, что природа пишет скорее произвольными мазками, чем шариковой ручкой».

Для представления знаний по ИК-спектроскопии авторами работ^{19,40} было предложено прикладное логическое исчисление (ПЛИ), опирающееся на теорию предикатов первого порядка⁴³ и теорию нечетких множеств.^{20,42} Основные положения ПЛИ следующие.

n-Местным предикатом называют логическую функцию $P(X_1, X_2, \dots, X_n)$, где $X_i \in M(X_i)$, принимающую значения «истинно» или «ложно» при постановке в нее объектов из определенной предметной области. Предикат отражает чаще всего свойства предметов и их отношения. Например, пусть предикат $P(X)$ отражает свойство веществ иметь сильную полосу ИК-поглощения около 1700 см⁻¹, и пусть $X_1 = \{\text{ацетон}\}$, а $X_2 = \{\text{пропилен}\}$. Тогда, очевидно, $P(X_1) = 1$ («истинно»), а $P(X_2) = 0$ («ложно»). Если $M(X_i)$ — множество, то нечеткое подмножество F множества $M(X_i)$ определяется как множество упорядоченных пар $\{X_i, \mu_A(X_i)\}$, где $\mu_A(X_i)$ — характеристическая функция принадлежности, которая изменяется в пределах от 0 до 1 и указывает степень

принадлежности элемента X_i подмножеству F . Например, для высказывания «в ИК-спектре имеется полоса около 1710 см⁻¹» функция принадлежности запишется так: $\mu_A(1710 \text{ см}^{-1}) = 1$. При этом предполагается, что существуют границы, например 1700 и 1720 см⁻¹, за пределами которых функция $\mu_A(X_i)$ равна нулю.

Если предикатным переменным придать смысл элементов нечетких множеств, то получим нечеткие предикаты. Каждому *n*-местному предикатному символу ставится в соответствие отображение M^n в множество значений характеристической функции принадлежности, которая в ПЛИ именуется функцией истинности μ_p , $1 \geq \mu_p \geq 0$.

Для описания спектрохимических знаний была разработана система предикатов, описывающих взаимосвязь ИК-спектр–структура. В частности, были введены предикаты, характеризующие ИК-спектр («в спектре содержится полоса, принадлежащая к колебаниям типа ζ структурной группы g ; полоса характеризуется частотой X_i , интенсивностью Y_i и полушириной Z_i ») и строение молекулы («молекула содержит фрагмент R_1 », «в молекуле содержится фрагмент R_1 и через n_1 связей от него находятся m_1 фрагментов R_2 », «в молекуле имеется цикл размера j , в который входит фрагмент R_3 » и т.д.). Специальные предикаты характеризуют агрегатное состояние вещества, тип растворителя и т.д. Для описания изменения данных вводятся предикаты второго порядка, с помощью которых можно отразить направление изменения переменных, входящих в предикаты первого порядка (например, «увеличивается», «уменьшается», «исчезает», «расщепляется на d компонентов» и т.д.).

Используя такую систему предикатов и обычные логические операции, можно формализовать значительную часть декларативных знаний. Например:

1) «валентные колебания ν_s соединений с одной двойной C=C-связью и алкильными заместителями (Alk) поглощают в области 1640–1680 см⁻¹»

$\text{STRUCT}(C=C, \text{Alk}) \rightarrow \text{SPECTR}(\nu_s, C=C, 1640, 1680)$;

2) «сопряжение двойных связей двух алкилзамещенных групп приводит к расщеплению полосы валентных колебаний ν_s на две полосы»

$Q_1(\text{STRUCT}, C=C, \text{Alk}(C=C)) \rightarrow$
 $\rightarrow Q_2(\text{SPECTR}, \nu_s, C=C, X_1(\text{SPL}, 2))$.

Логической формуле G и предикатному символу P ставится в соответствие множество внешних значений функции истинности μ_e , $1 \geq \mu_e \geq 0$. Внешнее значение функции μ_e отражает квантификацию знаний («часто», «иногда», «в некоторых случаях» и т.д.). Например, утверждение «валентные колебания связи C=C довольно часто наблюдаются в области 1640–1660 см⁻¹» отображается элементарной формулой $[0.8]\text{SPECTR}(\nu_s, C=C, 1640, 1660 \{10\})$, где $[0.8]$ — внешнее значение истинности, $\{10\}$ — длина отрезка (в см⁻¹) слева и справа от границ области 1640–1660 см⁻¹, где трапециевидная функция принадлежности спадает до нуля. Чтобы определить значение функции истинности μ_i элементарной формулы необходимо перемножить внешнее значение μ_e и значения функций принадлежности нечетких констант, а также переменных этой формулы. В данном примере для значения переменной $X = 1665 \text{ см}^{-1}$ с учетом свойств треугольника имеем $\mu(X) = 0.5$; при $\mu_e = 0.8$ величина μ_i будет равна $0.5 \cdot 0.8 = 0.4$.

Спектрохимические знания можно разделить на факты (выражения, содержащие константы) и закономерности (выражения, содержащие хотя бы одну логическую переменную). Со знаниями удобно оперировать, если они структурированы. Для структуризации знаний выделяется класс общих закономерностей, справедливых для большинства органических молекул (например, «малые циклы с тройной связью не встречаются в органической химии»). В качестве единиц

структурирования остальных знаний принимаются основные функциональные группы, обладающие характеристическими спектральными признаками ($C=C$, $C\equiv C$, $C=O$, $C\equiv N$, OH и т.д.). Молекулярный фрагмент описывается его ядром ($C=C$, $C\equiv C$ и т.д.), к которому последовательно присоединяются атомы окружения, образуя соответственно фрагменты первого, второго, третьего и т.д. уровней. Образующаяся иерархия знаний представляет собой граф в виде дерева, называемый сетью декларативных знаний. Описание узла сети содержит структуру фрагмента, возможные типы его колебаний, закономерности и факты для каждого типа колебаний, закономерности и факты, характерные для данного фрагмента.

Если рассматривать декларативные знания как систему аксиом прикладной теории, то задача СГА представляется утверждениями (теоремами), которые необходимо доказать или опровергнуть. Показано,⁴³ что для автоматического доказательства теорем (автоматического СГА) можно использовать классический метод резолюций, адаптированный к теории нечетких предикатов.

Предложенный формализм позволяет решать задачи СГА и структурной интерпретации ИК-спектров, причем по результатам интерпретации возможно самообучение декларативным знаниям, в процессе которого проверяются факты и закономерности, изменяются значения характеристических функций принадлежности, выявляются новые факты и закономерности.

Описанный подход был реализован в виде исследовательского прототипа ЭС на языке Пролог.⁴⁴ Система выявляет непротиворечивые наборы фрагментов (узлов сети), объясняющие спектр. Одновременно производится интерпретация ИК-спектра: каждая частота относится к определенному типу колебаний соответствующих группировок с указанием истинности μ_p именно такого отнесения.

Логика предикатов и алгоритмический язык Пролог также использовались для структурной интерпретации ИК-спектров в системе EXPEC.^{45–47} База знаний системы включает правила интерпретации. Фрагменты, входящие в базу, как и в предыдущей системе,⁴⁴ иерархически структурированы в виде сети декларативных знаний. При интерпретации спектра неизвестного соединения сеть просматривается, и для каждого фрагмента находится правило, по которому анализируется связь фрагмента со спектром. Каждому полученному результату сопоставляется его вероятность. Средства объяснения позволяют проводить трассировку процесса рассуждения программы.

База знаний системы формируется на основе корреляций спектр–структура, известных из литературы, и пополняется с помощью генератора правил интерпретации,⁴⁵ а также за счет ввода знаний экспертов. Для автоматической генерации правил используется библиотека ИК-спектров органических соединений класса СНО. Программа продуцирует множество потенциальных правил, выделяет из них наименее коррелированные и минимизирует их число.

Наличие генератора правил⁴⁵ — достоинство системы EXPEC. Процесс автоматической генерации правил состоит из четырех шагов: 1) отбор структурных фрагментов; 2) поиск в базе данных всех структур, содержащих фрагмент S_i ; 3) поиск интервалов характеристических признаков для S_i ; 4) выбор интервалов, пригодных для использования в правилах интерпретации. Проблема формирования базы знаний имеет исключительно важное значение для создания эффективно работающей ЭС, поэтому описанный выше прием автоматизации этого процесса представляется весьма перспективным.

Интерпретатор (машина вывода) отбирает из сети только те фрагменты, для которых подтверждено присутствие групп, находящихся в узлах сети. Например, правила, касающиеся обнаружения соединений фенола, могут быть применены только после подтверждения присутствия бензольного

кольца и гидроксильной группы. Для каждой подструктуры вычисляется вероятность ее присутствия, причем функция распределения вероятности для интервала выбирается в форме трапеции (например, в работах^{19,40} эта форма выбрана для функции принадлежности полосы к интервалу поглощения). Отобранные фрагменты ранжируются по убыванию вероятности их присутствия.

Разными исследователями предлагаются различные правила для вывода вероятных наборов фрагментов (см., например,⁴⁸). Представляется, что работы в этом направлении получат дальнейшее развитие и позволят вплотную подойти к созданию ЭС для молекулярного спектрального анализа, общающихся с химиком на профессионально ограниченном естественном языке. Однако на этом пути исследователи неизбежно столкнутся со многими трудностями, для преодоления которых потребуются привлечение специалистов из смежных областей, таких, например, как математическая лингвистика, математическая логика, инженерия знаний и т.д.

3. Применение искусственных нейронных сетей

С начала 1990-х годов внимание химиков привлекли возможности нового перспективного аппарата компьютерной химии — искусственных нейронных сетей (ИНС). Особенно быстро увеличивается число работ по использованию ИНС для интерпретации и классификации спектральных данных.

Нейронная сеть — это упрощенная модель человеческого мозга, состоящая из нескольких слоев нейронов, которые посылают сигналы другим нейронам в зависимости от полученных входных сигналов. Такие сети функционируют по принципу «черного ящика» и обладают общей способностью строить эмпирические модели систем, для которых теоретические зависимости между входом и выходом слишком сложны или вообще неизвестны. Модели получаются в результате обучения сети. В процессе обучения сеть представляется парами вход–выход, которые связаны моделируемым преобразованием. Сеть, обученная на этих примерах, способна предсказывать выходы по входам, которые она еще «не видела». Процедура обучения может потребовать больших затрат времени (десятки часов), но однажды обученная сеть выдает ответ или предсказание почти мгновенно. Например, сеть может быть обучена выдавать структурную информацию (выход), извлеченную из спектра (вход), или по структурной информации (вход) предсказывать спектр (выход). Применяя эту методологию к спектрам, можно установить, какие функциональные группы могут присутствовать в образце, а какие нет, при этом выдается оценка надежности результата.

Рассмотрим подробнее общие принципы построения нейронных сетей.^{24,49}

Строго говоря, ИНС — это вычислительная структура, состоящая из некоторого числа взаимосвязанных элементарных модулей, называемых вершинами, или нейронами. В упрощенном виде функционирование нейрона можно представить следующим образом.

Значения «сигналов» s_i ($i = 1 \rightarrow m$), поступающих на вход нейрона (эти значения будем называть входами), умножаются на весовые множители w_i и суммируются, образуя функцию Net:

$$\text{Net} = s_1w_1 + s_2w_2 + \dots + s_mw_m.$$

Затем с помощью нелинейной передаточной функции осуществляется преобразование величины Net в выходной сигнал нейрона. В качестве передаточной функции чаще всего используется сигмоидальная функция

$$\text{sf} = \frac{1}{1 + \exp(-s_iw_i)}.$$

Этот сигнал далее используется в качестве входа для других нейронов, а общая организация связей между нейронами определяет архитектуру сети. Большое число нейронов и их взаимосвязей усложняет поведение сети. Сеть может быть модифицирована с помощью настройки ее параметров, среди которых главными являются весовые множители. Настройка параметров, приводящая к желаемому поведению, называется обучением сети. Этот процесс выполняется итеративно — с помощью обучающего алгоритма на множестве примеров. На каждой итерации выходы сети сравниваются со значениями эталонов (примеров), при обнаружении различия между ними программа вносит поправки в весовые множители. Процесс обучения продолжается до совпадения выходов с эталонами.

В первых работах по структурной интерпретации ИК- и ЯМР-спектров чаще всего использовались многослойные ИНС, обучение которых осуществлялось в соответствии с алгоритмом обратного распространения ошибок (ОРО).²⁴ В этих сетях каждый нейрон соединен со всеми нейронами, находящимися в предшествующих и последующих слоях. Первый (входной) слой предназначен для приема входных данных и распределения их по следующему слою. Последний (выходной) слой или слой решения ответствен за выходы всей сети. Остальные слои, если они имеются, называются скрытыми (рис. 1). Информация распространяется от входного слоя через скрытый (скрытые) к выходному слою. Например, ИК-спектр, закодированный в виде значений интенсивности в точках, равномерно распределенных по оси абсцисс, используется в качестве входного вектора. Координатами выходного вектора являются некоторые числа, означающие наличие или отсутствие определенных структурных признаков (часто это нули и единицы). Число точек, которыми кодируется спектр, определяет размерность (число вершин) входного слоя, а число введенных структурных признаков (фрагментов) определяет размерность выходного слоя. Не существует универсального правила задания числа вершин в скрытых слоях: оно должно быть достаточно большим, чтобы обеспечить правильность обработки данных, однако с ростом числа вершин увеличивается время обучения, а также риск «перетренировки» сети. Последнее явление состоит в том, что в процессе обучения после некоторого числа итераций (их называют циклами или эпохами) производительность сети начинает падать.

Начальные значения весовых множителей w_i задаются случайным образом, после чего производится обучение сети на множестве спектров веществ, содержащих фрагменты, которые сеть должна научиться обнаруживать. (Такое множество называют тренировочным.) Имеется много вариантов алгоритма ОРО, но все они базируются на одном и том же принципе: минимизация общей ошибки, вычисляемой как сумма разностей (или квадратов разностей) между значе-

ниями выходов сети и желаемыми результатами, производится методом наискорейшего спуска.

Чтобы определить производительность (эффективность) сети используется тестовое множество примеров. Это множество должно отличаться от тренировочного множества, так как оно предназначено для оценки способности сети к обобщению (но не к обучению).

Размер и состав тренировочного множества существенно влияют на эффективность обучения. Так, с увеличением числа примеров возрастает полная производительность сети, но это происходит только до определенного момента.⁵⁰ Воздействие состава тренировочного множества оказывается гораздо более сложным. Исследованию влияния соотношения между присутствием и отсутствием структурных признаков в молекулах, т.е. соотношения в тренировочном множестве соединений с данным структурным признаком и без него, посвящены работы^{49, 51, 52}. В частности, авторы статьи⁴⁹ обнаружили, что увеличение степени присутствия заданных структурных признаков в тренировочном множестве повышает вероятность того, что эти фрагменты будут опознаны в анализируемом веществе, однако при этом снижается возможность правильного определения отсутствия фрагмента. Вместе с тем оказалось, что есть некоторое предельное значение числа соединений, содержащих данный фрагмент, при котором сеть не в состоянии производить хорошую классификацию. Формирование тренировочного множества, хорошо сбалансированного и содержащего разнообразные примеры, может оказаться затруднительным, особенно если сеть, подлежащая обучению, имеет много выходов.

Число выходов сети имеет большое значение: по данным работ^{50, 53} сети с одним выходом работают лучше, чем сети с многими выходами, более того, сети большего размера труднее поддаются оптимизации, чем малые сети, имеющие меньше выходов. Поскольку наилучший результат для каждого отдельного выхода может быть получен при специфическом только для него времени обучения, то процесс обучения ИНС останавливается на такой стадии, когда производительность сети будет оптимальна только в среднем, но не для каждого выхода. Если для обнаружения большого числа структурных признаков использовать много сетей, каждая из которых имеет один выход, то потребуется значительно больше времени, чем в случае обучения одной единой сети. Следовательно, достоинства и недостатки единой сети со многими выходами прямо противоположны достоинствам и недостаткам сетей с одним выходом.

Авторы работы⁵⁴ предложили иерархическую систему для спектроскопических исследований, которая по их замыслу должна была вобрать в себя все лучшее от обоих подходов. В такой системе сеть высшего уровня с несколькими выходами производит грубую классификацию спектров. В зависимости от рассматриваемого структурного признака иерархия распространяется на более глубокие уровни, на которых производится дальнейшее уточнение.

Выходы сети высшего уровня активизируют сети первого, затем второго уровня и так далее на всю глубину иерархии. Композицию иерархической системы нейронных сетей можно рассматривать как компромисс между одной сетью с большим числом выходов и многими сетями с одним выходом. Ее главные характеристики являются также промежуточными. Обнаружилось, что установить баланс составляющих тренировочного множества для каждой отдельной сети иерархии оказывается легче, чем для единой большой сети (в известной степени так же просто, как в случае сети с одним выходом). Кроме того, оптимизировать время обучения проще для меньшего числа выходов, а блочная концепция позволяет модифицировать систему без полной ее реконструкции.

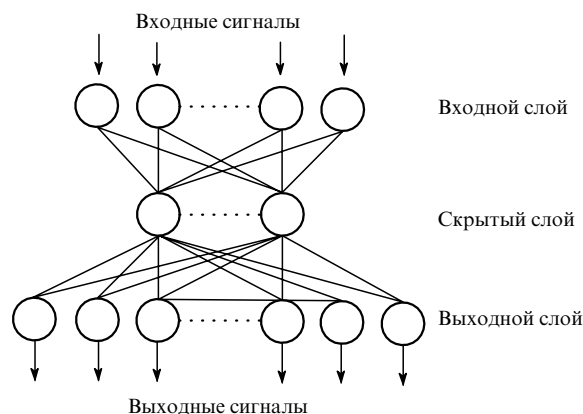


Рис. 1. Схема нейронной сети.

Иерархическая система дает информацию о структурных признаках с разной степенью подробности: от общих характеристик («соединение содержит группу СО») до более частных («соединение содержит кислотную группу»), причем число сетей, которые нужно активировать в каждом частном случае, остается небольшим. Например, нет необходимости выяснять, является ли данное вещество первичным или третичным спиртом, если группа ОН не была распознана сетью высшего уровня. Однако авторы работы⁴⁹ отмечают, что настройка иерархической сети является сложной операцией, требующей предварительных знаний и учета зависимости действия нейронных сетей нижнего уровня от производительности сетей верхнего уровня.

В качестве примера приведем результаты исследования иерархической системы как инструмента для обнаружения фрагментов по ИК-спектрам.⁴⁹ Для проведения экспериментов было отобрано 8620 соединений и их ИК-спектров, взятых из библиотеки Sadtler Bio-Rad. Расчеты проводились на рабочей станции RS6000 IBM. Система обучалась опознаванию тридцати трех структурных групп, иерархию которых иллюстрирует рис. 2. Видно, что для выявления карбонильной группы было предназначено две сети: одна для широко распространенных структурных признаков (карбоксил, кетон, сложный эфир и амид) и другая для реже встречающихся признаков (хлорангидрид кислоты, ангидрид, альдегид).

Использовались следующие параметры сети.

Входной вектор. ИК-Спектр в интервале 500–3596 см⁻¹ был представлен 259 точками через 12 см⁻¹. В качестве входа брали среднее взвешенное поглощение в окрестности каждой точки, задаваемое величиной, изменяющейся от 0 до 1.

Скрытые вершины. Сети включали только один скрытый слой, причем для всех сетей было принято 30 скрытых вершин.

Функция передачи. Использовалась классическая сигмоидальная функция.

Кодирование структуры: 0 — отсутствие, 1 — присутствие структурного признака.

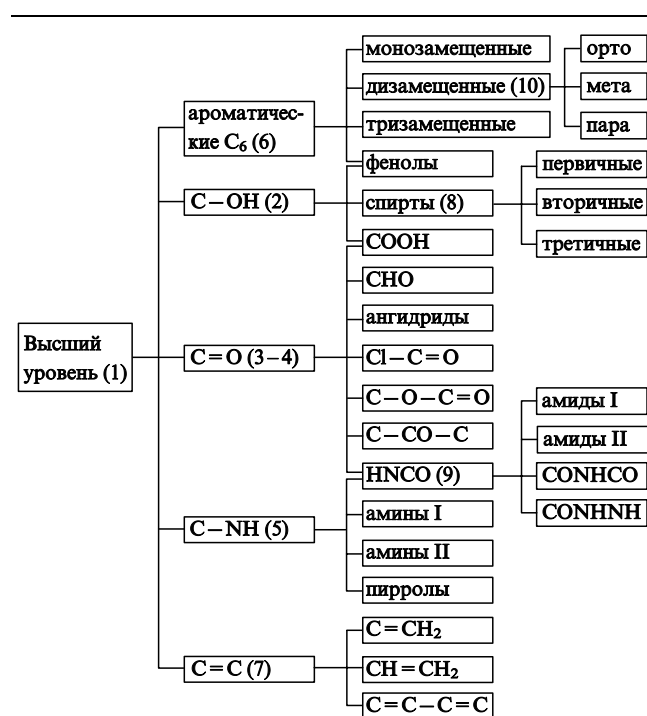


Рис. 2. Иерархическая организация сетей.
В скобках указаны числа активированных сетей.

Обучающий алгоритм. Все сети были обучены с помощью алгоритма ОРО при значении параметра, определяющего скорость обучения, $\alpha = 0.05$, и моменте $\mu = 0.90$ (см. ²⁴).

Принятие решения. Так как выходы сетей представляли собой вещественные значения в интервале 0–1, то использовались два порога для принятия решения о наличии или отсутствии в молекуле каждого структурного признака f : AL_f — порог присутствия, RL_f — порог отвержения признака. Если число на выходе оказывается между RL_f и AL_f , то ответ считается неопределенным и фрагмент не классифицируется. Такие пороги принятия решения были оптимизированы отдельно для каждой сети.

Формирование тренировочных и тестовых множеств. Для обучения 10 сетей, образующих иерархическую систему, методом случайного выбора из исходной совокупности были сформированы тренировочные множества с таким расчетом, чтобы процент соединений, содержащих заданный признак, был близок к заранее фиксированному значению. Для обучения и тестирования сети высшего уровня была использована вся коллекция спектров. Обучение сетей промежуточных уровней проводили по специально разработанным протоколам.

Как отмечалось выше, в иерархической структуре сеть, предназначенная для обнаружения фрагментов данного класса, активируется только если в том случае, она распознала предъявленный образец как элемент этого класса. Возможны два типа неверной классификации структурного признака: 1) фрагмент идентифицирован в соединении, где он на самом деле отсутствует; этот случай называют ложным присутствием (AdP); 2) фрагмент не идентифицирован в соединении, где он на самом деле присутствует; этот случай называют ложным отсутствием (PdA). Ложные отсутствия не распространяются через иерархию, так как признак помещается как отсутствующий. Однако ложные присутствия приводят к тому, что сети промежуточных уровней должны уточнять отнесение к классу фрагмента, на самом деле отсутствующего в соединении.

Специальные эксперименты показали, что иерархическая система является устойчивой и выдерживает изменения сетей без потери глобальных способностей иерархии. Обучение было наиболее успешным, когда тренировочное множество содержало 1000 примеров, а число циклов находилось в пределах от 5 до 20. Особенно хорошие результаты показаны для групп C=O и ОН, которые хорошо распознаются по ИК-спектрам.

Для оценки производительности иерархической системы как целого предложены следующие показатели, характеризующие иерархию:⁴⁹

Ptf — общая характеристика присутствия структурного признака, определяется как отношение числа соединений, в которых правильно идентифицировано присутствие фрагмента, к полному числу соединений, имеющих данный структурный признак;

Atf — то же применительно к отсутствующим структурным признакам;

Pte — итоговая ошибка принятия решения о присутствии структурного признака, определяется как отношение числа соединений, в которых признак присутствует, но идентифицирован как отсутствующий, к общему числу соединений, в которых он присутствует;

Ate — итоговая ошибка принятия решения об отсутствии признака, определяется аналогично предыдущему показателю.

Результаты тестирования системы с использованием введенных критериев представлены в табл. 1. Сопоставление их с данными, полученными в работе⁵⁵, показывает, что в общем иерархическая система работает лучше, чем сеть с многими выходами, однако результаты качественно уступают тем, которые достигаются на хорошо оптимизированных сетях с одним выходом.

Таблица 1. Производительность глобальной иерархии.

Фрагмент	Atf	Ptf	Ate	Pte
C—OH	93.3	88.3	6.7	16.4
CO—OH	91.2	79.8	1.0	18.0
Ph—OH	92.8	56.4	0.8	32.3
C _{sp} ³ —OH	91.3	78.6	1.1	18.2
CH ₂ —OH	91.1	71.5	0.7	27.2
CH—OH	90.6	57.2	1.2	41.9
C _{четв.} —OH	91.6	35.1	0.1	62.9
C=O	94.3	91.8	3.1	5.5
CO—OH	94.1	82.8	1.3	13.7
CO—OR	93.6	77.0	1.6	15.2
CO—R	92.9	50.7	1.7	39.0
CO—NH	91.5	85.5	3.0	11.8
CO—NH ₂	93.4	87.0	2.0	10.5
CO—NHR	94.9	55.5	0.8	35.8
CO—NH—CO	95.6	74.2	0.1	21.4
CO—NH—NH	95.6	47.6	0.3	52.4
CHO	86.8	74.0	6.1	22.1
OC—O—CO	92.2	82.8	1.2	10.0
Cl—CO	94.6	47.0	1.7	48.0
C—NH	92.0	82.7	6.6	15.6
C—NH ₂	91.1	52.4	3.0	43.0
C—NHR	93.4	36.3	1.8	60.3
CO—NH	91.6	66.0	1.3	29.8
Пиррол	93.8	14.3	0.4	79.8
Ароматические C ₆	79.0	86.4	16.3	5.9
C ₆ —OH	80.9	69.9	2.9	29.7
мономзамещенные	80.8	54.0	2.1	35.0
дизамещенные	70.0	53.3	10.7	37.8
орто	73.9	34.9	0.2	62.8
мета	73.7	16.7	0.3	83.3
пара	71.7	47.6	1.1	49.0
тризамещенные	76.4	34.8	6.0	59.1
C=C	93.6	42.6	6.3	56.9
C=CH ₂	93.2	30.4	0.9	68.7
CH=CH ₂	93.3	23.0	0.3	75.8
C _{sp} ² —C=C	89.4	34.4	1.8	62.0

Публикаций, посвященных применению ИНС для отбора молекулярных фрагментов пока немного, но исследования в этом направлении ведутся с нарастающей интенсивностью. Так, анализу ИК-спектров с помощью ИНС посвящены работы^{52–61}, в работе⁶² ИНС использовали для идентификации аминокислотных остатков в белках по «отпечаткам пальцев» в спектрах 2D ЯМР, а в исследованиях^{51,63} сети настраивались на совместное применение разных спектральных методов для выявления структурных фрагментов. Это направление представляется особенно интересным, поэтому рассмотрим подробнее результаты двух последних работ.

Авторы статьи⁵¹ провели исследование, в котором ИК-спектры, спектры ЯМР на ядрах ¹³C и молекулярные формулы органических веществ подавались на вход нейронной сети, предварительно обученной распознаванию молекулярных фрагментов. Использовалась ИНС, содержащая 512 входов и 85 выходов с числом нейронов в скрытом слое ~ 50. Тренировочное множество состояло из 1560 соединений, для которых были известны ИК-спектры и спектры ЯМР на ядрах ¹³C. Для ввода ИК-спектров было отведено 128 входов, для ЯМР-спектров — 294, а для молекулярных формул — 90. Такое распределение выбрано с учетом оценки объема информации, содержащейся в каждом типе данных. ИК-Спектры кодировали таким же образом, как и в работе⁴⁹. Спектры ЯМР на ядрах ¹³C кодировали признаками наличия или отсутствия сигнала на интервалах 2 м.д., причем для сигналов разной мультиплетности отводились определенные позиции входного вектора (например, для

кодирования химических сдвигов кватертов в диапазоне спектра 0–64 м.д. отводилось 32 интервала, и им соответствовали позиции 128–159 входного вектора). Молекулярная формула кодировалась числами присутствующих типов атомов, степенью ненасыщенности и молекулярной массой. Структурные формулы кодировались 0 или 1 в зависимости от наличия или отсутствия в молекуле каждого из 85 фиксированных структурных элементов, среди которых были функциональные группы ($-\text{O}-\text{C}=\text{O}$, $\text{C}\equiv\text{N}$, $\text{H}_2\text{C}=\text{CH}$ и

т.д.) и химические классы (стероиды, терпены, пиридины и т.д.). Оптимизация параметров и обучение сети с помощью процедуры ОРО производились на рабочих станциях (HP 9000 RISC и др.). Производительность системы изучалась при разных комбинациях исходных данных.

В результате выполненного исследования установлено, что при использовании только спектров ЯМР производительность системы выше, чем при использовании только ИК-спектров; учет молекулярной формулы повышает производительность системы при любом сочетании данных ИК- и ЯМР-спектроскопии; совместное использование спектров обоих видов значительно повышает производительность системы (т.е. заметно увеличивается число групп, которые идентифицируются с высокой достоверностью).

Полученные выводы не были неожиданными, однако значение работы⁵¹ заключается в том, что в ней предложена методика идентификации фрагментов при комплексном использовании ИК-спектров и спектров ЯМР на ядрах ¹³C, основанная на применении ИНС. Эту методику можно легко реализовать в блоке отбора фрагментов в ЭС.

Применение ИНС для совместной интерпретации ИК- и масс-спектров (МС) газов описано в работе⁶³. Сеть обучалась идентификации 26 функциональных групп. При использовании комбинированных данных доля правильно идентифицированных присутствующих групп составляла 86.4%; когда данные ИК-спектроскопии и масс-спектрометрии использовались отдельно, этот показатель оказался равным соответственно 88.4 и 78.2%. Только для 8 из 26 групп применение МС повысило качество распознавания, причем результаты заметно улучшились при определении связи C—Cl. Это подтверждает известный факт, что для выявления функциональных групп больше подходит ИК-спектроскопия, чем масс-спектрометрия. Предсказание отсутствия функциональных групп проводилось с точностью 95.5% в случаях использования комбинированных данных и данных только ИК-спектроскопии, а предсказание в случае использования только данных масс-спектрометрии было хуже — 87%. Работа⁶³ показывает, что принципы комбинирования данных, полученных различными спектральными методами, применительно к ИНС требуют специального изучения с целью выявления оптимальных вариантов.

Необходимо отметить, что разными группами исследователей, как правило, используются разные показатели производительности сети, поэтому сравнение результатов работ затруднено. В связи с этим весьма актуальным представляется установление каких-либо общепринятых показателей и стандартных тестовых множеств, что позволило бы легко сравнивать эффективность разных методов классификации.

Несмотря на большие возможности ИНС для обнаружения фрагментов, успешное применение сетей в значительной степени зависит от искусства пользователя, заключающегося в его умении так настроить параметры сети (пороги принятия решения, состав тренировочного множества и т.д.), чтобы они удовлетворяли конкретным требованиям при решении каждой задачи. Пока нет публикаций об использовании ИНС в какой-либо действующей ЭС, предназначенной для установления структуры молекул, и неясно, какое число структурных признаков может быть опознано с помощью ИНС (для сравнения заметим, что спектральные библиотеки известных ЭС содержат сотни фрагментов). Поэтому оценка

реальных возможностей этого подхода еще впереди. Однако с уверенностью можно утверждать, что использование ИНС для обнаружения отдельных классов химических соединений является весьма продуктивным.

4. Методы опознания крупных фрагментов

Чтобы ЭС справилась с задачей идентификации крупных молекул, она должна иметь средства выявления фрагментов большого размера. Попытки создания библиотек, содержащих большие фрагменты и интервалы изменения их характеристических спектральных признаков, совершенно бесперспективны из-за неизбежности комбинаторного взрыва и практической непредсказуемости интервалов. Реальный путь решения задачи — использование современных баз данных, которые содержат структурные формулы веществ и их спектры. В основе метода лежит гипотеза о том, что похожим структурам (подструктурам) соответствуют похожие спектры (подспектры).

В литературе описано два подхода к выявлению крупных фрагментов с использованием баз данных по ИК-, ЯМР- и масс-спектрам. В первом подходе для этой цели предлагается алгоритм, который сводится к следующим трем основным операциям:^{64–72} 1) поиск структур, имеющих спектры, похожие на предъявленный экспериментальный, и ранжирование их по убыванию степени сходства; 2) отбор некоторого числа структур в начале списка, имеющих, согласно основной гипотезе, наибольшее сходство со структурой анализируемого соединения; 3) выделение в этих структурах максимальных общих фрагментов, которые наиболее вероятно могут входить в структуру неизвестного соединения. Вероятность выявления правильного фрагмента возрастает, если информационно-поисковая система производит отбор по спектрам нескольких видов, а отобранные фрагменты проверяются по совпадению предсказанных для них подспектров (последнее легче всего можно осуществить по спектрам ЯМР на ядрах ^{13}C).

Реализация изложенного алгоритма требует создания надежных количественных критериев сходства структур и похожести спектров, а также методов выявления общего наибольшего подграфа в множестве химических графов. Разработки таких методов сформировали отдельное направление в теории графов, и по этому вопросу уже имеются многочисленные публикации.

Другой метод опознания крупных фрагментов получил практическую реализацию благодаря появлению больших баз данных по спектрам ЯМР на ядрах ^{13}C . В базах такого типа структуры описаны «сферическим» кодом,⁷³ а сигналы в спектрах отнесены к соответствующим атомам углерода. Это позволило авторам ряда работ (см., например,^{30, 31, 74}) создать алгоритмы, позволяющие автоматически формировать библиотеки корреляций подструктура–подспектр, которые и используются для решения поставленной задачи. Вначале программа сравнивает все подспектры библиотечных фрагментов с экспериментальным ЯМР-спектром анализируемого вещества, причем существенно, что в спектре помимо химического сдвига и мультиплетностей сигналов должна быть также указана интенсивность сигнала (т.е. число атомов углерода). В результате отбираются только подструктуры (фрагменты), которым соответствуют подспектры, совпадающие в пределах заданных отклонений с подспектрами запроса. Далее путем более детального анализа спектральной информации отбираются наиболее вероятные фрагменты, которые ранжируются по убыванию числа атомов углерода и скелетных атомов. Таким образом, самые крупные фрагменты оказываются в начале списка (использование описанного приема в ЭС рассмотрено ниже).

III. Генерация структурных формул изомеров

За годы, прошедшие со времени создания первых генераторов изомеров,^{3, 75–78} опубликовано не более двух десятков работ, посвященных поиску путей повышения быстродействия программ и методов более эффективного использования структурных ограничений. Сравнительно небольшое число работ свидетельствует не о потере актуальности этой проблемы и интереса к ней, а о больших трудностях, стоящих на пути ее решения. С увеличением числа скелетных атомов и фрагментов, участвующих в процессе генерации (их принято называть дискретными единицами структуры (ДЕС)),⁷ число возможных структурных формул и процессорное время растут экспоненциально. Поэтому усилия были направлены на то, чтобы использовать структурные ограничения уже в процессе генерации структур (а не после) и на разработку новых алгоритмов, позволяющих повысить эффективность генерации.

Эта интересная и амбициозная задача привлекла внимание математиков — специалистов в области теории графов и комбинаторики. В результате появились новые подходы и новые эффективные программы (COCOA,⁷⁹ MOLGEN,^{80–82} GENM^{83–86}). Параллельно с поисками эффективных алгоритмов генерации происходил стремительный рост быстродействия компьютеров, поэтому в целом размерность задач, с которыми способны работать современные генераторы изомеров, заметно увеличилась. Так, если в 1970-е годы программы могли за практически приемлемое время выполнять генерацию из 7–8 ДЕС, то в настоящее время этот интервал составляет 10–16 ДЕС в зависимости от валентности атомов, рассматриваемых как ДЕС, числа свободных связей фрагментов, степени ненасыщенности молекулы и ряда других факторов. Например, первой в нашей стране генераторной программой, написанной В.В.Разниковым в кодах ЭВМ М-20,⁷⁵ потребовалось 3420 с для построения 4347 изомеров состава $\text{C}_{15}\text{H}_{32}$, а современная программа GENM, установленная на ПК Pentium, справляется с этой задачей за несколько секунд. Другой пример: фортрановская программа МАИСС на ЭВМ Минск-32 генерировала 217 изомеров C_6H_6 за ~ 30 мин;⁷⁷ программа GENM делает это за доли секунды, т.е. реальная скорость генерации структур повысилась на несколько порядков. Но так как для практики наибольший интерес представляет распознавание структур крупных молекул, то несмотря на достигнутые успехи, стадия генерации остается критической для всех ЭС.

Основные требования к генератору структур были сформулированы и реализованы еще в 1970-е годы. Результаты этих исследований обобщены в работах^{7, 8}. Важнейшее из требований — независимо от применяемого алгоритма генерации, программы, пригодные для практического использования, должны обладать гибкими средствами наложения структурных ограничений. Типичными ограничениями являются следующие.

1. Фрагменты — дискретные единицы структуры обнаруживаемые при СГА. Обычно предполагается, что они не имеют пересекающихся частей, а число свободных валентных связей концевых атомов в них фиксировано. Поэтому такие фрагменты называют также макроатомами, они участвуют в генерации структур наряду со свободными атомами. Из атомов и макроатомов строят наборы ДЕС, причем суммарное число ДЕС в наборе определяет размерность задачи. В некоторых программах, например в GENOA,⁸⁷ допускается частичное наложение макроатомов, что, однако, усложняет алгоритм и увеличивает время генерации.

2. Обязательные и запрещенные фрагменты (соответственно ГУДЛИСТ и БЭДЛИСТ) — это накладываемые химиком структурные ограничения, которым должна удовлетворять каждая структура. Ограничения такого рода задаются из априорных знаний об исследуемом веществе

(происхождение образца, данные не спектральных экспериментов и т.д.). Особенность этих фрагментов состоит в том, что их структуры могут взаимно перекрываться и накладываться на структуры ДЕС.

3. Ограничения на размеры циклов и кратности связей позволяют формулировать общие требования к генерируемым структурам с учетом специфики набора ДЕС и априорной информации (наличие в макроатомах кратных связей, циклов, учет степени ненасыщенности молекулы и т.д.).

4. Фрагменты, маловероятные в органической химии. К числу таких фрагментов относятся малые циклы с тройными связями, структуры, противоречащие правилу Бредта, некоторые напряженные конденсированные структуры с малыми циклами, фрагменты типа $-O-O-O-$ и т.д.

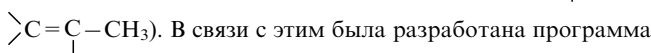
Поскольку «генераторостроение» стало специальной областью математической химии и комбинаторной математики, ограничимся лишь упоминанием некоторых новых программ, представляющих наибольший интерес для построения ЭС. Достаточно полную библиографию по современным алгоритмам генерации структур можно найти в работах^{88,89}.

Генератор CONGEN, разработанный в 1970-е годы,⁹⁰ основан на алгоритме генерации путем наращивания химического графа от его центра. В результате усовершенствования этого алгоритма в течение ряда лет были разработаны сначала программа MOLGRAPH,⁸⁰ а затем MOLGEN.^{81–83} Программа MOLGEN производит генерацию из неперекрывающихся фрагментов. Она отличается высокой производительностью и способна генерировать стереоизомеры. Этот генератор в настоящее время получил широкое распространение в Европе.

В нашей стране создана программа генерации структурных изомеров GENM,^{84–86,88} которая по своим параметрам несколько не уступает, а в некоторых случаях превосходит программу MOLGEN. Алгоритм содержит ряд оригинальных решений, позволяющих в соответствии с наложенными ограничениями отсекал бесперспективные ветви решения в процессе генерации. Опыт применения генератора GENM в НИОХ СО РАН, а также в системах РАСТР³⁶ и X-PERT²⁹ показал его высокую надежность и быстродействие.

Программа SMOG^{89,91} позволяет химику-органику анализировать различные структурные гипотезы в процессе исследования химических реакций. Эта программа способна генерировать структуры, содержащие атомы в необычных валентных состояниях или несущие формальный заряд, конструировать молекулы с участием как неперекрывающихся, так и перекрывающихся фрагментов и распознавать среди результатов генерации различные резонансные формы одной и той же ароматической молекулы. SMOG вполне может послужить основой для создания автоматизированного рабочего места химика-органика, однако быстродействия программы не достаточно, чтобы ее использовать в качестве «конвейерного» генератора структур в ЭС.

Когда фрагменты выявляются при независимом использовании разных методов исследования (ИК, ЯМР, МС и т.д.), может оказаться, что структуры некоторых фрагментов содержат общие части (например, $-CH=CH-C=O$ и



GENOA,⁸⁷ способная генерировать структуры из перекрывающихся фрагментов. Если в случае GENOA эта способность означала расширение возможностей программы CONGEN, то для ряда появившихся позднее программ^{30,31,74} наложение атомов, принадлежащих разным фрагментам, является основным условием выполнения генерации. Алгоритмы этого типа ориентированы на генерацию структур из набора фрагментов, обнаруживаемых программой по спектрам ЯМР на ядрах ^{13}C (см. раздел II.4). Программа пытается совместить атомы, принадлежащие

разным фрагментам, и предсказать подспектр нового, укрупненного фрагмента. Если предсказанный подспектр в пределах заданной величины допустимого отклонения не противоречит экспериментальному спектру, то делается попытка присоединить еще один фрагмент и т.д.

Таким образом, генерация структур представляет собой их последовательную сборку, в процессе которой производится проверка правильности всех возможных комбинаций фрагментов. При таком подходе знание брутто-формулы и даже молекулярной массы необязательно для выполнения генерации. Однако отсутствие брутто-формулы приходится компенсировать тем, что каждому сигналу экспериментального ЯМР-спектра должно быть приписано определенное число атомов углерода (это не всегда удается сделать, особенно, если молекула симметричная). В этом алгоритме в неявном виде работает $C_nH_{(m-k)}$ -часть брутто-формулы, где k — число атомов водорода при гетероатомах, сведения о которых «вытягиваются» в результате генерации структур. Однако если отобранные фрагменты не исчерпывают полную структуру, то генерация оказывается безуспешной. Применение генераторов этого типа подробно рассмотрено ниже.

Можно указать еще ряд работ (например,^{92–97}), в которых были опробованы различные принципы генерации. Оценить их истинные эффективность и быстродействие можно только путем решения тестовых задач на однотипных компьютерах, что представляется затруднительным.

В ряде случаев распознавание структуры не ограничивается выяснением структурной формулы вещества, а включает в себя также установление пространственной модели молекулы. С этой целью были разработаны алгоритмы и программы генерации стереоизомеров,^{83,98,99} а также программы построения пространственных моделей молекул.^{99–103} Исходной информацией для этих программ являются структурные формулы органических соединений. Построение пространственных структур может оказаться необходимым для прогнозирования спектров структур-кандидатов, и в этом случае построение пространственных моделей становится важной стадией процесса выявления истинной структурной формулы анализируемого вещества. Следует отметить, что доступ к одному из наиболее эффективных 3D-генераторов CORINA¹⁰¹ возможен через систему ИНТЕРНЕТ (<http://www.2.ccc.uni-erlangen.de/servies/3d.html>).

IV. Прогнозирование спектров вероятных структур

Заключительной стадией работы ЭС является предсказание спектров изомеров с целью выявления наиболее вероятных структур. Практика показывает, что ответный файл в зависимости от жесткости налагаемых структурных ограничений может содержать десятки, сотни или даже тысячи структурных формул, которые нельзя различить по интервалам спектральных характеристических признаков. Очевидно, что для этой цели требуется прогнозирование полных молекулярных спектров. В этом случае трудоемкие квантовые расчеты неприменимы, и речь может идти только о быстром приближенном предсказании спектров с использованием различных эмпирических корреляций, правил и параметров, а также спектральных баз данных. Опубликовано достаточно много работ, посвященных этой проблеме. Ввиду ее большой важности для методологии ЭС рассмотрим подходы к прогнозированию ИК-спектров и спектров ЯМР на ядрах ^{13}C и 1H .

1. Моделирование спектров ЯМР на ядрах ^{13}C

Спектроскопия ЯМР на ядрах ^{13}C — один из наиболее мощных методов установления структуры органических

молекул и проверки структурных гипотез. Из всех видов молекулярных спектров легче всего удается моделировать именно спектры ЯМР на ядрах ^{13}C , так как при неполной развязке от протонов они состоят из синглетов, дублетов, триплетов и квартетов (взаимодействие спинов ^{13}C и ^1H через одну связь), причем каждому анизохронному (магнитно эквивалентному) атому углерода соответствует сигнал с определенным химическим сдвигом и мультиплетностью. Применение режима DEPT обычно позволяет надежно определять мультиплетности сигналов даже при их сильном наложении. Можно выделить несколько уровней подробности предсказания спектров ЯМР на ядрах ^{13}C .

а. Предсказание числа сигналов разной мультиплетности

В работе ¹⁰⁴ впервые предложено использовать предсказание числа сигналов разной мультиплетности для отбраковки структур-кандидатов. Метод основан на топологическом анализе мультиграфа, отображающего структурную формулу молекулы. Все скелетные атомы разбиваются на группы топологически эквивалентных атомов. Если при этом окажется, что все атомы углерода в структуре уникальны, то прогнозируемое число сигналов в спектре равно числу атомов углерода в молекуле, при этом автоматически предсказывается число мультиплетов разного типа. Если же структура содержит несколько групп эквивалентных атомов углерода, то ожидаемое число сигналов в спектре равно числу групп. Авторы статьи ¹⁰⁴ проанализировали влияние стереохимических факторов на правильность предсказания и предложили ряд правил, позволяющих учитывать расщепление сигналов диметильных групп, находящихся вблизи хирального центра. Простое сравнение предсказанного и экспериментального чисел мультиплетов разного сорта позволяет существенно сократить число кандидатов. Однако при этом следует иметь в виду, что в спектрах возможно случайное совпадение неэквивалентных сигналов или появление расщеплений, не предусмотренных правилами. Для усовершенствования указанного подхода было предложено ²⁹ предсказывать числа сигналов различной мультиплетности в характеристических интервалах спектра, что повысило дискриминирующую способность метода.

б. Предсказание химических сдвигов

Моделирование спектров ЯМР на ядрах ^{13}C основано на предсказании химических сдвигов атомов углерода, принадлежащих данной структуре. Полученный модельный спектр сравнивается с экспериментальным, и заключение о правдоподобности структурной гипотезы делается на основании вычисленных средних или стандартных отклонений. Существуют два наиболее распространенных способа предсказания спектров ЯМР: один основан на построении эмпирических моделей, ^{105–110} а второй — на использовании баз спектральных данных. ^{111–115} В последнее время появились программы, ^{116–118} использующие оба подхода, а в исследованиях ^{119–128} описаны попытки моделирования спектров ЯМР на ядрах ^{13}C с использованием ИНС.

В работе ¹²⁹ впервые предложена основанная на правилах аддитивности линейная модель для расчета химических сдвигов атомов углерода в алифатических углеводородах с использованием инкрементов, учитывающих влияние окружения вплоть до δ -атомов. В дальнейшем линейные модели были построены для алканов, циклогексанов, спиртов, аминов, хлорсодержащих соединений и т.д. Полученные расчетные формулы и таблицы инкрементов обобщены в ряде монографий (см., например, ^{130, 131}). Описанный прием обеспечивает достаточно высокое качество предсказаний, поэтому были разработаны программы, которые анализируя предъявленную структурную формулу, выбирают для каждого атома углерода соответствующую аддитивную

модель и рассчитывают химический сдвиг. Недостаток таких программ состоит в том, что они предсказывают химические сдвиги только для атомов углерода в состоянии sp^3 -гибридизации, а инкременты невозможно модифицировать.

Эти недостатки были устранены в работе ¹¹⁰. Используя большие базы данных по структурам и отнесенным спектрам ЯМР на ядрах ^{13}C , авторам удалось найти линейные модели, которые применимы ко многим классам химических соединений, причем модели содержат параметры, зависящие от конфигурации и конформации. При этом принимается во внимание конфигурация $\text{C}=\text{C}$ -связей и наличие аксиальных и экваториальных заместителей циклогексанового кольца. Программа допускает модификацию параметров (базовых величин и инкрементов), а также ввод новых аддитивных правил. Проверка программы на множестве из 170 000 известных химических сдвигов показала, что заложенные в нее правила выполняются в 97% всех случаев. Стандартное отклонение предсказанных химических сдвигов от экспериментальных значений составило 5.5 м.д., а время расчета спектра молекул среднего размера (C_{12} – C_{16}) на современных ПК, по нашему опыту, занимает не более секунды. Очевидно, что такая программа вполне может использоваться для фильтрации ответных файлов, найденных ЭС.

В другом подходе ^{111–115} используются базы данных, содержащие структуры с указанием химических сдвигов атомов углерода. Для всех атомов углерода каждого эталонного соединения из базы данных генерируются атом-центрированные фрагменты (АЦФ) с заданным числом concentрических слоев. Эти фрагменты вместе с соответствующими химическими сдвигами хранятся в виде упорядоченного списка. Чтобы предсказать спектр предполагаемой структуры, программа выделяет в ней все возможные АЦФ, разыскивает в базе их аналоги и приписывает атомам углерода химические сдвиги, взятые из базовых фрагментов. Если некоторый АЦФ не найден в базе, то программа обычно производит «интерполяцию», используя имеющиеся «наиболее похожие» структуры. Этот подход дает очень хорошие результаты при использовании большой базы, отличающейся разнообразием структур (см., например, ⁷³). Обычно расхождение между предсказанными и экспериментальными химическими сдвигами находится в пределах 1 м.д., т.е. точность прогноза весьма высока.

Однако при всех несомненных достоинствах описанный подход имеет и ряд недостатков. Если база содержит только структурные формулы, но в ней отсутствует стереохимическая информация, то это может сильно повлиять на качество предсказания. Как показано в работе ¹³², игнорирование стереохимических эффектов может привести к различию предсказанных и реальных химических сдвигов более чем на 10 м.д. При экстраполяции использованные базовые фрагменты по своим спектральным свойствам могут оказаться не самыми близкими к фрагментам гипотетической структуры. Кроме того, при кодировании не учитывается тот факт, что атом, удаленный на одинаковое расстояние от другого атома, в разных соединениях может оказывать разное влияние. Наконец, время, необходимое для предсказания спектра, в зависимости от сложности молекулы оказывается относительно большим (десять секунд и даже минуты).

В работе ¹¹⁷ описан метод, представляющий собой комбинацию обоих упомянутых выше подходов. Окружения атома углерода в базовых и целевых (проверяемых) структурах кодируются вектором, и на первой стадии рассчитывается евклидово расстояние между векторами, описывающими окружение атома в целевой и базовых структурах. Результатом этой процедуры является формирование калибровочного множества атомов углерода, похожих по химическому окружению на атом углерода целевой молекулы. При обнаружении высокой степени сходства окружения целевому атому приписывается химический сдвиг базового

«двойника». В противном случае для построения линейной расчетной модели используется калибровочное множество. С этой целью вводятся дескрипторы, отражающие топологические, геометрические и электронные свойства окружения. Для построения моделей используются методы множественной линейной регрессии (МЛР), дробных наименьших квадратов^{28, 133} и анализ основных компонентов.²⁸ Проверка комбинированного подхода на тестовом множестве из 38500 атомов показала, что среднее отклонение составляет 1.69 м.д., а в случае прямого извлечения данных из базы эта величина равна 1.85 м.д.

Выше отмечалось, что ИНС могут быть использованы не только для структурной интерпретации спектров, но и для предсказания спектральной картины, отвечающей данной структуре. В ряде работ были сделаны попытки применить ИНС для прогнозирования химических сдвигов в ЯМР спектрах соединений отдельных классов — метилзамещенных циклогексанов,¹²¹ монозамещенных бензолов,¹²² моносахаридов,¹²⁵ рибонуклеазидов,¹²⁶ sp^2 - и sp^3 - α -атомов в ациклических алкенах^{127, 128} и т.д.

Для обучения ИНС предсказанию спектров соединений заданных классов формируется тренировочное множество, в которое включают структуры, закодированные способом, аналогичным описанному в разделе II.3. В процессе обучения на входы сети подаются закодированные эталонные структуры, а сигналы на выходе сравниваются с спектрами ЯМР на ядрах ^{13}C этих структур. Обучение заканчивается, когда отклонение предсказанных спектров от эталонных достигает некоторой приемлемой величины. Рассмотрим в качестве примера результаты, полученные в работе¹²⁸.

Авторы провели обучение сети предсказанию химических сдвигов sp^3 -атомов углерода, находящихся в α -положении к двойной связи в ациклических алкенах. Параллельно строилась модель, основанная на МЛР. Тестирование обеих моделей показало, что стандартное отклонение в случае ИНС находится в пределах 1.3–1.5 м.д., в то время как для модели МЛР оно составляет 3 м.д., т.е. ИНС способны более точно предсказывать спектры, чем линейные модели.

Использование ИНС возможно только для предсказания спектров соединений тех классов, на которых сеть была обучена, в то время как структуры ответных файлов, выдаваемых ЭС, могут содержать самые неожиданные атомные группировки. На наш взгляд, оптимальная стратегия применения прогнозирования спектров ЯМР на ядрах ^{13}C состоит в том, чтобы проводить его в два этапа: сначала следует осуществить скрининг структур, используя быстрые приближенные методы прогнозирования (например, линейные модели¹¹⁰), затем использовать более точное (ориентированное на химические классы) прогнозирование применительно к тем структурам, для которых среднее или стандартное отклонение находится в пределах погрешности расчетного метода, и, следовательно, требуется дальнейшее уточнение.

2. Предсказание спектров ЯМР на ядрах ^1H

Как отмечалось выше, сравнительная простота предсказания спектров ЯМР на ядрах ^{13}C связана с тем, что они состоят из линий с четко выраженной мультиплетностью. Прогнозировать спектры ЯМР на ядрах ^1H гораздо сложнее, так как наряду с сигналами первого порядка они содержат мультиплеты высших порядков. По этой же причине затруднено автоматическое сравнение экспериментальных и предсказанных спектров. Поэтому публикаций по методам прогнозирования спектров ЯМР на ядрах ^1H сравнительно немного.

Программа Proton Shift, разработанная¹³⁴ для предсказания химических сдвигов в спектрах ЯМР на ядрах ^1H на основе простых линейных моделей,¹³¹ автоматически определяет подструктуры, для которых применимы определенные правила, а остальную часть молекулы идентифицирует

как набор заместителей при каждой выбранной подструктуре. Чтобы расширить область применения программы, использовали различные приемы экстраполяции, включая расчленение больших подструктур и замену недостающих заместителей родственными подструктурами. Программа учитывает *цис-транс*-изомерию двойных связей, а также аксиальное и экваториальное расположение заместителей в шестичленных циклах. Линейные модели построены на базе большого набора инкрементов, причем общее число расчетных параметров близко к 3000. Проверка на ограниченном множестве из 583 химических сдвигов показала, что среднее отклонение предсказанных величин от измеренных составляет 0.06 м.д. при стандартном отклонении 0.18 м.д.

Авторы работы¹³⁵ оценили качество системы Proton Shift на большом тестовом множестве и нашли, что стандартное отклонение химических сдвигов составило 0.3 м.д., а это ближе к реальным значениям. С помощью генератора структур ASSEMBLE¹³⁶ было сгенерировано 24 308 изомеров, в которых протоны находились в 111 605 различных окружениях. Оценить химические сдвиги удалось в 89% случаев, что можно считать удовлетворительным. Среднее время предсказания химических сдвигов одной структуры на компьютере Macintosh Quadra 650 составило 0.30 с. При сравнении предсказанных и экспериментальных спектров последние вручную вводятся как совокупность химических сдвигов групп CH , CH_2 и CH_3 . Каждый экспериментальный химический сдвиг сопоставляется программой с ближайшим значением соответствующего химического сдвига в предсказанном спектре. По результатам сравнения вычисляются стандартные отклонения, значения которых используются для ранжирования сгенерированных изомеров. Из работы¹³⁵ непонятно, как производится сопоставление химических сдвигов в спектрах высоких порядков. По-видимому, описанный прием может быть использован лишь для предварительной грубой отбраковки структур в ответном файле, уже ранжированном после прогноза спектров ЯМР на ядрах ^{13}C .

Программа ACD/HNMR Predictor¹³⁷ предназначена для автоматического расчета кривой поглощения в спектре ЯМР на ядрах ^1H по структурной формуле молекулы. Программа опирается на внутреннюю базу данных, содержащую 180 000 экспериментальных химических сдвигов и 25 000 экспериментальных констант спин-спинового взаимодействия (КССВ). Алгоритм количественного учета спин-спиновых взаимодействий в органических соединениях использует параметры, найденные для более чем 3000 структурных фрагментов. Каждое имеющееся в базе значение δ_{H} отнесено к определенному структурному фрагменту, а каждая константа $J_{\text{H-H}}$ — к паре взаимодействующих структурных фрагментов с учетом связывающего их скелета. Внутренняя база отнесенных величин δ_{H} и $J_{\text{H-H}}$ была создана в результате анализа 48 000 экспериментальных спектров ЯМР на ядрах ^1H . Для фрагментов, отсутствующих в базе данных, значения δ_{H} и $J_{\text{H-H}}$ вычисляются по специальным алгоритмам.

Расчет спектра производится для базовой частоты спектрометра, заданной как параметр. Для рассчитанных химических сдвигов и констант указывается 95%-ный доверительный интервал вероятности, позволяющий оценивать надежность найденных величин. Расчет спектра ведется с учетом взаимодействий второго порядка и дальних КССВ, так что программа позволяет моделировать спектры сильно связанных спиновых систем, содержащих до восьми магнитно-неэквивалентных ядер (если имеются магнитно-эквивалентные ядра, то возможен учет и большего числа взаимодействующих спинов). Программа различает *цис-транс*-изомеры алкенов, а также *син-анти*-изомеры амидов, оксимов, гидразонов и нитрозоаминов.

После ввода структуры, для которой необходимо предсказать спектр, программа автоматически «расщепляет» молекулу на множество уникальных фрагментов, которые затем сравниваются с фрагментами, находящимися во внут-

ренной базе данных. Если какой-либо фрагмент из структуры совпадает с одним из фрагментов, принадлежащих базе данных, то его экспериментальные значения δ_H включаются в набор химических сдвигов структуры. Если некоторые фрагменты не обнаруживаются в базе данных, значения химических сдвигов экстраполируются с использованием параметров похожих фрагментов.

По завершении формирования списков химических сдвигов и КССВ программа строит и диагонализует матрицу спинного гамильтониана, в результате чего рассчитываются число линий, их положения и интенсивности, а также производится отнесение δ_H к атомам водорода данной структуры. Программа выдает спектр в том виде, в каком он выводится из прибора (включая оцифровку положений линий и интегральную кривую). Процессорное время компьютера типа Pentium, необходимое для расчета химических сдвигов, спектрального контура, интегральной кривой и оцифровки максимумов, составляет от нескольких секунд до десятков секунд.

Программа ACD/HNMR Predictor создавалась для интерпретации экспериментальных спектров и проверки гипотез. Однако когда предварительное ранжирование структур ответного файла ЭС уже выполнено на основе предсказания спектров ЯМР на ядрах ^{13}C , в дополнительной проверке нуждаются, как правило, лишь 5–10 структур, стоящих в начале списка, и, следовательно, сравнение расчетных спектров с экспериментальными с целью отсева структур-кандидатов может быть легко выполнено самим химиком.

В эффективности подобного использования этой программы автор обзора убедился на примере анализа ответных файлов для многих задач, решенных с помощью ЭС.³¹ Во всех рассмотренных случаях удавалось либо выявить истинную структуру, либо резко сузить круг альтернатив.

3. Прогнозирование ИК-спектров

Природа ИК-спектров такова, что в принципе их можно использовать для оценки структур-кандидатов. Это связано с тем, что ИК-спектр зависит не только от того, какие группы входят в молекулу, но и от того, как они расположены, т.е. ИК-спектр отражает свойства молекулы как целого. Поэтому значительные усилия многих исследователей были направлены на разработку методов прогнозирования ИК-спектров органических молекул.

Полуэмпирический расчет ИК-спектра с использованием компьютерной базы данных по молекулярным параметрам (геометрия, силовые постоянные, электрооптические параметры) можно рассматривать как эффективный метод прогноза спектров. Подходящая для этих целей база данных, а также вычислительные методы и программы были разработаны под руководством Л.А.Грибова.¹³⁸ Расчет ИК-спектра проводится в гармоническом приближении валентно-силового поля с использованием полуэмпирического подхода для вычисления интенсивностей полос.¹³⁹ Силовые и электрооптические параметры для базы данных находятся на основе квантово-химических расчетов с последующим решением механической и электрооптической обратных спектральных задач. Рассчитанные спектры представляются в форме спектрального распределения коэффициента поглощения и могут непосредственно сравниваться с экспериментальными кривыми. Процессорное время, необходимое для расчета ИК-спектра молекулы среднего размера, составляет несколько минут, но затраты времени химика на подготовку исходных данных зависят от степени сложности задачи и могут составлять от 15 мин до нескольких часов. Опыт показывает, что этот подход может быть успешно применен не только для выбора преимущественных изомерных структур, но и для выявления наиболее вероятных стереоизомеров и конформеров.¹⁴⁰

Наряду с указанными преимуществами описанный метод прогноза ИК-спектров имеет по крайней мере три недостатка: он не обладает такой быстротой, как методы предсказания спектров ЯМР и, следовательно, не может быть использован для «конвейерного» расчета десятков структур (о сотнях и тысячах не приходится говорить); предсказание спектра не может быть выполнено автоматически без участия химика; метод может быть применен для распознавания молекул малого и среднего размеров.

Два первых недостатка вполне очевидны и их влияние будет постепенно уменьшаться по мере усовершенствования алгоритмов и увеличения быстродействия компьютеров. Третий недостаток является принципиальным: он обусловлен не ограничением производительности программ размерами молекул (например, программа, рассмотренная в статье¹³⁸, позволяет работать со структурами, содержащими более 100 атомов), а природой колебаний молекул. В общем случае, чем больше размеры молекулы, тем богаче ее ИК-спектр. Следствием этого является тот факт, что полосы поглощения достаточно крупных молекул почти равномерно покрывают область «отпечатков пальцев», что делает бессмысленным сравнение рассчитанных спектров структур-кандидатов с экспериментальными спектрами. Следовательно, прямой расчет ИК-спектров целесообразно применять в случаях, когда быстрые методы предсказания не позволяют сделать однозначный выбор.

Прогноз ИК-спектров трех изомеров близкого строения, для которых стандартные отклонения предсказанных спектров ЯМР на ядрах ^{13}C оказались в пределах допустимых погрешностей метода расчета, иллюстрирует рис. 3. Структуры расположены в порядке уменьшения сходства предсказанных и экспериментального спектров ЯМР. Сравнение расчетных и экспериментального ИК-спектров показывает, что расчетный спектр структуры, представленной на рис. 3, *b*, ближе к экспериментальному, чем расчетные спектры других

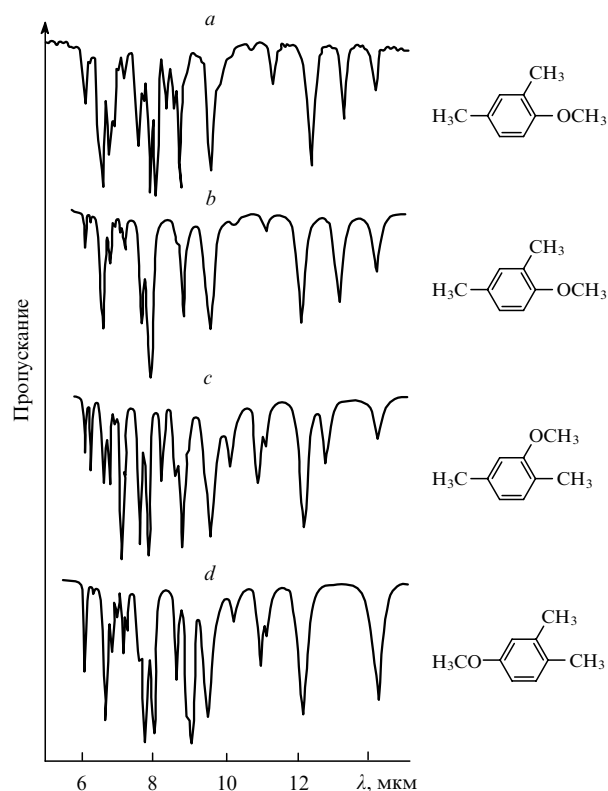


Рис. 3. Экспериментальный (*a*) и расчетные (*b–d*) ИК-спектры структур-кандидатов.¹⁴¹

структур (см. интервал 10–14 мкм). При использовании для прогнозирования спектров ЯМР на ядрах ^1H с помощью программы, предложенной в работе¹³⁷, получен такой же результат. Таким образом, три метода прогнозирования спектров указывают на приоритет первой структуры (2,4-диметиланизол).

Полуэмпирические методы для быстрого и автоматического прогнозирования ИК-спектров использовать невозможно, поэтому исследователи стали искать другие приемы. В работе¹⁴² предложен простейший способ моделирования спектральной кривой ИК-поглощения. В рамках системы DARC была создана база данных, содержащая 700 подструктур с фиксированием интервалов частоты в ИК-спектре, на которых следует ожидать поглощения. Для каждого интервала указываются ожидаемые интенсивность и форма полос. Метод прогнозирования состоит в том, что базовые фрагменты разыскиваются в проверяемой структуре, и спектр молекулы графически представляется в виде суммы полос на интервалах, соответствующих обнаруженным фрагментам. Очевидно, что при таком подходе неизбежно теряются многие полосы скелетных колебаний в области «отпечатков пальцев», а также обертоны и составные полосы, а в области функциональных групп не учитывается влияние симметрии молекулы на интенсивность полос. Очевидно также, что библиотека фрагментов далеко не исчерпывает все возможные атомные комбинации.

Следует иметь в виду, что характеристические спектральные признаки изомеров, включенных в ответный файл ЭС, как правило, соответствуют экспериментальным спектрам. Именно нехарактеристические ИК-полосы, проявляющиеся в области «отпечатков пальцев», являются носителями структурной информации, которая может быть использована для различения изомеров, имеющих близкую структуру. На практике окончательный выбор наиболее вероятной структуры делается на основании тщательного сравнения частот и интенсивностей не более пяти полос в экспериментальном и предсказанном ИК-спектрах (см. рис. 3). Поэтому подход, предложенный в работе¹⁴², следует считать абсолютно неприемлемым для анализа ответных структурных файлов.

Попытка прогнозирования кривой ИК-поглощения органических соединений по закодированным числовыми кодами структурным формулам с использованием ИНС предпринята в исследовании¹⁴³. С помощью этого подхода удалось удовлетворительно прогнозировать поглощение в интервалах, характерных для функциональных групп, но в целом он оказался непригодным для предсказания области «отпечатков пальцев». По-видимому, неудача объясняется слишком грубым методом кодирования, при котором, в частности, полностью игнорируется пространственная структура молекул.

К аналогичному выводу приводит и анализ статьи¹⁴⁴, авторы которой пытались моделировать ИК-спектры с помощью ИНС, используя для обучения сети экспериментальные спектры, а также спектры, рассчитанные полуэмпирическими методами квантовой химии.

На фоне неудач обнадеживающим выглядит новый подход, предложенный в работах^{145, 146}. Учитывая сильную зависимость ИК-спектров от пространственной структуры молекулы, авторы поставили задачу попытаться обучить ИНС на тренировочном множестве, построенном из трехмерных моделей органических молекул. Очевидно, при этом нельзя представлять модели декартовыми координатами атомов, так как каждая модель должна описываться вектором одинаковой размерности. Для описания трехмерной структуры молекул, имеющих различное число атомов, одинаковым числом дескрипторов, проведено математическое преобразование 3D-структуры. Это было сделано на базе уравнения, используемого для анализа экспериментальных данных по электронной дифракции. В результате уда-

лось найти способ кодирования любой пространственной структуры вектором, имеющим 32 или 64 координаты (3D-MoRSE-код). Используя программу CORINA¹⁰¹ для быстрого преобразования структурных формул в трехмерные модели, авторы изучили корреляцию между пространственными структурами и ИК-спектрами с помощью ИНС. Структурные формулы тренировочного множества и целевые структуры сначала преобразовывались в пространственные модели, а затем кодировались кодом 3D-MoRSE.

Оказалось, что предложенная методология, судя по примерам, обеспечивает приемлемое моделирование ИК-спектров заданных классов соединений во всем интервале поглощения, включая область «отпечатков пальцев». Для примера на рис. 4 представлены экспериментальный и предсказанный с помощью ИНС инфракрасные спектры производного хинолина (обучение сети методом ОРО производилось на тренировочном множестве, содержащем более 50 производных хинолина и изохинолина). Сходство экспериментальной и расчетной спектральных кривых, достигнутое для этой неаддитивной сложной молекулы, показывает, что описанный подход имеет шансы стать инструментом для выбора наиболее вероятных структур. Однако следует иметь в виду, что успех в данном случае достигнут при обучении сети на соединениях одного ряда.

Интересно отметить, что в рамках развитого подхода оказалось в принципе возможным решение и «обратной» задачи — по ИК-спектру получить пространственную модель молекулы. В работе¹⁴⁶ эта возможность проиллюстрирована на примере 1-фенилбутан-1-ола, а в работе¹⁴⁷ по ИК-спектру была найдена пространственная модель такой сложной молекулы, как молекула холестерина, которая оказалась достаточно близкой к структуре, рассчитанной полуэмпирическим методом квантовой химии.

Другой, на наш взгляд также перспективный, метод прогнозирования ИК-спектров, предложен в работе¹⁴⁸. Авторы поставили задачу создать метод моделирования ИК-спектров в области $2250\text{--}550\text{ см}^{-1}$ на основе поиска ближайших соседей целевой молекулы в библиотеке, содержащей структуры вместе с их ИК-спектрами. В основу механизма прогнозирования положено разумное предположение, что подобные структурные признаки порождают подобные ИК-спектры. Таким образом, предполагается, что соединения, которым присвоен высший ранг сходства, обладают большой вероятностью иметь ИК-спектры, подобные ИК-спектру целевой молекулы. Предсказанный спектр конструируется из спектров соседей (ближайших в пространстве структур), если они удовлетворяют определенным требованиям структурного сходства. Структуры и спектры рассматриваются сами по себе, т.е. корреляции подструктура – подспектр не используются. При таком подходе, наибо-

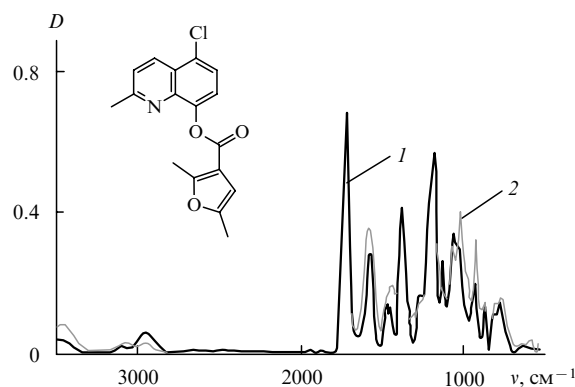


Рис. 4. Экспериментальный (1) и предсказанный с помощью ИНС (2) ИК-спектры тризамещенного хинолина.¹⁴⁶

лее вероятно сохранить нехарактеристические скелетные колебания, обертоны и комбинационные полосы, а также полосы, изменяющиеся в зависимости от симметрии молекулы, которые обычно не учитываются в корреляционных таблицах. Достоинством метода ближайших соседей является и то, что он не требует предварительного разбиения эталонных веществ на классы и можно легко обнаруживать целевые молекулы, которые не имеют аналогов в множестве эталонов, а следовательно, предупреждать о невозможности моделирования спектра.

Для реализации описанной идеи необходимо было выбрать такие способы кодирования фрагментов, структур и спектров, которые позволили бы ввести количественные меры сходства структур и подобия спектров. Авторы статьи¹⁴⁸ использовали два способа представления структур: с помощью АЦФ с одним слоем атомов,¹⁴⁹ и с применением расширенного варианта теоретико-графового дескриптора, обозначенного в работе¹⁵⁰ как ES-вектор.

Код АЦФ дает разумный компромисс между чрезмерным ростом числа фрагментов и подробным описанием молекулы. Он характеризует каждый атом его собственным типом (C, O и т.д.) и типами атомов и связей его первых ближайших соседей (атомы водорода не учитываются). Размер фрагментов изменяется от двух атомов и одной связи (концевая группа) до пяти атомов и четырех связей в случае атомов, имеющих четыре валентные связи. Фрагменты размещаются в строгом порядке, соответствующем порядку следования координат n -мерного вектора, а каждая молекула характеризуется своим вектором. Наличие или отсутствие фрагмента в молекуле сопоставляется двоичный признак (соответственно 1 или 0) и помечается число вложений данного фрагмента в структуру.

Алгоритмически было выделено 180 различных фрагментов в множестве из почти 9000 соединений с известными ИК-спектрами, а структуры кодировались вектором, имеющим $n = 180$. Отбирались молекулы, состоящие из атомов C, N, O, Br, Cl и содержащие одинарные, двойные, тройные и другие связи. В среднем структура описывалась восемью различными фрагментами, а эталонные молекулы были полифункциональными.

Код АЦФ позволяет точно описывать различные функциональные группы, однако он не способен различать варианты взаимной ориентации функциональных групп внутри молекулы (например, типы замещения бензольного кольца). Поэтому в качестве дополнительного дескриптора, способного моделировать структуру на средней и большую глубину, был использован ES-вектор.¹⁵⁰ Он не так чувствителен, как АЦФ, к изменению атомных групп, однако оба кода удачно дополняют друг друга. Код ES-вектора основан на статистике подсчета числа путей в окрашенном молекулярном графе, отображающем кодируемое соединение. Такое описание является подробным, но сильно избыточным, поэтому на следующем этапе анализировались основные компоненты²⁸ закодированных данных. Чтобы охарактеризовать каждую молекулу, использовались вклады первых восьми основных компонентов, обуславливающих 99% изменений полного кода. Опыт показал, что предложенная в работе¹⁴⁸ система кодирования структур является вполне приемлемой.

Различные способы оценки степени сходства структур рассмотрены в обзоре¹⁵¹. В частности, сходство структур X и Y , характеризуемых векторами $X = \{x_1, \dots, x_i, \dots, x_n\}$, $Y = \{y_1, \dots, y_i, \dots, y_n\}$ может быть выражено с помощью коэффициента Танимото

$$S_T = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sum_i y_i^2 - \sum_i x_i y_i}.$$

Компонентами векторов X и Y могут быть нули и единицы (наличие или отсутствие фрагмента) или целые

числа, показывающие число вложений фрагмента в структуру. Применительно к ИК-спектрам, нуль-единичное представление является более информативным, чем целочисленное, так как отсутствие какого-либо фрагмента часто полностью изменяет ИК-спектр, в то время как несоответствие количества фрагментов в структуре обычно ведет лишь к постепенному изменению спектральной картины. Поэтому отбор структур для последующей сортировки производится с помощью 0,1-векторов, а их ранжирование проводится с учетом сходства, рассчитанного на базе числа вложений. Ранжирование молекул с использованием ES-векторов основано на расчете евклидовых расстояний до вектора целевой молекулы.

Для проведения экспериментов ИК-спектры в интервале $2250 - 550 \text{ см}^{-1}$ с разрешением 2 см^{-1} были закодированы 128 точками, что соответствовало разрешению 13.5 см^{-1} с усреднением на этом интервале по семи точкам. Интенсивности спектров нормировали так, чтобы оптическая плотность самой сильной полосы была равна 1.0.

Надежная оценка сходства ИК-спектров представляет собой далеко не простую задачу. То, что видит глаз спектроскописта, трудно реализовать на языке чисел. Чтобы решить эту задачу, в работе¹⁴⁸ использовались коэффициент корреляции Пирсона и коэффициент ранговой корреляции Спирмана (самый сильный пик получает самый высокий ранг, средний пик — средний, а самый слабый — низший ранг).

Предсказание спектра выполнялось согласно следующему алгоритму.

1. Произвести поиск ближайших соседей в пространстве структур. Сравнить целевую структуру со всеми эталонами коллекции и выделить k ближайших соседей для каждой из трех моделей структурного сходства (коэффициенты Танимото для 0,1- и целочисленных векторов, евклидовы расстояния для ES-векторов). В результате получатся три списка совпадений, каждый из которых содержит по k структур.

2. Ранжировать выделенные молекулы в соответствии с тремя мерами сходства. Произвести поиск одинаковых структур в трех списках и суммировать индивидуальные ранги. Молекулы, которые попали только в один или два списка, получают ранг $k + 1$ в тех списках, где их не оказалось. Отсортировать молекулы в порядке возрастания суммарных рангов и получить конечный список.

3. Отобрать для предсказания молекулы из конечного списка, которые превосходят порог сходства по признаку наличия фрагментов или которые не превосходят порога расстояний применительно к расстоянию между SE-векторами.

4. Если в конечном списке имеется меньше двух молекул, прервать процедуру и показать химику результаты. Перейти к шагу 6.

5. Проверить ИК-спектры отобранных молекул на спектральную однородность с помощью коэффициентов корреляции. Если более 50% спектров (например, 2 из 2, 2 из 3, 3 из 4 и т.д.) проявляют большое сходство, т.е. коэффициент корреляции превышает некоторый порог, то для предсказания используется усредненный спектр. Если проверка на однородность не удастся, то передать управление химику. Перейти к шагу 6. В любом случае показать всю доступную информацию.

6. Система отказывается от автоматического предсказания, потому что нет ни одной молекулы или потому что осталась только одна молекула. Химик получает от системы сообщение о том, что нужно проанализировать ситуацию и решить, продолжать ли процедуру или прервать ее. Если будет принято решение о выполнении предсказания, то химик должен указать, спектры каких молекул из имеющих высший ранг нужно использовать.

Таким образом, процедура предсказания может быть выполнена либо в автоматическом режиме, либо при участии спектроскописта. Качество предсказаний выше, если

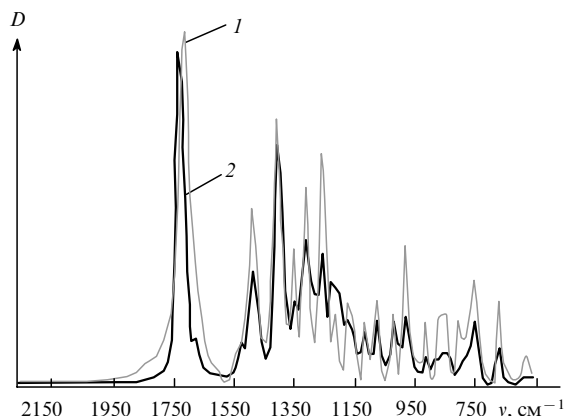


Рис. 5. Экспериментальный (1) и предсказанный на основе поиска ближайших соседей целевой молекулы (2) ИК-спектры.

окончательное решение остается за химиком. Процессорное время зависит от числа фрагментов в структуре и на компьютере Pentium 75 МГц составляет несколько секунд, что вполне приемлемо.

К сожалению, авторы ограничились демонстрацией только одного примера (рис. 5), который показывает, что сходство предсказанного и экспериментального спектров вполне приемлемо. Однако никаких попыток проверки предложенного метода для проведения сортировки реальных ответов ЭС сделано не было, что не позволяет дать ему статистически обоснованную оценку. Привлекательной стороной предложенного подхода, на наш взгляд, является то, что в алгоритме отсутствуют элементы «черного ящика», следовательно, причины изменения качества прогноза могут быть легко выявлены спектроскопистом.

V. Направления развития экспертных систем

Из экспертных систем, созданных в 1970–1980-е годы, наибольшую известность получила система DENDRAL,⁸ упоминающаяся во многих публикациях по химии и в книгах по искусственному интеллекту как «классический» образец ЭС. Популярность DENDRAL, с одной стороны, в значительной степени связана с тем, что данный проект системы разрабатывался под руководством лауреата Нобелевской премии Е.Дж.Ледерберга и известных специалистов в области искусственного интеллекта. С другой стороны, авторами проекта опубликовано порядка 40 статей под общим заголовком «Применение искусственного интеллекта в химии», что не могло не привлечь внимания многих специалистов. Однако Грей,¹⁵² критически проанализировав идеологию системы DENDRAL, пришел к заключению, что как система искусственного интеллекта она далека от совершенства. По этому вопросу состоялась острая дискуссия^{153, 154} с руководителями проекта. Вопреки расхожему мнению о широком использовании программы CONGEN, Грей утверждал, что она применяется редко (ему известно восемь работ, в которых она нашла применение). По-видимому, одна из причин не слишком большой популярности системы — отсутствие у нее такого важного качества, как дружелюбие к пользователю.

В настоящем обзоре основное внимание уделяется современным ЭС, появившимся в 1990-е годы. Сравнительный анализ возможностей современных ЭС и систем первых поколений проведен в работе¹⁵⁵. В последнее десятилетие усилия исследователей были направлены на совершенствование методов извлечения из спектров структурной информации, поиск оптимальных стратегий компьютерного распознавания структур, повышение уровня дружелюбия систем. Практика показала, что ЭС, не обладающие высо-

кой степенью дружелюбия, не имеют шансов стать рабочим инструментом химика-органика. Дружелюбная система должна:

- 1) иметь удобные и простые средства обмена информацией с пользователем, что позволяет химику контролировать любую стадию решения задачи;
- 2) предоставлять химику возможность получить объяснение каждого решения, принятого программой;
- 3) при возникновении «аварийной» ситуации на каком-либо шаге решения задачи выдавать химику совет, адекватный ситуации;
- 4) иметь высокоразвитые графические средства представления спектральной и структурной информации, обеспечивающие быстрое восприятие химиком текущей ситуации и принятие соответствующего решения.

Рассмотрим некоторые ЭС, которые, на наш взгляд, отражают современные тенденции и новые подходы. Но прежде определим основные понятия, которыми будем оперировать. Пусть эмпирическая формула неизвестного соединения определена и число изомеров, отвечающих эмпирической формуле, равно N . В принципе, значение N может быть найдено в результате генерации структур с использованием подходящей программы. Назовем решением задачи множество из n структур ($1 \leq n \leq N$), отобранных в результате работы ЭС. Решение будет правильным, если оно содержит верную структуру, в противном случае оно будет неправильным. Правильное решение и верная структура могут быть найдены только путем проверки каждой структуры, входящей в ответный файл, с использованием предсказания спектров и дополнительных экспериментов. Если ответный файл не содержит ни одной структуры ($n = 0$), то это означает, что задача не имеет решения, которое согласуется с экспериментальными данными, базой знаний и системными опциями.

В литературе описана экспертная система PACTP (см., например,²²), предназначенная для установления структуры молекул по ИК- и ЯМР-спектрам. Эта система способна не только выявлять вероятные структуры, но и генерировать и изображать все стереоизомеры, отвечающие данной структурной формуле, а также строить пространственные модели стереоизомеров. Система PACTP использовалась в качестве прототипа для построения на ее основе дружелюбной системы X-PERT,²⁹ работающей в среде WINDOWS. Рассмотрим наиболее важные элементы этой ЭС, которые в том или ином виде присутствуют и в ряде других систем.

1. Система X-PERT

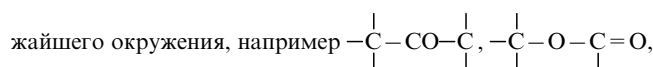
Система X-PERT предназначена для установления структуры органических молекул, содержащих до 30–35 скелетных атомов (C, N, O, P, S, B, Si, Hal) по ИК-спектрам и спектрам ЯМР на ядрах ^1H и ^{13}C . База знаний системы представляет собой набор библиотек молекулярных фрагментов с интервалами изменений их признаков в спектрах. Библиотеки формируются с учетом отнесения фрагментов к разным классам химических соединений и их специфичности для разных видов спектроскопии. Библиотеки разных категорий играют разную роль. Библиотеки, относящиеся к первой категории, являются универсальными и используются как для отбора фрагментов, так и для фильтрации структур. Библиотеки, привязанные к типам спектров, используются главным образом в качестве фильтров. Каждому фрагменту сопоставляются его признаки в ИК- и ЯМР-спектрах.

Универсальные библиотеки организованы по иерархическому принципу и состоят из четырех библиотек.

1. Библиотека основных функциональных групп (ОФГ), например $\text{C}=\text{O}$, $\text{C}\equiv\text{N}$ и т.д. Для этих групп указаны максимально возможные характеристические спектральные интервалы независимо от ближайшего окружения фрагмента. Они

используются преимущественно для предварительной оценки качественного состава образца и анализа неаддитивных соединений.

2. Библиотека структурно-группового анализа содержит функциональные группы с описанием допустимого ближайшего окружения, например



—CH₂—OH и т.д. Библиотека используется для СГА и фильтрации структур.

3. Библиотека ароматических фрагментов (АРОМ) содержит 12 типов замещения бензольного кольца.

4. Библиотека фосфорсодержащих фрагментов.

К специализированным относятся следующие библиотеки:

Библиотека данных ЯМР на ядрах ¹³C; три библиотеки данных ЯМР на ядрах ¹H, ориентированные на обнаружение групп CH₃, CH₂ и CH соответственно с учетом ближайшего окружения; библиотека, содержащая фрагменты (главным образом циклические), имеющие сильные специфические полосы поглощения (например, четырех- и пятичленные циклические кетоны, лактоны, лактамы и т.д.); библиотека фрагментов, содержащих группу OH.

Каждому фрагменту сопоставляется его «визитная карточка», в которой содержится следующая информация: структура фрагмента; интервалы изменения признаков в ИК- (частота, полуширина и интенсивность полос поглощения) и ЯМР-спектрах (химические сдвиги, КССВ, мультиплетности, числа атомов С или Н); структурные дескрипторы, задающие требования к окружению фрагмента (условия аддитивности) и обеспечивающие сохранение его спектральных признаков при переносе из одной молекулы в другую. Дескрипторы относятся к атомам фрагмента, имеющим свободные связи. Дескриптор показывает возможность присоединения к данному атому атома водорода, гетероатомов (они могут быть перечислены), ароматических циклов, кратных связей. Отмечается также возможность (или невозможность) образования кратной связи и вхождения фрагмента в циклы определенного размера.

Дескриптор может принимать три значения: «разрешено» (Р), «запрещено» (F) и «обязательно» (О). Программа обнаруживает в структуре фрагмента классы топологически эквивалентных атомов со свободными связями и помечает их своими цветами. При этом каждому окрашенному атому в таблице дескрипторов сопоставляется столбец соответствующего цвета. Это позволяет легко формировать, интерпретировать и корректировать значения дескрипторов, характеризующих эквивалентные атомы. Общее число фрагментов в библиотеках равно 450.

Экспериментальные данные — брутто-формула (молекулярная масса) и доступные спектры — могут быть введены в табличной форме вручную или импортированы программами OPUS и WIN-NMR. Система проверяет данные на непротиворечивость, задает уточняющие вопросы и выдает рекомендации по проверке надежности данных. Если брутто-формула неизвестна или вызывает сомнения, то программа генерирует все возможные молекулярные формулы по заданной молекулярной массе с учетом ограничений, накладываемых на качественный и количественный состав вещества. В случае неоднозначного ответа программа выбирает наиболее правдоподобную формулу методом, описанным в работе ¹⁵⁶.

Дальнейшая работа системы проходит в диалоговом режиме, позволяющем сочетать опыт и интуицию химика со способностью ЭВМ выполнять с большой скоростью поиск необходимой информации и производить сложные логико-

комбинаторные вычисления. При этом обеспечивается выполнение основного требования, предъявляемого к ЭС: в любой момент прохождения задачи программа способна объяснить причины принятия того или иного решения, а пользователь может получить подсказку и активизировать ту опцию, которая реализует требуемое условие.

Перед началом работы программа оценивает сложность задачи и выдает рекомендации относительно метода ее решения. Если возможно полностью автоматическое решение, то программа генерирует все структурные изомеры, а затем отбирает вероятные структуры путем пропуска их через систему фильтров.

Стандартный метод решения предполагает в качестве первой стадии генерацию наборов фрагментов. Вначале производится отбор фрагментов, которые по их брутто-формулам, спектральным интервалам и степени ненасыщенности могут присутствовать в молекуле. При этом пользователь сам выбирает наиболее подходящие для этого библиотеки фрагментов. Отбор фрагментов производится с учетом того, что надежность обнаружения фрагмента равна 100% при попадании экспериментальной полосы в среднюю часть характеристического интервала. Она линейно уменьшается при приближении к краям интервала, а за его пределами снижается до нуля. Трапециевидная функция принадлежности ¹⁹ интервала выбирается таким образом, чтобы с увеличением длины интервала уменьшение надежности на его краях происходило более плавно.

Отобранные фрагменты с индексами надежности высвечиваются на экране. Процедура «показать спектры» представляет в наглядной графической форме экспериментальные спектры вместе с интервалами отобранных фрагментов, что позволяет химику самому оценить надежность фрагментов и отредактировать список, вводя при необходимости собственные фрагменты. Далее на экран выводится таблица, содержащая интерпретацию всех имеющихся экспериментальных спектров. Химик имеет возможность просмотреть логические соотношения, связывающие фрагменты, и используя свой опыт и априорную информацию, внести изменения в «рассуждения» компьютера. Окончательно согласованные логические уравнения поступают на вход «машинного вывода». Результат решения логических уравнений (функция f(A), см. раздел II.1) выдается в виде наборов фрагментов, удовлетворяющих спектру, брутто-формуле и степени ненасыщенности анализируемого соединения. Программа автоматически добавляет в наборы микрофрагменты типа CH₃, CH₂ и т.д., которые не вошли в структуры отобранных фрагментов. Наборы могут быть также скорректированы специалистом или заново сформированы по его усмотрению. Особая процедура «повторение фрагментов» позволяет учесть повторное вхождение фрагмента в структуру, что уменьшает размерность задачи.

Следующим этапом является генерация структур из наборов фрагментов. Предусмотрена возможность наложения структурных ограничений, перечисленных в разделе III (ГУДЛИСТ, БЭДЛИСТ, ограничения на размеры циклов, кратности связей и т.д.). Кроме того, имеется опция, отменяющая или разрешающая дескрипторный контроль в процессе генерации. Учет дескрипторов помогает избежать появления большого числа неверных структур. Отмена контроля может оказаться нужной при повторном решении задачи, когда найденные структуры вызывают сомнения химика. Генерация выполняется программой, основанной на алгоритмах, разработанных в исследованиях ^{84,85}. Счетчики структур позволяют химику следить за процессом и оценивать эффективность ограничений.

В процессе генерации структур из фрагментов и свободных атомов могут образоваться такие подструктуры, присутствие которых противоречит спектрам и (или) положениям органической химии, поэтому все созданные

[†] См. также недавно вышедшую работу М.Е.Эляшберг, Yu.Z.Karasev, E.R.Martirosian. *Anal. Chim. Acta*, **388**, 353 (1999).

структуры подвергаются проверке с использованием следующих шести фильтров (см. раздел IV).

1. Проверка на соответствие характеристическим признакам в спектрах.
2. Грубое моделирование спектров ЯМР ^{13}C (см. раздел IV.1.a).
3. Проверка на соответствие правилам структурной химии и стереохимии (маловероятные фрагменты, правило Бредта).
4. Проверка каждого изомера на соответствие экспериментальным спектрам.
5. Поиск и удаление идентичных структур.
6. Предсказание химических сдвигов в ^{13}C ЯМР-спектрах с целью выявления наиболее вероятных структур.

Необходимые для фильтрации библиотеки выбираются системой на основании обобщенного портрета ответного файла. Проверка происходит в два этапа. На первом этапе формируется список фрагментов, противоречащих спектрам. На втором этапе «просеиваются» через структурные формулы только противоречивые фрагменты (это дает экономию времени) и происходит отсев тех структур, в которых обнаружен хотя бы один из этих фрагментов. Опции фильтрации предоставляют пользователю возможность указывать параметры, определяющие жесткость проверки (например, степень размытости границ интервалов).

Проверка с помощью двух первых фильтров может быть выполнена в стандартном и контролируемом химиком режимах. Первый режим обеспечивает автоматическое прохождение процесса фильтрации (так называемая слепая фильтрация). В этом случае выдаются сообщения только о выполняемой в данный момент работе и текущем числе «хороших» и «плохих» структур. Второй режим позволяет визуализировать процесс фильтрации и предоставляет пользователю возможность управлять им, а также получать всю информацию, объясняющую принятые системой решения. Каждая проверяемая структурная формула выдается на дисплей. Если в структуре не найдено ни одного из противоречивых фрагментов (напомним, что просеиваются только они), то структура помечается как правильная. Если же обнаруживаются противоречивые фрагменты, то в структурной формуле они выделяются цветом, а также выдается информация о признаках (частотах, химических сдвигах, параметрах полос или сигналов), не подтвержденных спектрами. Для этого на экран выводится «визитная карточка» фрагмента с противоречащими признаками, выделенными цветом.

В зависимости от числа и вида противоречивых фрагментов, найденных в структуре, пользователь принимает решение: удалить структуру как маловероятную или оставить ее для дальнейшей проверки. Описанный режим работы обеспечивает высокий уровень дружелюбия программы, позволяет визуализировать процесс фильтрации и делает его полностью контролируемым и понятным.

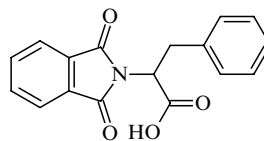
При проверке числа сигналов различной мультиплетности в спектрах ЯМР на ядрах ^{13}C в контролируемом режиме выдаются данные о различии чисел предсказанных и наблюдаемых сигналов.

После прохождения первых пяти фильтров ответный файл автоматически поступает на вход программы SpekEdit, которая предсказывает химические сдвиги в спектрах ЯМР на ядрах ^{13}C . Далее вычисляются стандартные отклонения (S) расчетных спектров от экспериментального, и структуры ранжируются по возрастанию S .

Оценка возможностей системы будет даже ниже, здесь же рассмотрим лишь один пример, иллюстрирующий работу программы.²⁹

Неизвестное вещество имеет молекулярную массу 295 и возможный состав $\text{C}_{1-50}\text{H}_{0-120}\text{N}_{0-30}\text{O}_{0-30}$; таким параметрам соответствуют 197 различных молекулярных формул. Используя данные ИК- и ЯМР-спектров, программа нашла единственную молекулярную формулу $\text{C}_{17}\text{H}_{13}\text{N}_1\text{O}_4$, после

чего были сформированы наборы фрагментов: 1) $\text{C}=\text{O}$, C_6H_5 , 1,2- C_6H_4 ; 2) $\text{C}=\text{O}$, C_6H_5 , 1,3- C_6H_4 . Из этих наборов сгенерировано 6995 структур. С использованием спектральной и структурной фильтрации выявлена единственная и правильная структура:



2. Система SpecSolv

SpecSolv[‡] является новым модулем широко известной системы SpecInfo,¹⁵⁷ предназначенной для хранения и структурной интерпретации ЯМР-, ИК- и масс-спектров. В частности, SpecInfo содержит более 200 000 отнесенных спектров ЯМР на ядрах ^{13}C , которые и послужили базой для разработки программы.

Система SpecSolv построена так, что для установления структуры молекулы используется только спектр ЯМР на ядрах ^{13}C , причем знание молекулярной массы и молекулярной формулы необязательно. Исходными служат данные спектров ЯМР на ядрах ^{13}C , в которых указаны химические сдвиги, мультиплетности и интенсивности сигналов (последние в виде числа атомов углерода). База знаний системы представляет собой библиотеку фрагментов с их подспектрами. Библиотека создана методом автоматического «нарезания» фрагментов из структур, хранящихся в SpecInfo, и включает более 400 000 фрагментов, кодированных HOSE-кодом⁷³ на глубину в три сферы, а также свыше 100 000 фрагментов, кодированных на глубину в две сферы от центрального атома углерода. Библиотека тщательно выверена специалистами.

Информация о фрагменте представлена в форме матрицы связности, HOSE-кодов всех скелетных атомов и подспектра. Для каждого атома углерода хранятся параметры его спектра ЯМР на ядрах ^{13}C (химический сдвиг, мультиплетность, интенсивность и среднеквадратичное отклонение R химического сдвига). Значения R вычислены для наборов эквивалентных подструктур в базе SpecInfo. Все фрагменты полностью определены, включая указания на замыкание колец во внешней сфере.

Первым этапом работы программы является поиск фрагментов. Все подспектры библиотеки фрагментов сравниваются с экспериментальным спектром и отбираются фрагменты, подспектры которых подтверждаются спектром. Для молекулы, содержащей ~25 тяжелых атомов, типичным является отбор ~500 фрагментов.

Процедура поиска фрагментов делится на три этапа: предварительный, основной и уточняющий поиск. Результатом предварительного поиска является кодирование подспектров интервалами, на которых присутствуют сигналы определенной мультиплетности. На втором шаге проверяется, содержит ли экспериментальный спектр соответствующее число сигналов на участках, выделенных в результате предварительного поиска. На последнем шаге проверяются точные значения отклонений химических сдвигов, представленных в обоих спектрах. Типичное отклонение составляет 2–5 м.д. Фрагменты, прошедшие проверку, затем ранжируются в соответствии с фактором совпадения, который определяется как среднее отклонение химического сдвига каждого атома от сигналов экспериментального спектра.

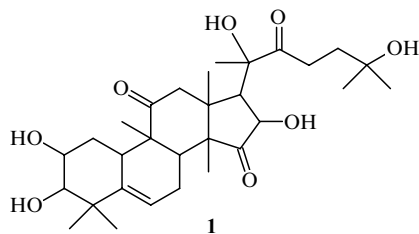
[‡] Первая работа, посвященная описанию системы SpecSolv,³⁰ появилась в 1996 г. под интригующим заголовком «Полностью автоматическое установление структуры молекул — мечта спектроскопистов становится явью».

Несмотря на то, что программа способна работать без знания молекулярной формулы, задание возможных типов атомов необходимо, причем особенно важно указать возможные гетероатомы, ибо методом ЯМР на ядрах ^{13}C непосредственно их нельзя обнаружить.

Существенно, что химику можно не просматривать все отобранные фрагменты, так как отсев неправильных фрагментов должен происходить в процессе генерации структур. Однако если химик располагает априорной информацией (а чаще всего это имеет место), то можно ожидать, что просмотр списка позволит исключить из него заведомо неправильные фрагменты, и это ускорит решение. Главная особенность алгоритма генерации в системе SpecSolv заключается в том, что фрагменты соединяются друг с другом путем наложения общих атомов. В качестве первого берется самый крупный фрагмент, подспектр которого лучше всего вписывается в экспериментальный спектр. Программа последовательно пытается присоединить к нему какой-либо другой фрагмент. В случае успеха образуется промежуточный укрупненный фрагмент. Прежде чем будет сделана попытка присоединить очередной фрагмент, производится прогнозирование спектра ЯМР на ядрах ^{13}C укрупненного фрагмента. Это достигается простым комбинированием подспектров, что обеспечивает высокую скорость прогноза. Промежуточные структуры, для которых отклонения химических сдвигов превышают заданное значение, отбрасываются. Таким образом, происходит сборка окончательной структуры с одновременной перекрестной оценкой правильности промежуточных структур. Отсутствие такого проверочного механизма в генераторах, допускающих сборку структур из перекрывающихся фрагментов (например, GENOA⁸⁷), является их существенным недостатком.

Процесс сборки продолжается до тех пор, пока будут отнесены все химические сдвиги экспериментального спектра, а в результирующей структуре будут отсутствовать свободные связи. Для получения ответной структуры обычно достаточно провести циклы генерации на базе первых десяти наиболее перспективных фрагментов (самых крупных с минимальным отклонением химических сдвигов). Это возможно, так как отобранные фрагменты ранжированы в соответствии с фактором близости их подспектров к экспериментальному спектру. В результате полного перебора во всем пространстве фрагментов чаще всего находится одна структура, хотя может быть получено несколько изомеров сходного строения (вообще говоря, если брутто-формула не задана, то возможна генерация и неизомерных структур).

В работе³⁰ приведены примеры распознавания довольно крупных молекул, например, таких как **1**.



Однако очевидно, что генерация структур возможна, только если библиотека фрагментов содержит большие перекрывающиеся «куски» таких молекул. Следовательно, для успешной идентификации в базе данных системы должно находиться достаточно много крупных молекул, родственных распознаваемой. Однако даже если подходящие фрагменты в базе найдутся, но они не будут иметь общих атомов, то генерация правильной структуры окажется принципиально невозможной, т.е. результат существенно зависит от разнообразия и качества базы данных. Программа также очень «чувствительна» к растворителю, в котором снят экспериментальный спектр, а кроме того (и это существен-

ный недостаток), она не учитывает стереоизомерные формы. Эти факторы могут оказать решающее влияние на результаты сравнения предсказанных подспектров промежуточных подструктур и полного модельного спектра с экспериментальным. По данным авторов, система опознала 80% соединений, поступивших в лабораторию для анализа. Система адаптирована для работы на ЭВМ VAX 6610.

Как видим, на самом деле система SpecSolv пока еще не реализует «мечту спектроскопистов» — не решается каждая пятая задача. При невозможности решить задачу система выдает сообщение об этом и останавливается. Поэтому необходимы новые подходы, позволяющие продолжить анализ и в тех случаях, когда программа заходит в тупик. Способ преодоления тупиковой ситуации удалось реализовать в системе ACD/Structure Elucidator.³¹

3. Система ACD/Structure Elucidator

В системе ACD/Structure Elucidator стратегия распознавания, развитая авторами SpecSolv,³⁰ объединена со ставшей уже традиционной методологией, применяемой в системах РАСТР,²² X-PERT,²⁹ CHEMICS³⁷ и др.

В системе предусмотрено два режима выявления структур: стандартный и режим классической ЭС. Первый режим во многом использует общую стратегию, развитую в работе³⁰. В качестве основных экспериментальных данных, как и в системе SpecSolv, используется спектр ЯМР на ядрах ^{13}C , в котором указываются химические сдвиги и интенсивности сигналов, причем если точное число атомов углерода, отнесенных к сигналу, определить затруднительно, указываются его нижний и верхний пределы. Для облегчения измерения интенсивности сигналов целесообразна регистрация спектра ЯМР на ядрах ^{13}C с добавлением парамагнитного реагента, например $\text{Cr}(\text{acac})_2$, который уменьшает и выравнивает время спин-решеточной релаксации T_1 для всех атомов углерода, включая четвертичные. Спектр может быть введен непосредственно с прибора или вручную в виде таблицы спектральных параметров. Наличия одного только спектра ЯМР может оказаться достаточно для выявления структуры, однако если в качестве дополнительных данных введены ИК-спектры и спектры ЯМР на ядрах ^1H , то решение задачи облегчается. Базу данных системы составляет библиотека, включающая более 400 000 фрагментов с их подспектрами ЯМР на ядрах ^{13}C , которые сгенерированы из 135 000 структур, охарактеризованных спектрами ЯМР на ядрах ^{13}C (база структур). «Объем знаний» у этой ЭС заметно меньше, чем у системы SpecSolv.

Работа программы начинается с поиска спектра ЯМР на ядрах ^{13}C в базе. При обнаружении спектра выдается соответствующая структура и программа прекращает работу. В противном случае она приступает к восстановлению структуры из фрагментов.

На первом этапе отбираются фрагменты, имеющие подспектры, не противоречащие экспериментальному спектру. Фрагменты располагаются по уменьшению чисел скелетных атомов и среднего отклонения их подспектров от соответствующих сигналов экспериментального спектра. Поскольку спектр ЯМР на ядрах ^{13}C не дает данных, позволяющих подтвердить или опровергнуть присутствие, например, групп OH, NH, $\text{C}\equiv\text{N}$ или отличить δ_{C} группы $\text{O}=\text{C}-\text{N}$ от сигнала ароматического кольца (150–160 м.д.), то при наличии ИК-спектра производится фильтрация отобранных подструктур с помощью библиотеки функциональных групп, имеющих наиболее характерные признаки в ИК-спектрах. Аналогичная фильтрация может быть проведена с использованием фрагментов, имеющих «яркие» признаки в спектрах ЯМР на ядрах ^1H . При фильтрации такого рода число отобранных фрагментов уменьшается (иногда в два раза), что способствует продвижению правильных фрагментов в начало списка. Далее, начиная с первого фрагмента, про-

изводится проверка возможности объединения фрагментов за счет общих атомов с целью последовательной сборки структуры. Как и в системе SpecSolv, объединение фрагментов сопровождается прогнозированием подспектра ЯМР на ядрах ^{13}C укрупненного фрагмента, показывающим, в правильном ли направлении идет сборка.

Решение задачи заканчивается, если в результате проверки всех возможных комбинаций фрагментов программа генерирует одну или несколько структур. Самые перспективные фрагменты (т.е. наиболее крупные, имеющие хорошо «вписывающиеся» подспектры) находятся в начале списка, поэтому часто ответная структура генерируется уже через несколько минут после начала работы программы. В случае хорошего совпадения предсказанного для нее спектра ЯМР на ядрах ^{13}C с экспериментальным процесс может быть остановлен до завершения перебора всех комбинаций фрагментов.

Если попытка собрать структуру заканчивается неудачей (в базе не оказалось подходящих фрагментов, имеющих общие атомы, или же правильные фрагменты вообще не отобраны), то программа автоматически переключается в режим классической ЭС, работающей с дискретными единицами структуры.

Как отмечалось в разделе III, для работы такой системы необходимо знание молекулярной формулы или хотя бы молекулярной массы. Если молекулярная формула неизвестна, то программа сначала отбирает функциональные группы, которые могли бы присутствовать в анализируемой молекуле. Для этого используется библиотека, содержащая небольшое число наиболее характерных функциональных групп с их спектральными признаками. Отобранные функциональные группы разыскиваются в структуре ранее выделенных фрагментов. Элементный состав и степени ненасыщенности функциональных групп, подтвержденных этой проверкой, а также числа атомов углерода и водорода, подсчитанные по данным спектров ЯМР, используются программой для автоматического определения ограничений для генератора брутто-формул. Одна или несколько вероятных брутто-формул, установленных этим генератором, передаются далее программе генерации структурных формул, которая поочередно проверяет их.

Эта программа представляет собой генератор структур из атомов и фрагментов, не имеющих общих атомов, и основана на алгоритмах, предложенных в работах^{84–86}. Формирование наборов фрагментов производится следующим образом.

Сначала выделяются N первых (базисных) фрагментов в первичном списке отобранных и отфильтрованных фрагментов (обычно $N = 10–30$ по усмотрению химика). Очевидно, что это самые крупные фрагменты, и среди них почти всегда имеются такие, которые включают часть структурной формулы анализируемого вещества. Затем используется то обстоятельство, что согласно алгоритму ранжирования в конце первичного списка оказываются малые фрагменты, чаще всего представляющие собой функциональные группы с ближайшим окружением.

Подспектр первого базисного фрагмента сравнивается с экспериментальным спектром, и в последнем выделяются химические сдвиги, для которых не нашлось соответствующего сигнала в подспектре. Затем отбираются малые фрагменты, способные интерпретировать эти сдвиги, и к первому базисному фрагменту добавляются такие комбинации малых фрагментов, которые вместе с ним обеспечивают максимальную интерпретацию экспериментального спектра. Если после этого еще остаются не интерпретированными отдельные сигналы в спектре ЯМР на ядрах ^{13}C , то программа вводит в набор соответствующие им микрофрагменты (CH_3 , CH_2 , CH , $\text{C}=\text{O}$ и т.д.). Каждый из полученных наборов проверяется на соответствие текущей брутто-формуле молекулы. Аналогичным образом формируются наборы, отве-

чающие второму, третьему и т.д. базисному фрагменту. Если не нашлось малых фрагментов, подходящих для формирования наборов, то наборы формируются с помощью микрофрагментов, и число наборов оказывается равным N .

Химику предоставлена возможность вводить в наборы свои собственные фрагменты, формировать наборы на основе одного из базисных фрагментов и использовать все виды ограничений, описанных в разделе III. Чтобы химику было легче включиться в процесс формирования наборов, программа выдает «обобщенный портрет» исследуемой молекулы, т.е. распределение типичных функциональных групп по частоте их встречаемости в отобранных фрагментах. Большая частота встречаемости говорит о том, что данная группа, возможно, действительно присутствует в молекуле, а отсутствующие группы используются для формирования БЭДЛИСТА.

Перед началом генерации структур из текущего набора программа оценивает время генерации (для этого измеряется процент выполненной работы за указанное в опциях время, например 20 с), и если оно оказывается больше приемлемого, то при первом проходе набор опускается, и программа переходит к тестированию следующего набора. Все генерируемые структуры проверяются на соответствие имеющимся спектрам и правилам органической химии и стереохимии. Специальная программа разыскивает таутомерные структуры и удаляет из списка нестабильные формы, что нередко сильно сокращает ответный файл.

Затем структуры автоматически поступают на вход программы предсказания спектров ЯМР на ядрах ^{13}C , для чего используется программа ACD/CNMR Predictor.¹¹⁸ По завершении предсказаний структуры ранжируются по возрастанию средних отклонений полученных спектров от экспериментального. Опыт показывает, что на ПК Pentium 133 МГц расчет спектра занимает от 0.5 до 1 с, так что для прогнозирования спектров 1000 структур требуется всего 5–15 мин в зависимости от числа атомов углерода. Тестирование системы показало, что в абсолютном большинстве случаев правильная структура оказывается на первом или втором месте или по крайней мере в первой десятке. Однако если отрыв предпочтительной структуры от последующих находится в пределах точности расчета спектра ЯМР на ядрах ^{13}C , то требуется дальнейшая проверка правильности решения.

Для этого осуществляется автоматический прогноз спектра ЯМР на ядрах ^1H с помощью программы ACD/HNMR Predictor (см. раздел IV.2).¹³⁷ Как правило, прогнозируются спектры только первой десятки структур, так что затраты процессорного времени сравнительно невелики ($\sim 10–60$ с на структуру в зависимости от сложности рассчитываемых спиновых систем). Экспериментальный и рассчитанные спектры ЯМР на ядрах ^1H сравниваются химиком, который делает окончательный выбор преимущественной структуры или принимает решение о необходимости дополнительных экспериментов. Рассчитанные спектры вместе с интегральной кривой и с индикацией δ_{H} всех линий выводятся на экран. Если для нескольких структур предсказанные спектры оказываются одинаково близки к экспериментальному, то сделать правильный выбор помогает определение величины ΔCH_3 для синглетов δ_{CH_3}

$$\Delta\text{CH}_3 = |\delta_{\text{CH}_3}^{\text{эксп}} - \delta_{\text{CH}_3}^{\text{расчет}}|$$

и сравнение интегралов всех мультиплетов. При $\Delta\text{CH}_3 \geq 0.2–0.3$ м.д. структура в большинстве случаев может быть отклонена. Для разрешения сомнительных ситуаций необходимы дополнительные эксперименты, в частности использование 2D спектроскопии ЯМР.

Таким образом, система ACD/Structure Elucidator отличается от других систем тем, что в ней могут быть использованы две взаимодействующие стратегии поиска решения, каждая из которых опирается на соответствующий генера-

тор структур. В случае если, в обоих режимах программа не сможет автоматически найти удовлетворительного решения, предусмотрена возможность пошагового решения задачи под контролем химика, которому предоставлены широкие возможности для вмешательства в «рассуждения» программы. Работа системы на тестовом множестве из 220 задач показала, что в 90% случаев она позволяет найти правильное решение.

4. 2D Спектроскопия ЯМР в экспертных системах

Созданные системы постоянно развиваются. Помимо усовершенствования алгоритмов и расширения баз знаний предпринимаются попытки использовать в ЭС данные 2D спектроскопии ЯМР, методы которой получили широкое распространение и, в сущности, стали уже рутинными.

Так, для расширения возможностей системы CHEMICS в нее введены средства использования данных двумерной спектроскопии ЯМР.^{158, 159} Разработаны методы обработки данных, получаемых в результате спектроскопических экспериментов 2D-INADEQUATE (C—C-корреляции через одну связь) и ^1H — ^{13}C NMR COSY (C—H-корреляции через одну связь)/ ^1H — ^1H NMR COSY (H—H-корреляции через три связи). Из данных экспериментов 2D спектроскопии обнаруживаются группы CH_n ($n = 0-3$), между которыми имеется химическая связь, и эта дополнительная информация соотносится с отобранными системой библиотечными фрагментами, при этом проверяется возможность образования связи между фрагментами. Полученные данные используются для управления процессом генерации структур из фрагментов.

В результате усовершенствования системы CASE¹⁶⁰ создана новая система SESAMI.¹⁶¹ В качестве базы данных в этой системе используется библиотека спектров ЯМР на ядрах ^{13}C , содержащая 5100 АЦФ с одним слоем окрестности. На первом этапе SESAMI формирует исчерпывающий список фрагментов, который затем «прореживается» специальной подпрограммой. Другая подпрограмма формирует дополнительные ограничения, налагаемые на структуры. Сокращенный список фрагментов и ограничения поступают на вход программы генерации структур СОСОА.⁷⁹ Предварительный отбор сгенерированных структур происходит на основе предсказания их спектров ЯМР на ядрах ^{13}C , а окончательное решение остается за специалистом.

Увеличение возможностей системы SESAMI достигается за счет включения программ¹⁶² обработки результатов экспериментов 2D ЯМР. Система принимает информацию о связности сигналов и строит фрагменты, совместимые с этими данными. Программа СОСОА осуществляет генерацию структур. Дальнейшее усовершенствование системы ведется в направлении создания блока, основанного на применении ИНС и ориентированного на выявление фрагментов при комбинированном использовании ИК-спектров и спектров ЯМР на ядрах ^{13}C .

Наряду с усилением известных ЭС средствами 2D ЯМР, в 1990-е годы начались исследования по созданию компьютерных систем, предназначенных для глубокого извлечения структурной информации из спектров 2D ЯМР и ее эффективного использования с целью генерации молекулярных структур. Потребность в разработке таких систем обусловлена необходимостью устанавливать структуры очень крупных молекул природных соединений (порядка 50–60 скелетных атомов). Представляет интерес подход, развитый в работах^{32, 163, 164}. Авторы исходят из вполне логичного тезиса о том, что коль скоро затрачен огромный труд на выделение природного продукта, то проведение серии также достаточно трудоемких экспериментов 2D ЯМР представляется вполне оправданным. Для выявления структуры используются главным образом те фрагменты, которые удастся собрать непосредственно в результате проведения 2D-экспериментов. В результате была создана экспертная

система CISOC-SES, для работы которой необходимы данные экспериментов 1D и 2D ЯМР: 1D ^1H , 1D ^{13}C , DEPT, DQF-COSY, HMQC, HMBC, HETCOR, COLOC, 2D-INADEQUATE. При таком подходе используются главным образом структурные данные экспериментов 2D ЯМР (структура восстанавливается как бы «из первых принципов»), поэтому авторы называют его *ab initio*.

Для получения информации от экспериментов 2D ЯМР требуется квалифицированный выбор методик (последовательностей импульсов) и корректное использование данных. Поэтому эффективность таких систем будет в значительной степени определяться тем, насколько успешно удастся в них реализовать «симбиоз» химика и компьютера. Поскольку системы, использующие отбор фрагментов, извлекают при этом весьма ценную структурную информацию, то следует ожидать, что подход *ab initio* в конечном счете объединится с фрагментарным подходом.

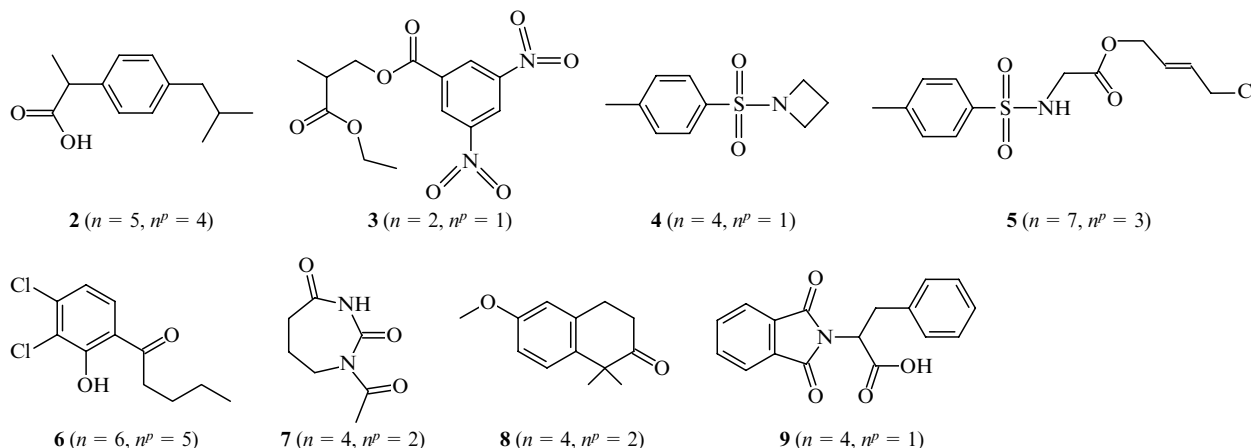
VI. Стратегия установления структуры молекул

Архитектура ЭС в целом уже определена, поэтому весьма актуальна разработка оптимальной стратегии применения ЭС. Эта проблема вряд ли может быть решена теоретически, поскольку нельзя представить все разнообразие ситуаций, которые могут возникнуть в процессе выявления структуры. Эффективные приемы могут быть установлены только в результате анализа протоколов решений, зарегистрированных для достаточно большого набора задач.

В работе¹⁴¹ была предпринята попытка построения методологии установления структуры молекул при использовании ЭС. Особое внимание уделено изучению затруднений и противоречий, возникающих при решении задач, и разработке способов их преодоления. Авторы провели достаточно широкомасштабный вычислительный эксперимент, используя систему X-PERT.²⁹ В ходе эксперимента было решено ~ 200 структурных задач, представленных главным образом в руководствах^{165–168} или взятых из спектральных коллекций и аналитической практики. Анализируемые молекулы содержали от 10 до 23 скелетных атомов, что характерно для веществ, обычно используемых в промышленных лабораториях. Подавляющее большинство эмпирических формул включали гетероатомы (от 1 до 10). Более 80% от общего числа веществ относилось к классу циклических, ~ 45% из них содержали бензольные кольца.

Как правило, экспериментальные данные включали ИК-спектры и спектры ЯМР ^1H и ^{13}C , а также эмпирическую формулу или молекулярную массу вещества. Чтобы выявить возможные пути решения, каждая задача решалась многократно с использованием различных методов и стратегий, причем особое внимание уделялось выяснению причин затруднений и неудач. Чтобы облегчить логико-статистический анализ полученных данных, протоколы решения задач автоматически регистрировались в текстовом файле, предусмотренном в активной базе данных. Наиболее вероятные структуры выявлялись на основании предсказания химических сдвигов в спектрах ЯМР на ядрах ^{13}C с использованием программы, разработанной авторами статьи¹⁶⁹. Вычислительные эксперименты проводились на персональном компьютере PC AT 486 (66 МГц).

Анализ протоколов показал, что интерактивный режим практически во всех рассмотренных случаях позволяет найти правильное решение. Для 40% задач решения оказались однозначными, а в 70% случаев число структур-кандидатов n не превышало 5. Доля задач, для которых n превышало 10 составила 15%. В результате применения предсказания спектра ЯМР на ядрах ^{13}C к неоднозначным решениям общий процент однозначных и правильных ответов вырос с 40 до 75%, т.е. программа ранжирования поместила истинную структуру на первое место в файле.



Например, для молекул **2–9** получены ответные структуры с указанием величин n , уменьшившегося размера ответного файла после применения предсказания спектра ЯМР на ядрах ^{13}C (n^p), а также порядкового номера истинной структуры в ранжированном ответном файле n' . Значение n^p определялось как число структур, для которых среднеквадратичное отклонение s не превышает 5.5 м.д. (точность программы ¹⁶⁹).

Около 60% задач было решено автоматически с применением СГА или прямой генерацией структур (ПГС) из брутто-формулы. В случаях, когда структура была однозначно установлена с помощью ПГС и последующей фильтрации структур, полная информация о структуре молекулы была получена на основе характеристических признаков, отсутствующих в экспериментальных спектрах. Более того, в результате фильтрации были распознаны некоторые структуры, содержащие фрагменты, которых нет в базе данных.

Стратегия решения задачи, в сущности, представляет собой совокупность методов преодоления всевозможных трудных ситуаций. Авторы работы ¹⁴¹ проанализировали типичные трудные ситуации, встречающиеся при использовании ЭС.

1. Примеры трудных ситуаций

Атомы переменной валентности. Химические элементы переменной валентности способны образовывать различные функциональные группы в зависимости от их валентного состояния (например, SH , >S=O , >SO_2 , >PH , >P=O). Каждая функциональная группа порождает свое семейство изомеров. Поэтому, строго говоря, структурные задачи должны повторно решаться, пока не будут перебраны все валентные состояния атомов, присутствующих в эмпирической формуле. Задача существенно упрощается, если информация о происхождении образца или данные независимых экспериментов позволяют фиксировать валентности всех элементов, указанных в брутто-формуле.

Специфический характер ИК-спектра. Известно, что ИК-спектр чрезвычайно чувствителен к малейшим изменениям молекулярной структуры и электронной плотности. Это позволяет обнаруживать даже незначительные изменения в структуре молекулы. Однако частоты или интенсивности могут быть интерпретированы только, если известна структурная формула молекулы. Пока структура молекулы не распознана, сделать это невозможно. Поэтому простые перестановки функциональных групп в изомерах, мало влияющие на аддитивность молекулы, могут приводить к выходу характеристических частот и интенсивностей полос за пределы, указанные в библиотеках фрагментов.

Например, в некоторых случаях характерные для производных бензола полосы при 3000, 1600 и 1500 cm^{-1} оказываются неактивными или настолько слабыми, что вовсе не

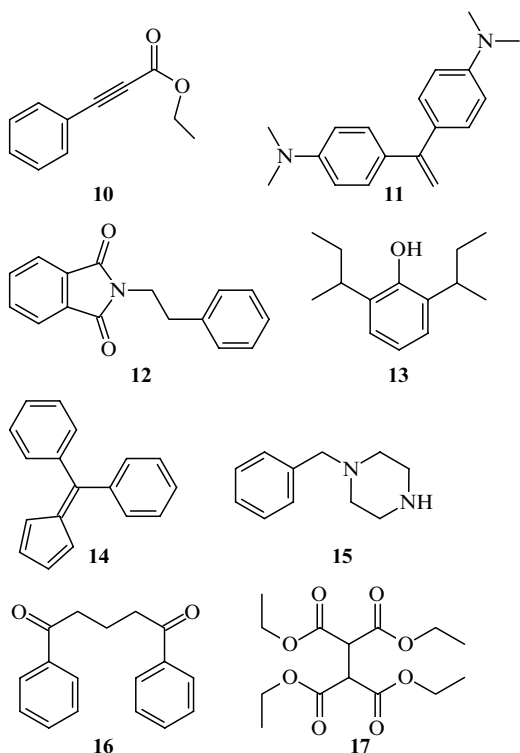
обнаруживаются в спектре. Однако если спектры ЯМР такого вещества содержат сигналы в ароматической области, то X-PERT оценивает ситуацию и предлагает химику подтвердить или опровергнуть присутствие ароматического фрагмента (да/нет). Другая проблема возникает, если вещество находится в растворителе типа CCl_4 или CHCl_3 . В этом случае диагностически важные ароматические полосы в интервале 700–800 cm^{-1} не могут быть использованы для определения типов бензольного замещения, что ведет к увеличению размеров ответного файла.

Спектральная картина ИК-поглощения валентных колебаний ОН-группы зависит от многих факторов (наличие соседних атомов и функциональных групп, свойства растворителя и т.д.). Внутримолекулярная водородная связь $\text{O} \cdots \text{H} \cdots \text{X}$ исключительно сильно влияет на параметры некоторых ИК-полос ($\nu_s(\text{OH})$, $\nu_s(\text{C}=\text{O})$ и др.), смещая их за пределы характеристических интервалов. Поэтому для решения вопроса о наличии или отсутствии ОН-группы часто необходимо вмешательство химика (X-PERT дает об этом сообщение).

Оказалось, что совместное использование ИК- и ЯМР-спектров нередко приводит к потере правильных фрагментов, характеристические признаки которых не были подтверждены ИК-спектром в силу необычной спектральной картины. Поэтому иногда следует «отключать» ИК-спектр при первом пуске программы. Но это, конечно, не означает, что ЭС может вообще обойтись без ИК-спектров, так как имеется довольно много функциональных групп, не содержащих атомов С и Н (например, NO_2 , SO , SO_2 , NO , ONO , $\text{P}=\text{O}$, $\text{P}=\text{S}$ и т.д.), которые не могут быть непосредственно обнаружены из спектров ЯМР без подтверждения по полосам в ИК-спектре.

Влияние особенностей структуры на сложность задачи. На первый взгляд может показаться, что чем больше молекула, тем большего значения n следует ожидать. В действительности n является очень сложной функцией разных структурных параметров. Поэтому вполне возможно, что крупная молекула будет идентифицирована однозначно, в то время как решение, выданное при распознавании сравнительно небольшой молекулы, будет содержать много структур. Рассмотрим некоторые наиболее значимые структурные факторы, влияющие на решение задачи.

Если структура анализируемого соединения обладает определенной симметрией, то решение задачи облегчается, так как при проверке структур на соответствие предсказанного числа сигналов и их мультиплетностей в спектре ЯМР на ядрах ^{13}C часто отбрасывается большое число неправильных структур. Например, для молекул **10–17** были найдены единственные и правильные ответные структуры. Заметим, что структура **12**, включавшая фрагмент $\text{O}=\text{C}-\text{N}-\text{C}=\text{O}$, была распознана однозначно, несмотря на то, что этот фрагмент отсутствует в базе знаний.



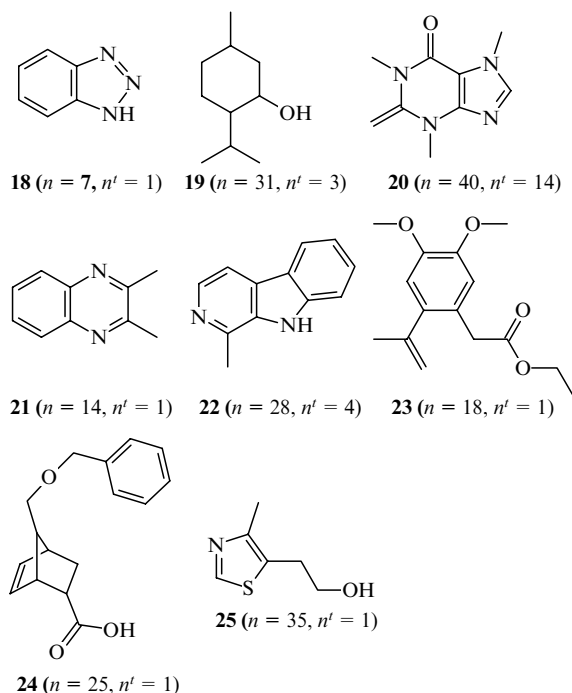
Наличие симметрии также позволяет химику использовать более эффективно процедуру «повторение фрагментов», которая формирует наборы, содержащие допустимое (с точки зрения эмпирической формулы) число идентичных фрагментов, что делает возможным генерацию крупных структур. Например, для эмпирической формулы $C_{14}H_{22}O_8$ программа создала набор, включавший четыре сложноэфирные группы и четыре этильные группы, в результате чего единственная структура **17** была выделена системой фильтров.

Как известно, число сигналов в спектре ЯМР на ядрах ^{13}C , рассчитанное как число групп топологически эквивалентных атомов углерода, не всегда соответствует экспериментальным данным. Например, в спектре ЯМР на ядрах ^{13}C молекулы **18** вместо ожидаемых шести сигналов наблюдаются только три. В этом случае правильная структура была однозначно определена с использованием прогнозирования спектра ЯМР на ядрах ^{13}C .

Случайное вырождение сигналов в спектре ЯМР или расщепление сигналов атомов углерода *gem*-диметильных, изопропильных и других групп, находящихся вблизи стереоцентров, затрудняет компьютерное распознавание структуры и требует вмешательства химика. Если молекула содержит несколько стереоцентров, то ответный файл часто содержит больше структур, чем обычно. Этот факт объясняется сложностью спектров ЯМР на ядрах 1H (мало или вовсе нет сигналов первого порядка, различные значения δ_H геминальных протонов) и дополнительными расщеплениями сигнала в спектрах ЯМР на ядрах ^{13}C . Например, решение для соединения **19**, имеющего три стереоцентра, содержит 31 структуру, причем истинная структура помещена программой предсказания δ_C на третье место.

Большое число структур-кандидатов в ответном файле характерно и для гетеро- и карбоциклических соединений, например, для **18–25**. Этот факт связан с большим числом перестановок, которые возможны при наличии в гетероциклических структурах нескольких заместителей. Примером может служить соединение **20**. Набор, включающий все фрагменты этой структуры, был построен вручную:

$3(\text{>N-CH}_3)$, $2(\text{>C=O})$, >C=C< , >C=N- . В результате



программа сгенерировала 181 структуру, а после фильтрации было выделено 40 структур, содержащих пяти- и шестичленные циклы.

Для предсказанного спектра ЯМР на ядрах ^{13}C истинной структуры **20** величина $S = 11.6$ м.д. (максимальное отклонение равно 25.6 м.д.). В результате истинная структура помещена ранжирующей программой на 14-е место. Таким образом, даже набор фрагментов, исчерпывающий правильную структуру, приводит к достаточно «размытому» решению, поэтому для выявления истинной структуры требуется привлечение дополнительной информации (например, от 2D ЯМР) и более надежных методов предсказания спектров.

Анализ способов преодоления трудных ситуаций позволил предложить методологические подходы, ведущие к правильным решениям.

2. Стратегии решения задач

Чтобы решение задачи, полученное с помощью ЭС, было правильным (т.е. истинная структура присутствовала бы в ответном файле) и содержало минимум возможных структур (идеально одну), необходимо (но не достаточно) обеспечить отбор как можно большего числа фрагментов, входящих в анализируемую структуру. Для этого у системы должна быть обширная база знаний, и характеристические спектральные интервалы фрагментов должны быть достаточно широкими, чтобы не потерять правильные фрагменты.

При выполнении этих условий оказывается, что количество отобранных фрагментов слишком большое; отбирается много лишних (неправильных) фрагментов; возникает необходимость обнаружения и исключения неправильных фрагментов. Редактирование списка отобранных фрагментов вряд ли возможно без участия химика. Эту процедуру можно исключить только при автоматическом построении всех логически допустимых наборов и генерации из них структурных формул. В результате количество сгенерированных структур будет чрезмерно большим, а генерация и фильтрация — длительными, в связи с чем задача даже может оказаться практически неразрешимой. Кроме того, ответный файл может содержать структуры, противоречащие априорной информации. Следовательно, в любом случае, проверка списка отобранных фрагментов химиком представляется не только желательной, но и практически неизбежной.

Таким образом, требование наличия максимального числа правильных фрагментов в списке отобранных подструктур ведет к увеличению общего числа фрагментов и, как следствие, к увеличению размеров ответного файла. При этом, вероятно, решение будет правильным, но «размытым». Неопределенность решения следует рассматривать как плату за правильность.

Селективность отбора фрагментов существенно зависит от дескрипторов окружения фрагментов. Чем жестче требования к окружению фрагментов, тем уже интервалы характеристических спектральных признаков, следовательно, выше селективность СГА. Но в то же время сужение интервалов повышает вероятность потери правильных фрагментов из-за того, что какой-либо экспериментальный признак выходит за пределы характеристического интервала.

Существует также четкая зависимость между размерами структурного файла и строгостью дескрипторов, задающих окружение фрагментов. Дело в том, что в процессе генерации на диске сохраняются только те структуры, которые полностью отвечают требованиям дескрипторов окружения. Это резко уменьшает число структур, требующих дальнейшей проверки. С одной стороны, существуют задачи, которые вообще не решаются, если контроль по дескрипторам не осуществляется в процессе генерации структур (время решения становится неприемлемым). Но, с другой стороны, существует опасность, что дескрипторный контроль может привести к потере правильной структуры. Например, если на стадии СГА отобран фрагмент $O=C-C=C$, определенный в библиотеке как ациклический (в то время как истинная структура содержит в цикле $C=C$), то все структуры, содержащие этот фрагмент в цикле (и истинная структура в том числе), отсеиваются в процессе генерации структур. В результате решение будет неправильным. Желательно, чтобы идеальная ЭС была селективной на стадии СГА, и в то же время выдавала однозначное и правильное решение. Практическая реализация обоих требований едва ли возможна, следовательно, должен быть найден компромисс.

Поскольку приоритетное требование к решению состоит в том, что оно должно быть правильным, то можно примириться с неоднозначным, зато правильным решением. Поэтому оптимальная стратегия выявления структуры может быть сформулирована следующим образом: решение задачи осуществляется в интерактивном режиме, который представляет собой последовательность итераций, сопровождающихся повышением жесткости накладываемых ограничений. Например, если размерность задачи не очень велика, то для отбора фрагментов сначала следует использовать библиотеку основных функциональных групп (ОФГ, см. раздел V.1). В этом случае будет обеспечена максимальная свобода окружения фрагментов в процессе генерации структур. Если необходимо распознать крупную молекулу, то дополнительно используют библиотеки СГА и АРОМ.

Для регулирования жесткости ограничений при отборе фрагментов и генерации структур в рамках ЭС химик может включать (отключать) следующие опции: использование любого типа спектров из доступных экспериментальных данных; использование мультиплетностей и КССВ в ЯМР-спектрах; использование интенсивностей и полуширин полос в ИК-спектрах; контроль по дескрипторам окружения при генерации структур; контроль по дескрипторам окружения в процессе фильтрации структур; установка различных допусков для предсказанного числа сигналов различной мультиплетности в спектре ЯМР на ядрах ^{13}C ; использование структурных ограничений в процессах генерации и фильтрации структур (пределы размеров циклов, максимальная кратность связи, выполнимость правила Бредта, ГУДЛИСТ, БЭДЛИСТ, список фрагментов, маловероятных в органической химии, и т.д.).

Если при первом пуске программы (обычно это делается в автоматическом режиме) решение не найдено, то при

разумном использовании перечисленных выше переключаемых параметров в процессе итераций (проб и ошибок), как правило, можно выяснить причины противоречий, что облегчает поиск правильного решения. Особенно эффективным методом поиска противоречий является использование режима контролируемой фильтрации (см. раздел V.1). Сохранение всех промежуточных результатов в активной базе данных дает пользователю возможность возвращаться к решению задачи начиная с любого шага.

Быстродействие и автоматизм программ прогнозирования спектров ЯМР на ядрах ^{13}C (0.5–1 с на спектр) позволяют эффективно применять изложенную выше стратегию. Так, для достижения правильного решения допустимо даже такое ослабление жесткости начальных ограничений, при которых ответный файл может содержать тысячи структур. Процессы предсказания химических сдвигов и ранжирования структур по их правдоподобию производятся полностью автоматически, поэтому необходимые на это затраты процессорного времени ПК оправданы и приемлемы.

Выбор подходящей стратегии применительно к данной задаче зависит от оптимального комбинирования двух главных режимов работы системы, которые можно назвать «Археолог» и «Скульптор». Режим «Археолог» обеспечивает выполнение СГА и генерацию структур («склеивание» фрагментов и атомов). Затем производятся фильтрация и процедура отбрасывания неверных структур, что соответствует стратегии режима «Скульптор» (отсекается все лишнее).

Анализ большого числа решений позволил разделить все задачи на следующие категории:

1) задачи, решаемые в автоматическом режиме прямой генерацией структур из молекулярной формулы с последующим применением стратегии режима «Скульптор»;

2) задачи, решаемые автоматически с применением стратегий режимов «Археолог» и «Скульптор»;

3) задачи, решаемые в полуавтоматическом режиме (химик редактирует рассуждения программы, используя различного рода информацию, легко извлекаемую из экспериментальных данных);

4) задачи, которые могут быть решены только в интерактивном режиме (ручное формирование наборов фрагментов, наложение различных ограничений, обусловленных происхождением образца и т.д.).

Попадание задачи в ту или иную категорию зависит от различных факторов (например, размера молекулы, наличия крупных фрагментов, имеющих характеристические спектральные признаки, наличия симметрии и т.д.), из которых наиболее важным является число скелетных атомов в эмпирической формуле. Во многих случаях СГА может быть выполнен автоматически, однако появление хотя бы одного противоречия между аксиомами теории СГА и экспериментальными данными делает вмешательство специалиста практически неизбежным. Следует иметь в виду, что даже совпадение предсказанных спектров с экспериментальными не дает гарантии правильного выбора структуры. Может потребоваться проведение дополнительных экспериментов, в частности рентгеноструктурного анализа или подтверждения структуры с помощью техники INADEQUATE 2D ЯМР. Хотя это и связано с дополнительными трудозатратами, но они несоизмеримы с затратами на встречный синтез, выделение и очистку продукта.

Таким образом, программы ИИ следует рассматривать как мощные средства усиления интеллекта исследователя, но не как метода, способного вытеснить его из процесса установления структуры. Эту мысль четко выразили авторы монографии¹⁷⁰: «для установления структуры сложных природных соединений разработаны автоматические компьютерные методы или их интерактивные версии, такие как SESAMI, но не существует ничего, что могло бы заменить

упорный труд, опыт и интуицию химика». С этим нельзя не согласиться.

VII. Заключение

Первые работы, опубликованные 30 лет назад, заложили основы самостоятельного научного направления, сложившегося к настоящему времени, — разработка ЭС для установления структур молекул органических соединений спектральными методами. Это направление тесно переплетается со многими разделами химии (органической, физической, аналитической, квантовой химией, стереохимией, молекулярной спектроскопией, молекулярной механикой), математики (теорией графов, комбинаторикой, математической логикой, теорией нечетких множеств, математической статистикой), кибернетики (ИИ, программированием, теорией информации, распознаванием образов, теорией ИНС, компьютерной графикой, математической лингвистикой), психологии (взаимодействие в системах человек–компьютер) и философии науки (роль принципа дополнительности при изучении строения вещества).¹⁷¹ Поэтому следует ожидать, что ЭС, вбирающие в себя новые достижения смежных наук, будут быстро совершенствоваться.

Три основных этапа процесса распознавания структуры — СГА, генерация структур и прогнозирование спектров структур-кандидатов — превратились в самостоятельные развивающиеся научные направления. В большей степени удалось автоматизировать генерацию структур и прогнозирование спектров ЯМР. Для выявления молекулярных фрагментов разработан ряд методов, однако эта стадия остается наиболее уязвимой и зависящей от априорной информации. В последние годы начались интенсивные исследования по применению ИНС для обнаружения фрагментов.

До сих пор не удалось найти способа быстрого и надежного прогнозирования ИК-спектров. Для решения этой проблемы необходимо исследовать возможности ИНС (уже имеются обнадеживающие результаты) и разработать новые подходы. Известные методы расчета ИК-спектров необходимо существенно усовершенствовать, чтобы ускорить расчеты. В работах^{172–174} показано, что расчеты ИК-спектров могут быть также использованы для безэталонного количественного спектрального анализа органических соединений. Поэтому встраивание в ЭС баз данных (например, такого типа, как в работе¹³⁸) позволит создать системы как для качественного, так и для количественного спектрального анализа органических веществ.

С учетом опыта решения многих структурных задач установлена общая стратегия выявления структур молекул с помощью ЭС: проводится ряд последовательных итераций, сопровождаемых постепенным повышением жесткости накладываемых ограничений. Применение этой стратегии позволяет получать правильные решения.

Большинство ЭС пока способны распознавать молекулы небольшого и среднего размера (до 20–25 скелетных атомов), но это не означает, что применение ЭС ограничено. Даже достаточно малой молекулярной формуле могут соответствовать миллионы органических структур (например, эмпирической формуле $C_9H_{10}O_3$ соответствует почти 48.5 млн структур), а так как в настоящее время известно «всего» ~ 13 млн. веществ, то нетрудно прийти к выводу, что поле для применения ЭС очень широко.

Так как реальные возможности наиболее совершенных генераторов структур ограничены 10–12 молекулярными фрагментами, то весьма перспективными для распознавания крупных молекул являются ЭС, способные обнаруживать большие фрагменты. Следует ожидать также быстрого развития ЭС, в которых используются данные спектроскопии 2D ЯМР.

Литература

1. М.Е.Эляшберг, Л.А.Грибов. *Журн. прикл. спектроскопии*, **8**, 296 (1968)
2. М.Е.Эляшберг, Л.А.Московкина. *Журн. прикл. спектроскопии*, **8**, 998 (1968)
3. J.Lederberg, G.L.Sutherland, B.G.Buchanan, E.A.Feigenbaum, A.V.Robertson, A.M.Duffield, C.Djerassi. *J. Am. Chem. Soc.*, **91**, 2973 (1968)
4. S.I.Sasaki, H.Abe, T.Ouki, M.Sakamoto, S.Ochiai. *Anal. Chem.*, **40**, 2220 (1968)
5. D.B.Nelson, M.E.Munk, K.B.Gasli, D.L.Horald. *J. Org. Chem.*, **34**, 3800 (1969)
6. В.М.Татевский. *Классическая теория строения молекул и квантовая механика*. Химия, Москва, 1973
7. М.Е.Эляшберг, Л.А.Грибов, В.В.Серов. *Молекулярный спектральный анализ и ЭВМ*. Наука, Москва, 1980
8. R.K.Lindsay, B.G.Buchanan, E.A.Feigenbaum, J.Lederberg. *Applications of Artificial Intelligence for Organic Chemistry. The DENDRAL Project*. McGraw-Hill, New York, 1980
9. N.A.B.Gray. *Computer-Assisted Structure Elucidation*. Wiley, New York, 1986
10. *Искусственный интеллект. Применение в химии*. (Под ред. Т.Пирса, Б.М.Хони). Москва, Мир, 1988
11. J.Klassens, G.Kateman. *Fresenius Z. Anal. Chem.*, **326**, 203 (1987)
12. N.A.B.Gray. *Anal. Chim. Acta*, **210**, 9 (1988)
13. G.W.Small. *Anal. Chem.*, **59**, 535A (1987)
14. М.Е.Эляшберг. В кн. *Математические методы и ЭВМ в аналитической химии*. Наука, Москва, 1989. С. 132
15. М.Е.Эляшберг, В.В.Серов, Л.А.Грибов. *Talanta*, **34**, 21 (1987)
16. L.A.Gribov, M.E.Elyashberg, V.V.Serov. *J. Mol. Struct.*, **50**, 371 (1978)
17. А.М.Денисов. *Введение в теорию обратных задач*. Изд-во МГУ, Москва, 1994
18. А.Н.Тихонов, В.Я.Арсенин. *Методы решения некорректных задач*. Наука, Москва, 1979
19. V.V.Serov, L.A.Gribov, M.E.Elyashberg. *J. Mol. Struct.*, **129**, 183 (1985)
20. А.Кофман. *Введение в теорию нечетких множеств*. Радио и связь, Москва, 1982
21. H.J.Luinge. *Vibrat. Spectrosc.*, **1**, 3 (1990)
22. М.Е.Эляшберг. *Журн. аналит. химии*, **47**, 966 (1992)
23. W.A.Warr. *Anal. Chem.*, **65**, 1087A (1993)
24. J.Zupan, J.Gasteiger. *Neural Networks for Chemists*. VCH, Weinheim, 1993
25. E.W.Robb, M.S.Madison, M.E.Munk. *Mikrochim. Acta.*, **II**, 505 (1991)
26. C.Affolter, J.T.Clerc. *Chemometr. Intell. Lab. Syst.*, **21**, 121 (1993)
27. D.Ricard, C.Cachet, D.Cabrol-Bass, T.P.Forrest. *J. Chem. Inf. Comput. Sci.*, **33**, 202 (1993)
28. М.А.Шараф, Д.Л.Иллмен, Б.Р.Ковальский. *Хемомеритика*. Химия, Ленинград, 1989
29. М.Е.Эляшберг, Е.Р.Мартirosian, Yu.Z.Karasev, H.Thiele, H.Somberg. *Anal. Chim. Acta*, **337**, 265 (1997)
30. M.Will, W.Fachinger, J.R.Richert. *J. Chem. Inf. Comput. Sci.*, **36**, 221 (1996)
31. *ACD/Structure Elucidator*. Advanced Chemistry Development Inc., www.acdlabs.com
32. C.Peng, S.Yuan, C.Zheng, Y.Hui. *J. Chem. Inf. Comput. Sci.*, **34**, 805 (1994)
33. L.A.Gribov, M.E.Elyashberg. *J. Mol. Struct.*, **5**, 179 (1970)
34. В.В.Серов, М.Е.Эляшберг, Л.А.Грибов. *Докл. АН СССР*, **232**, 592 (1977)
35. М.Е.Эляшберг, Л.А.Грибов, В.Н.Колдашов, И.В.Плетнев. *Докл. АН СССР*, **268**, 112 (1983)
36. М.Е.Эляшберг, В.В.Серов, Е.Р.Мартirosian, L.A.Zlatina, Yu.Z.Karasev, V.N.Koldashov, Yu.Yu.Yampolskiy. *J. Mol. Struct.*, **230**, 191 (1991)
37. K.Funatsu, Y.Susuta, S.Sasaki. *Anal. Chim. Acta*, **201**, 55 (1989)
38. G.N.Andreev, O.K.Argirov, P.N.Penchev. *Anal. Chim. Acta*, **284**, 131 (1993)
39. G.N.Andreev, O.K.Argirov. *J. Mol. Struct.*, **347**, 439 (1995)

40. В.В.Серов, М.Е.Эляшберг. *Журн. структ. химии*, **27** (2), 32 (1986)
41. L.J.Bellamy. *The Infrared Spectra of Complex Molecules*. Wiley, Chichester, 1975
42. *Нечеткие множества и теория возможностей*. (Под ред. Р.Р.Ягера). Радио и связь, Москва, 1986. С. 6
43. Ч.Чень, Р.Ли. *Математическая логика и автоматическое доказательство теорем*. Наука, Москва, 1983
44. В.В.Серов, М.Е.Эляшберг, В.М.Петров. В кн. *Математические методы и ЭВМ в аналитической химии*. Наука, Москва, 1989. С. 150
45. H.J.Luinge, G.J.Kleywegt, H.A.van't Klooster, J.H.van der Maas. *J. Chem. Inf. Comput. Sci.*, **27**, 95 (1987)
46. H.J.Luinge, J.H.van der Maas. *Anal. Chim. Acta*, **223**, 135 (1989)
47. H.J.Luinge. *Trends Anal. Chem.*, **9**, 66 (1990)
48. H.Huixiao, X.Xinquan. *J. Chem. Inf. Comput. Sci.*, **30**, 203 (1990)
49. C.Cleva, C.Cachet, D.Cabrol-Bass, T.P.Forrest. *Anal. Chim. Acta*, **348**, 255 (1997)
50. C.Klawun, C.L.Wilkins. *J. Chem. Inf. Comput. Sci.*, **36**, 69 (1996)
51. M.E.Munk, M.S.Madison, E.W.Robb. *J. Chem. Inf. Comput. Sci.*, **36**, 231 (1996)
52. M.E.Munk, M.S.Madison, E.W.Robb. *Mikrochim. Acta*, **II**, 505 (1991)
53. N.Sbirrazzuoli, C.Cachet, D.Cabrol-Bass, T.P.Forrest. *Neural Comput. Applic.*, **1**, 229 (1993)
54. J.R.M.Smith, P.Schoenmakers, A.Stehmann, F.Sijstermans, G.Kateman. *Chemometr. Intell. Lab. Syst.*, **18**, 27 (1993)
55. C.Klawun, C.L.J.Wilkins. *Chem. Inf. Comput. Sci.*, **34**, 984 (1994)
56. E.W.Robb, M.E.Munk. *Mikrochim. Acta*, **I**, 131 (1990)
57. R.J.Fessenden, L.Gyorgi. *J. Chem. Soc., Perkin Trans. 2*, 1755 (1991)
58. O.C. van Est, P.Schoenmakers, J.R.M.Smits, W.P.M.Nijssen. *Vib. Spectrosc.*, **4**, 263 (1993)
59. M.Meyer, T.Weigelt. *Anal. Chim. Acta*, **265**, 183 (1992)
60. M.Meyer, H.Hobert. *Anal. Chim. Acta*, **282**, 407 (1993)
61. U.M.Weigel, R.J.Herges. *J. Chem. Inf. Comput. Sci.*, **32**, 723 (1992)
62. B.J.Hare, J.H.Prestegard. *J. Biomol. NMR*, **4**, 35 (1994)
63. C.Klawun, C.L.Wilkins. *J. Chem. Inf. Comput. Sci.*, **36**, 249 (1996)
64. И.И.Строков, И.В.Гриценко, К.С.Лебедев. *Изв. СО АН СССР. Сер. хим. наук*, (2), 78 (1987)
65. К.С.Лебедев, Е.А.Отмахова, И.В.Гриценко. *Изв. СО АН СССР. Сер. хим. наук*, (5), 73 (1990)
66. К.С.Лебедев, О.Н.Шарапова, И.К.Коробейничева, В.А.Кохов. *Сиб. хим. журн.*, (1), 50 (1993)
67. К.С.Лебедев. *Журн. аналит. химии*, **48**, 851 (1993)
68. К.С.Лебедев, И.В.Гриценко. *Журн. структ. химии*, **34** (3), 51 (1993)
69. К.С.Лебедев, Б.Г.Дерендяев. *Химия в интересах устойчивого развития*, **3**, 269 (1995)
70. K.S.Lebedev, D.Cabrol-Bass. *J. Chem. Inf. Comput. Sci.*, **38**, 410 (1998)
71. K.Varmuza, P.N.Penchev, H.Scisibrany. *J. Chem. Inf. Comput. Sci.*, **38**, 420 (1998)
72. L.Chen, W.Robien. *J. Chem. Inf. Comput. Sci.*, **34**, 934 (1994)
73. W.Bremser. *Anal. Chim. Acta*, **103**, 355 (1978)
74. M.Carabedian, I.Dagane, J.E.Dubois. *Anal. Chem.*, **60**, 2186 (1988)
75. В.В.Разников, В.Л.Тальрозе. *Журн. структ. химии*, **11**, 357 (1970)
76. Y.Kudo, S.Sasaki. *J. Chem. Inf. Comput. Sci.*, **16**, 43 (1976)
77. В.В.Серов, М.Е.Эляшберг, Л.А.Грибов. *Докл. АН СССР*, **224**, 109 (1975)
78. C.A.Shelley, M.E.Munk. *Anal. Chim. Acta*, **133**, 507 (1981)
79. B.D.Christie, M.E.Munk. *J. Chem. Inf. Comput. Sci.*, **28**, 87 (1988)
80. A.Kerber, R.Laue, D.Moser. *Anal. Chim. Acta*, **235**, 221 (1990)
81. C.Benecke, R.Grund, A.Kerber, R.Laue, T.Wieland. *J. Chem. Inf. Comput. Sci.*, **72**, 403 (1995)
82. C.Benecke, R.Grund, R.Hohberger, A.Kerber, R.Laue, T.Wieland. *Anal. Chim. Acta*, **314**, 141 (1995)
83. T.Wieland, A.Kerber, R.Laue. *J. Chem. Inf. Comput. Sci.*, **36**, 413 (1996)
84. S.G.Molodtsov. *Commun. Math. Chem. (MATCH)*, **30**, 213 (1994)
85. S.G.Molodtsov. *Commun. Math. Chem. (MATCH)*, **30**, 203 (1994)
86. S.G.Molodtsov. *Commun. Math. Chem. (MATCH)*, **37**, 157 (1998)
87. R.E.Carhart, D.H.Smith, N.A.B.Gray, J.G.Nourse, C.Djerassi. *J. Org. Chem.*, **46**, 1709 (1981)
88. С.Г.Молодцов. Дис. канд. хим. наук. НИОХ СО РАН, Новосибирск, 1997
89. М.С.Молчанова. Дис. канд. хим. наук. ИОХ РАН, Москва, 1997
90. L.M.Masinter, N.S.Sridharan, J.Lederberg, D.H.Smith. *J. Am. Chem. Soc.*, **96**, 7702 (1974)
91. M.S.Molchanova, V.V.Shcherbukhin, N.S.Zefirov. *J. Chem. Inf. Comput. Sci.*, **36**, 888 (1996)
92. Ch.-L.Hu, L.Xu. *Anal. Chim. Acta*, **298**, 78 (1994)
93. J.-M.Nuzillard, W.Naanaa, S.Pimont. *J. Chem. Inf. Comput. Sci.*, **35**, 1068 (1995)
94. R.Barone, F.Barberis, M.Chanou. *Commun. Math. Chem. (MATCH)*, **32**, 19 (1995)
95. I.P.Bangov. *J. Chem. Inf. Comput. Sci.*, **34**, 318 (1994)
96. V.Kvasnicka, J.Pospichal. *J. Math. Chem.*, **9**, 181 (1992)
97. V.Kvasnicka, J.Pospichal. *J. Chem. Inf. Comput. Sci.*, **36**, 516 (1996)
98. Л.А.Златина, М.Е.Эляшберг. *Журн. структ. химии*, **32**(4), 92 (1991)
99. L.A.Zlatina, M.E.Elyashberg. *Commun. Math. Chem. (MATCH)*, **27**, 191 (1992)
100. E.V.Gordeeva, A.R.Katritzky, V.V.Shcherbukhin, N.S.Zefirov. *J. Chem. Inf. Comput. Sci.*, **33**, 102 (1993)
101. J.Gasteiger, C.Rudolph, J.Sadowski. *Tetrahedron. Comput. Methodol.*, **3**, 537 (1990)
102. J.Sadowski, J.Gasteiger. *Chem. Rev.*, **93**, 2567 (1993)
103. J.Sadowski, J.Gasteiger, G.Klebe. *J. Chem. Inf. Comput. Sci.*, **34**, 1000 (1994)
104. C.A.Shelley, M.E.Munk. *Anal. Chem.*, **50**, 1522 (1978)
105. J.T.Clerc, H.A.Sommerauer. *Anal. Chim. Acta*, **95**, 33 (1977)
106. L.Chen, W.Robien. *Anal. Chem.*, **65**, 12282 (1993)
107. K.L.Jensen, A.S.Barber, G.W.Small. *Anal. Chem.*, **63**, 1081 (1991)
108. D.L.Clouser, P.C.Jurs. *Anal. Chim. Acta*, **295**, 221 (1994)
109. P.C.Jurs, J.W.Ball, L.S.Anker, T.L.Friedman. *J. Chem. Inf. Comput. Sci.*, **32**, 272 (1992)
110. A.Furst, E.Pretsch. *Anal. Chim. Acta*, **229**, 17 (1990)
111. W.Bremser. *Magn. Reson. Chem.*, **23**, 271 (1985)
112. L.Chen, W.Robien. *Chemometr. Intell. Lab. Syst.*, **19**, 217 (1993)
113. C.W.Crandell, N.A.B.Gray, D.H.Smith. *J. Chem. Inf. Comput. Sci.*, **32**, 48 (1992)
114. H.N.Cheng, L.J.Kasehagen. *Anal. Chim. Acta*, **285**, 223 (1994)
115. C.W.von der Lieth, J.Seibl, I.Kohler, H.J.Oferkuch. *Magn. Reson. Chem.*, **23**, 1048 (1985)
116. A.Furst, E.Pretsch, W.Robien. *Anal. Chim. Acta*, **233**, 213 (1990)
117. R.C.Schweitzer, G.W.Small. *J. Chem. Inf. Comput. Sci.*, **37**, 249 (1997)
118. *ACD/CNMR Predictor*. Advanced Chemistry Development Inc., www.acdlabs.com
119. V.Kvasnicka. *J. Math. Chem.*, **6**, 63 (1991)
120. J.P.Doucet, A.Panaye, E.Feuilleaubois, P.Ladd. *J. Chem. Inf. Comput. Sci.*, **33**, 320 (1993)
121. A.Panaye, J.P.Doucet, B.Fan, E.Feuilleaubois, S.R.El Azzouzi. *Chemometr. Intell. Lab. Syst.*, **24**, 129 (1994)
122. V.Kvasnicka, S.Sklenak, J.Pospichal. *J. Chem. Inf. Comput. Sci.*, **32**, 742 (1992)
123. L.S.Anker, P.C.Jurs. *Anal. Chem.*, **64**, 1157 (1992)
124. Y.Miyashita, H.Yoshida, O.Yaegashi, T.Kimura, H.Nishiyama, S.Sasaki. *J. Mol. Struct. (Theochem)*, **311**, 241 (1994)
125. B.E.Mitchell, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **36**, 58 (1996)
126. D.L.Clouser, P.C.Jurs. *J. Chem. Inf. Comput. Sci.*, **36**, 168 (1996)
127. O.Ivanciuc, J.-P.Rabine, D.Cabrol-Bass, A.Panaye, J.P.Doucet. *J. Chem. Inf. Comput. Sci.*, **36**, 644 (1996)
128. O.Ivanciuc, J.-P.Rabine, D.Cabrol-Bass, A.Panaye, J.P.Doucet. *J. Chem. Inf. Comput. Sci.*, **37**, 587 (1997)
129. D.M.Grant, E.G.Paul. *J. Am. Chem. Soc.*, **86**, 2984 (1964)
130. E.Breitmaier, W.Voelter. *Carbon-13 NMR Spectroscopy*. VCH, Weinheim, 1987
131. E.Pretsch, T.Clerc, J.Seibl, W.Simon. *Tables of Spectral Data for Structure Determination of Organic Compounds*. Springer-Verlag, Berlin, 1989
132. N.A.B.Gray, J.G.Nourse, C.W.Crandell, D.H.Smith, C.Djerassi. *Org. Magn. Res.*, **15**, 375 (1981)
133. K.R.Beebe, B.R.Kowalski. *Anal. Chem.*, **59**, 1007A (1987)

134. S.R.Burgin, E.Pretch. *Anal. Chim. Acta*, **296**, 295 (1994)
135. S.R.Burgin, M.E.Munk, E.Pretch. *J. Chem. Inf. Comput. Sci.*, **36**, 239 (1996)
136. C.A.Shelley, M.E.Munk. *Anal. Chim. Acta*, **296**, 295 (1994)
137. *ACD/HNMR*. Advanced Chemistry Development Inc., www.acdlabs.com
138. L.A.Gribov, K.A.Zinovyev. *J. Mol. Struct.*, **268**, 191 (1992)
139. L.A.Gribov, W.J.Orville-Thomas. *Theory and Methods of Calculation of Molecular Spectra*. Wiley, Chichester, 1988
140. М.Е.Эляшберг, Ю.З.Карасев, В.А.Дементьев, Л.А.Грибов. *Интерпретированные колебательные спектры углеводородов — производных циклогексана и циклопентана*. Наука, Москва, 1988
141. М.Е.Elyashberg, E.R.Martirosian, Yu.Z.Karasev, H.Thiele, H.Somberg. *Anal. Chim. Acta*, **348**, 443 (1997)
142. J.E.Dubois, G.Mathieu, P.Peguet, A.Panaye, J.P.Doucet. *J. Chem. Inf. Comput. Sci.*, **30**, 290 (1990)
143. C.Affolter, J.T.Clerc. *Chemometr. Intell. Lab. Syst.*, **21**, 151 (1993)
144. U.M.Weigel, R.J.Herges. *Anal. Chim. Acta*, **331**, 63 (1996)
145. J.H.Schuur, P.Selzer, J.Gasteiger. *J. Chem. Inf. Comput. Sci.*, **36**, 334 (1996)
146. J.Gasteiger, J.Sadowski, J.Schuur, P.Selzer, L.Steinhauer, V.Steinhauer. *J. Chem. Inf. Comput. Sci.*, **36**, 1030 (1996)
147. J.Gasteiger, J.Schuur, P.Selzer, L.Steinhauer, V.Steinhauer. *Fresenius J. Anal. Chem.*, **359**, 50 (1997)
148. K.Baumann, J.T.Clerc. *Anal. Chim. Acta*, **348**, 327 (1997)
149. G.W.Adamson, M.F.Lynch, W.G.Town. *J. Chem. Soc.*, **C**, 3702 (1970)
150. J.T.Clerc, A.Tercovics. *Anal. Chim. Acta*, **235**, 93 (1990)
151. P.Willett, J.M.Barnard, G.M.Downs. *J. Chem. Inf. Comput. Sci.*, **38**, 983 (1998)
152. N.A.B.Gray. *Chemometr. Intell. Lab. Syst.*, **5**, 11 (1988)
153. B.G.Buchanan, E.A.Feigenbaum, J.Lederberg. *Chemometr. Intell. Lab. Syst.*, **5**, 33 (1988)
154. N.A.B.Gray. *Chemometr. Intell. Lab. Syst.*, **5**, 37 (1988)
155. M.E.Munk. *J. Chem. Inf. Comput. Sci.*, **38**, 997 (1998)
156. M.E.Elyashberg, Yu.Z.Karasev, E.R.Martirosian. *Chimia*, **52**, 324 (1998)
157. W.Bremser. *Angew. Chem., Int. Ed. Engl.*, **27**, 247 (1988)
158. K.Funatsu, Y.Susuta, S.Sasaki. *J. Chem. Inf. Comput. Sci.*, **29**, 6 (1989)
159. K.Funatsu, S.Sasaki. *J. Chem. Inf. Comput. Sci.*, **36**, 190 (1996)
160. M.E.Munk, B.D.Christie. *Anal. Chim. Acta*, **216**, 57 (1989)
161. M.E.Munk, V.K.Velu, M.S.Madison, E.W.Robb, M.Christie, M.Razinger. In *Recent Advances in Chemical Information II*. (Ed. H.Collier). Royal Society of Chemistry, Cambridge, 1993. P. 247
162. M.Christie, M.E.Munk. *J. Am. Chem. Soc.*, **113**, 3750 (1991)
163. C.Peng, S.Yuan, C.Zheng, Y.Hui, H.Wu, K.Ma, X.Han. *J. Chem. Inf. Comput. Sci.*, **34**, 814 (1994)
164. C.Peng, S.Yuan, C.Zheng, Z.Shi, H.Wu. *J. Chem. Inf. Comput. Sci.*, **35**, 539 (1995)
165. P.L.Fuchs, C.A.Bunnell. *Carbon-13 NMR Based Organical Spectral Problems*. Wiley, New York, 1979
166. E.Pretsch, J.Seibl, A.Manz, W.Simon. *Aufgabensammlung zur Strukturaufklärung Organischer Verbindungen mit Spektroskopischen Methoden*. Springer-Verlag, Berlin, 1985
167. R.B.Bates, W.A.Beavers. *Carbon-13 NMR Spectral Problems*. Humana Press, Clifton, 1981
168. J.T.Clerc, E.Pretsch, J.Seibl. *Structural Analysis of Organic Compounds by Combined Application of Spectroscopic Methods*. Akademiai Kiadó, Budapest, 1981
169. E.Pretsch, A.F.Furst, M.Badertscher, S.R.Burgin, M.E.Munk. *J. Chem. Inf. Comput. Sci.*, **32**, 291 (1992)
170. N.F.Atta-ur-Rahman, M.I.Choudhary. *Solving Problems with NMR Spectroscopy*. Academic Press, New York, 1996. P. 391
171. L.A.Gribov. *J. Mol. Struct.*, **300**, 415 (1993)
172. L.A.Gribov, M.E.Elyashberg, Yu.Z.Karasev. *Anal. Chim. Acta*, **316**, 217 (1995)
173. M.E.Elyashberg, L.A.Gribov, Yu.Z.Karasev, E.R.Martirosian. *Anal. Chim. Acta*, **353**, 105 (1997)
174. Л.А.Грибов, Д.И.Сиделов, И.В.Маслов. *Журн. аналит. химии*, **53**, 706 (1998)

EXPERT SYSTEMS FOR THE DETERMINATION OF STRUCTURES OF ORGANIC MOLECULES BY SPECTRAL METHODS

M.E.Elyashberg

Research Institute of Organic Synthesis

12, Ul. Radio, 107005 Moscow, Russian Federation, Fax +7(095)261-0777

The current state of studies aimed at the creation of expert systems for establishing structures of organic molecules based on their IR and ^1H and ^{13}C NMR spectra is analysed. The complex methods used to identify molecular fragments, to generate structures and to predict their spectra are considered. The principles of construction of modern expert systems and the general strategy of the solution of structural problems are discussed.

Bibliography — 174 references.

Received 25th February 1999