Э к о л о г и я

УДК 621.3:543.51:547.411

ИЗВЛЕЧЕНИЕ И СТРУКТУРИРОВАНИЕ ЗНАНИЙ ПРИ ИДЕНТИФИКАЦИИ СУПЕРЭКОТОКСИКАНТОВ ПО МАСС-СПЕКТРАМ НИЗКОГО РАЗРЕШЕНИЯ

Е. И. Абрамов[1], ***Н. М. Богатов***[2], ***Н. Н. Буков***[3], ***В. Т. Панюшкин***[4]

EXTRACTION AND STRUCTURIZATION OF KNOWLEDGE DURING IDENTIFICATION OF SUPERECOTOXIC SUBSTANCES ACCORDING TO LOW-RESOLUTION MASS SPECTRA

Abramov E. I., Bogatov N. M., Bukov N. N., Panyushkin V. T.

The paper discusses the problem of designing a knowledge base for integrated expert systems aimed to identify superecotoxic substances according to low-resolution mass spectra. The system of identification is offered, which is based on the sequence of expert's elementary steps while analyzing mass-spectra of chloro-organic pesticides.

Введение

Идентификация органических соединений необходима для решения задач экологии, фармакологии и медицины, химии, пищевой технологии и др. Особенно актуальна эта проблема при анализе суперэкотоксикантов, получивших свое название за способность разрушительно влиять на организм человека даже в микроскопических дозах [1]. В выбросах промышленных предприятий суперэкотосиканты перемешаны в различных сочетаниях, и зачастую в малых концентрациях. Поэтому проблема их распознавания достаточно сложна. Для обнаружения и идентификации суперэкотоксикантов используются методы масс-спектрометрии как достаточно чувствительные и информативные. Эти исследования стоят дорого, так как используется современная сложная экспериментальная техника и привлекаются квалифицированные эксперты для анализа. Применение систем, основанных на искусственном интеллекте, повышает эффективность и достоверность анализа [2].

Цель данной работы — представить некоторую методологическую основу конкретного подхода к процессам выявления и представления высказываний эксперта о свойствах проблемной области, содержащих элементы нечёткости, неопределённости и неточности на примере анализа масс-спектров хлорорганических пестицидов (ХОП).

1. Методика анализа

Формализуем процесс извлечения и структурирования знаний при анализе масс-

[1]Абрамов Евгений Иванович, доцент кафедры общей физики и информационных систем Кубанского государственного университета.

[2]Богатов Николай Маркович, д-р физ.-мат. наук, профессор, заведующий кафедрой общей физики и информационных систем Кубанского государственного университета.

[3]Буков Николай Николаевич, доцент кафедры общей и неорганической химии Кубанского государственного университета.

[4]Панюшкин Виктор Терентьевич, д-р хим. наук, профессор, заведующий кафедрой общей и неорганической химии Кубанского государственного университета.

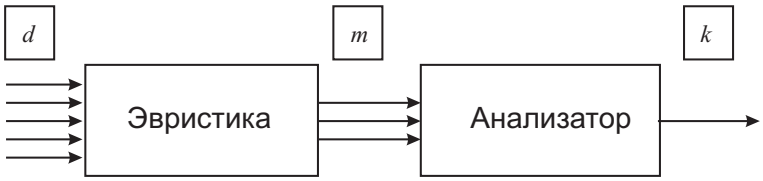

Рис. 1. Условная схема идентификации суперэкотоксикантов по масс-спектрам низкого разрешения: d — размерность входного масс-спектра, m — размерность вектора, после применения эвристики, k — идентифицированное химическое соединение

спектров. Для этого выделим систему эвристик, которой пользуется эксперт, представляющих в целом частный опыт эксперта, и сам анализ, имеющий собственную структуру. Здесь и далее под эвристикой понимается совокупность приемов и методов, заменяющих формальный перебор. Структура анализа — это система последовательных элементарных шагов. Структуры системы эвристик и элементарных шагов анализа тесно связаны (рис. 1). Первичной является структура элементарных шагов анализа. Для выполнения элементарного шага анализа, эксперту фактически предлагается представить в явном виде ту структуру, от которой зависит система эвристик. Система эвристик позволяет уйти от бесконечного перебора. Многократные стрелки иллюстрируют эффект сокращения размерности входных данных.

На первом этапе представим систему элементарных шагов анализа в виде простого перечня, структурированного пустыми строками. Например:

1. Выделение из экспериментального масс-спектра группы пиков с максимальным отношением m/z.

2. Выделение из этой группы только пиков с приращением 2 по m/z.

Второй этап состоит в группировке эвристик с элементарными шагами анализа. Например, первому шагу соответствует эвристика — поиск молекулярного иона, а второму шагу — анализ вещества на присутствие хлора, брома и серы. Результатом этого этапа является перечень шагов анализа, в котором для некоторых шагов перечислены эвристики. Следующий этап состоит в формировании некоторой структуры, в которой делается акцент на следующие элементы:

а) возможные альтернативы при выполнении этого элементарного шага;

б) типичные и нетипичные действия для данного шага.

С точки зрения общей методологии, для анализа некоторой сложной структуры $\boldsymbol{A}$ (в нашем случае системы эвристик, используемых экспертом) строится другая структура $\boldsymbol{B}_i$, от которой зависит структура $\boldsymbol{A}$; $\boldsymbol{B}_i$ состоит из отдельных элементарных шагов, выполняемых экспертом и связанных с ними эвристик. Далее строится модель $\boldsymbol{C}_j$ представления для каждого $\boldsymbol{B}_i$ и рекурсивно уточняется структура $\boldsymbol{A}$.

2. Экспериментальная часть

Для идентификации сложных органических молекул по масс-спектрам низкого разрешения разработана оригинальная программа распознавания действующих веществ ХОП по мультиплетной (изотопной) структуре сигналов молекулярного иона (M^+) [3]. Программа основана на структурной зависимости интенсивностей сигналов мультиплета M^+ от количества входящих в состав молекулы ХОП атомов хлора [4], позволяющая определять брутто-формулы химических соединений. Эта возможность объясняется тем, что интенсивности изотопных пиков мультиплета M^+ определяются только элементным составом и относительным содержанием изотопов элементов, входящих в состав исследуемого вещества. Данная программа может быть успешно использована в тех случаях, когда интенсивность пика или группы пиков (мультиплета) молекулярного иона достаточно высока, то есть для пиков I–II уровней интенсивностей, согласно классификации [4].

Результаты расчета интенсивностей сигналов мультиплетов M^+ для молекул, содержащих до восьми атомов хлора, приведены в табл. 1. Так как в состав молекул некоторых ХОП могут входить и атомы брома и серы, имеющие изотопный сигнал с приращением 2, табл. 1 дополнена значениями интенсивностей для этих элементов.

Таблица 1

Расчетные интенсивности сигналов мультиплета молекулярного иона с различным содержанием атомов хлора, брома и серы в молекуле

N	Число атомов Cl, Br, S	$\sum m_i$	Положение и интенсивность пика в мультиплете							
			M^+	M^++2	M^++4	M^++6	M^++8	M^++10	M^++12	M^++14
1	Cl	35	100,0	32,0						
2	Cl2	70	100,0	65,0	11,0					
3	Cl3	105	100,0	97,0	31,0	3,0				
4	Cl4	140	77,0	100,0	49,0	11,0	1,0			
5	Cl5	175	63,0	100,0	64,0	20,0	3,0	0,2		
6	Cl6	210	52,0	100,0	88,0	34,0	8,0	1,0	0,05	
7	Cl7	245	45,0	100,0	96,0	51,0	16,0	3,0	0,3	0,02
8	Cl8	280	35,0	89,0	100,0	64,0	26,0	7,0	1,0	0,1*
9	Br	79	100,	97,0						
10	Br2	158	51,0	100,0	49,0					
11	Br3	237	34,0	100,0	98,0	32,0				
12	Br4	316	18,0	69,0	100,0	65,0	16,0			
13	ClBr	114	77,0	100,0	24,0					
14	Cl2Br	149	61,0	100,0	46,0	6,0				
15	Cl3Br	184	52,0	100,0	64,0	17,0	2,0			
16	ClBr2	193	44,0	100,0	69,0	13,0				
17	Cl2Br2	228	39,0	100,0	89,0	31,0	4,0			
18	ClBr3	272	27,0	86,0	100,0	48,0	8,0			
19	S	32	100,0	4,0	0,02					
20	S2	64	100,0	9,0	0,2	0,002				
21	S3	96	100,0	13,0	1,0	0,01	0,001			
22	S4	128	100,0	18,0	1,0	0,04	0,001	0,0001		

* — интенсивности пика M^++16 меньше 0,02% от максимального пика мультиплета в таблице не приводится, т.к. на практике не регистрируется

Рассмотрим последовательность элементарных шагов, выполняемых экспертом при идентификации исследуемого соединения по масс-спектру низкого разрешения:

Шаг 1. Определяется максимальный по значению m/z пик в масс-спектре исследуемого соединения (*процедура определения молекулярного иона*).

Шаг 2. Определяется количество пиков, входящих в мультиплет молекулярного иона:

– выделяется все пики, начиная с максимального, с уменьшением значения m/z на единицу и две до тех пор, пока разница по m/z не станет > 2;

– формируются сигналы с приращением $\Delta m/z = 2$, так как изотопы Cl, Br и S отличаются на две единицы атомной массы (*процедура определения мультиплета молекулярного иона*).

Шаг 3. Выполняется анализ выделенных пиков:

– если выделен всего один пик с интенсивностью I–III уровня, то он принимается за M^+. В этом случае может быть сделан вывод об отсутствии в составе молекулы атомов хлора;

– если пиков несколько, то за M^+ принимается самый низкий пик по m/z и выполняется нормировка всех выделенных пиков, принимая максимальный по интенсивности за 100%;

– анализируется форма мультиплета (*процедура анализа мультиплета молекулярного иона*).

Шаг 4. Выполняется сравнение интенсивностей сигналов мультиплета M^+ с таблицей различных содержаний хлора, брома и серы

Таблица 2

Экспериментальные интенсивности сигналов мультиплета* молекулярных ионов ДВ некоторых ХОП

Брутто-формула ДВ пестицида	Положение и интенсивность пика в мультиплете							
	M^+	M^++2	M^++4	M^++6	M^++8	M^++10	M^++12	M^++14
$C_{14}H_9Cl_5$	57,1	100,0	61,9	19,0	2,4	0,0		
$C_6H_6Cl_6$	52,0	100,0	80,0	32,0	8,0	0,0	0,0	
$C_{12}H_8Cl_6$	50,0	100,0	85,7	32,1	10,7	0,0	0,0	
$C_{10}H_5Cl_7$	44,5	100,0	96,2	51,5	16,6	3,2	0,35	0,017
$C_{10}H_6Cl_8$	29,4	92,9	100,0	64,7	23,5	0,0	0,0	0,0

* — указаны только сигналы с $\Delta m/z = 2$

(табл. 1) (*процедура расчета фактора несовпадений*).

Шаг 5. Анализируется фактор несовпадения:

– если минимальное значение одно, то определяется количество атомов хлора (брома и серы) в анализируемом соединении;

– если минимальных значений несколько, то выбирается соответствующее по значению m/z молекулярному иону, определенному для данного соединения (*процедура определения брутто-формулы*).

В экспериментальном мультиплете M^+ наряду с сигналами с приращением +2 будут проявляться и сигналы с приращением +1, обусловленные наличием в составе молекул органических веществ атомов водорода, углерода, азота и кислорода. Так как их интенсивность гораздо меньше, чем рассматриваемых атомов галогенов и серы, при распознавании мультиплета их вкладами можно пренебречь. По сигналу M^++1 определяется количество атомов углерода, входящих в состав исследуемой молекулы.

Сравнение результатов расчета с экспериментальными значениями интенсивностей (табл. 2) сигналов мультиплета M^+ для молекул действующих веществ ХОП: ДДТ ($C_{14}H_9Cl_5$), альдрин ($C_{12}H_8Cl_6$), гексахлоран ($C_6H_6Cl_6$), гептахлор ($C_{10}H_5Cl_7$) и хлордан ($C_{10}H_6Cl_8$) показывает малое (менее 1%) расхождение данных для теоретических изменений интенсивностей сигналов мультиплета. Учитывая реальную чувствительность масс-спектрометров низкого разрешения (как наиболее распространенного прибора у исследователей-экологов) этим расхождением можно пренебречь.

Более важной является проблема учета всех сигналов мультиплета M^+ в тех случаях, когда интенсивность сигнала M^+ мала (менее 10%), и в составе мультиплета M^+ могут быть не учтены сигналы с низкой интенсивностью, как M^++16 для позиции 8 и M^++4–10 для позиций 19–22 (табл. 1).

Как видно из табл. 2, в экспериментальном спектре ДДТ мультиплет M^+ несколько отличается от теоретического (табл. 1) с сохранением общего образа мультиплета.

Для решения проблемы распознавания образа мультиплета M^+ был введен фактор несовпадения ($\mathrm{dist}(a, x)$ — мера Кульбака [5]) формы интенсивностей пиков мультиплета молекулярного иона определяемого соединения и изотопных пиков, приведенных в табл. 1

$$\mathrm{dist}(a, x) = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^{N} \left[\frac{a_i}{n_1} - \frac{x_i}{n_2} \right] \log \frac{n_2 a_i}{n_1 x_i},$$

где x_i и a_i — относительные интенсивности пиков мультиплета определяемого вещества (например, ДДТ) и спектров изотопных мультиплетов из табл. 1. Количество атомов хлора определяется по номеру позиции N в табл. 1 с помощью минимизации фактора несовпадения $\mathrm{dist}(a^j, x)$

$$N = \min(\mathrm{dist}(a^j, x)), \quad j = 1, 2, \ldots, p$$

(p — количество позиций в табл. 1).

Результаты расчетов фактора несовпадения для вышеперечисленных ХОП приведены в табл. 3. Характер его изменения на примере ДДТ показан на рис. 2. Как видно из рис. 2, фактор несовпадения является минимальным для распознаваемого мультиплета молекулярного иона ДДТ. Возможные близкие значения фактора несовпадения для других позиций обычно легко различимы в масс-спектрах по значениям M^+ [4].

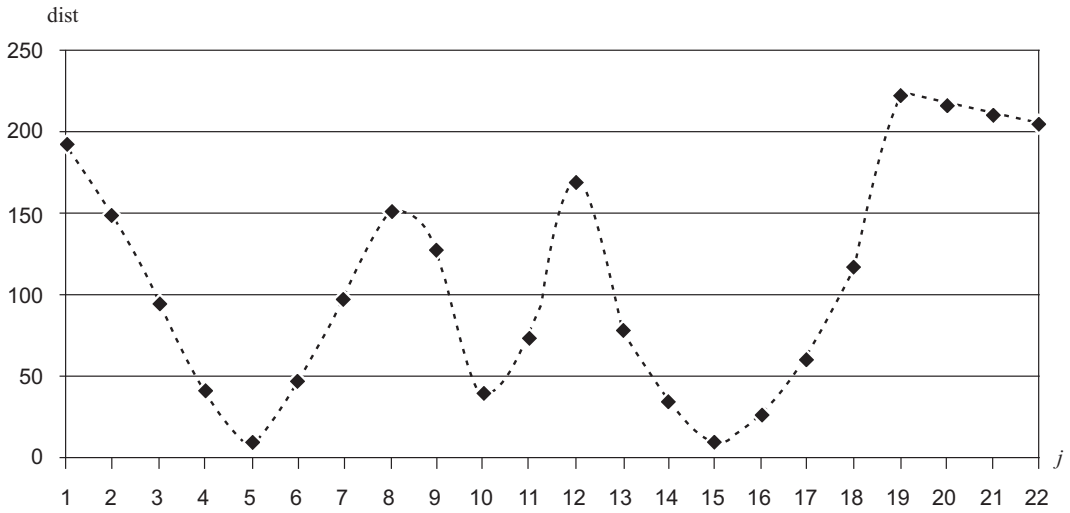

Рис. 2. Графическое представление фактора несовпадения dist для мультиплета молекулярного иона ДДТ ($C_{14}H_9Cl_5$). По оси абсцисс отложены номера пестицидов из табл. 3

Таблица 3

Значения фактора несовпадения при сравнении мультиплетов молекулярных ионов рассматриваемых пестицидов

Пестицид № позиции	ДДТ $C_{14}H_9Cl_5$	ГХЦГ $C_6H_6Cl_6$	Альдрин $C_{12}H_8Cl_6$	Гептахлор $C_{10}H_5Cl_7$	Хлордан $C_{10}H_6Cl_8$
1	194,19	236,00	246,57	291,00	319,76
2	150,19	192,00	202,57	247,00	275,76
3	95,19	137,00	147,57	192,00	228,88
4	42,19	84,00	94,57	139,00	181,88
5	**10,52**	41,00	51,57	96,00	138,88
6	48,81	**7,00**	**11,29**	48,00	92,88
7	98,81	57,00	46,43	**2,00**	46,88
8	150,81	109,00	98,43	54,00	**20,29**
9	128,19	170,00	180,57	225,00	263,88
10	40,48	72,00	80,57	125,00	167,88
11	74,57	44,00	39,14	53,00	69,88
12	169,81	128,00	117,43	81,00	45,18
13	79,19	121,00	131,57	176,00	218,88
14	35,19	77,00	87,57	132,00	174,88
15	10,67	36,00	44,57	89,00	131,88
16	27,67	45,00	51,57	86,00	126,88
17	61,81	28,00	23,14	49,00	77,88
18	119,81	78,00	72,86	53,00	42,59
19	222,19	264,00	274,57	319,00	347,76
20	217,19	259,00	269,57	314,00	342,76
21	212,19	254,00	264,57	309,00	337,76
22	207,19	249,00	259,57	304,00	332,76

Например, мультиплет молекулярного иона ДДТ ($C_{14}H_9Cl_5$) достаточно близок к мультиплету вещества, содержащего три атома хлора и один атом брома (позиция 15, табл. 1). Однако поиск пестицида с таким содержанием атомов галогенов и/или молекулярной массой равной 352–362 у.е. [6] не дал положительного результата. Для Тедиона ($C_{12}H_6Cl_4O_2S$) и Мильбекса ($C_{12}H_6Cl_4N_2S$), входящих в этот диапазон масс, имеем гораздо более высокий фактор несовпадения.

В свою очередь, ГХЦГ и альдрин, как и другие пестициды с одинаковыми наборами атомов галогенов, но различными молекулярными массами, легко различимы по значениям M^+ (в данном случае 288 и 362, соответственно).

3. Результаты

Использование предлагаемого подхода идентификации суперэкотоксикантов по масс-спектрам низкого разрешения позволило извлечь и структурировать знания, применяемые экспертом, получить список возможных кандидатов и выбрать максимально достоверный, основываясь на интуиции эксперта, формализованной в виде последовательности элементарных шагов.

Показано, что методология разработки базы знаний идентификации суперэкотоксикантов по масс-спектрам низкого разрешения во всех своих фазах и независимо от приемов, применяемых при ее реализации, строится с учетом интуиции эксперта.

Основной путь сокращения числа интуитивных ошибок состоит в многократном повторении операций и проверок. Решения о том, где и как применять правила, согласно которым ведется разработка базы знаний, принимаются интуитивно. Поэтому более глубокая проблема разработки базы знаний состоит в построении экспертом теории компьютерных решений.

Создана оригинальная программа с использованием предложенной методологии, которая на практике показала эффективность применения экспертных эвристик для снижения размерности анализируемых масс-спектров суперэкотоксикантов. Основной результат применения предлагаемого подхода состоит в точной идентификации хлорсодержащих пестицидов по структуре мультиплета молекулярного иона в масс-спектрах низкого разрешения.

Литература

1. *Майстренко В. Н., Хамитов Р. З., Будников Г. К.* Эколого-аналитический мониторинг супертоксикантов. М.: Химия, 1996. 319 с.
2. *Джон М. Т.* Программирование искусственного интеллекта в приложениях / Пер. с англ. А. И. Осипова М.: ДМК Пресс, 2004. 312 с.
3. *Абрамов Е. И., Богатов Н. М., Буков Н. Н., Панюшкин В. Т.* Экспертная идентификация органических соединений по масс-спектрам низкого разрешения с использованием алгоритма «ионных серий» // Информационные технологии моделирования и управления: Междунар. сб. научн. трудов. Вып. 13. Воронеж: Научная книга, 2004. С. 3–7.
4. *Зенкевич И. Г., Иоффе Б. В.* Интерпретация масс-спектров органических соединений. Л.: Химия, 1986. 176 с.
5. *Гайдышев И.* Анализ и обработка данных: специальный справочник. СПб: Питер, 2001. 752 с.
6. Справочник по пестицидам / Под ред. А. В. Павлова. Киев: Урожай, 1968. 432 с.

Статья поступила 30 июня 2005 г.
Кубанский государственный университет
© Абрамов Е. И., Богатов Н. М., Буков Н. Н., Панюшкин В. Т., 2005