РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 6 1995

КРАТКИЕ СООБЩЕНИЯ

УДК 519.25 : 577.15/17 : 681.142.1

А. Л. Осипов, Р. Д. Семенов
(Новосибирск)

МОДЕЛИ ПРОГНОЗИРОВАНИЯ ТОКСИКОЛОГИЧЕСКИХ СВОЙСТВ ХИМИЧЕСКИХ ВЕЩЕСТВ

Предложены и исследованы математические модели предсказания классов опасности и количественных параметров токсичности органических соединений, основанные на моделях многомерной регрессии, теории распознавания образов и статистических решений. Разработана программная реализация предложенных подходов и проведены вычислительные эксперименты по предсказанию LD_{50} на экспериментальном материале, показывающие высокую эффективность системы.

Введение. Поиск новых высокоэффективных и безопасных для человека и окружающей среды химических веществ является важнейшей проблемой мировой науки. Значительное место в этой проблеме занимает задача прогноза параметров токсичности (LD_{50}, CK_{50} и др.) органических соединений, необходимость практического решения которой тесно связана со следующими обстоятельствами:

на этапе разработки и эксплуатации технологических процессов — с отставанием и неполнотой обоснования санитарно-гигиенических нормативов на используемое сырье, продукты и отходы;

на этапе поисковых исследований (синтез и биологические испытания) — с необходимостью возможно более ранней оценки токсичности новых химикатов с целью дополнительной фильтрации токсичных целевых соединений.

Таким образом, токсикометрия занимает значительное место в принятии радикальных решений по профилактике неблагоприятных воздействий химических веществ в окружающей среде. На стадии синтеза новых соединений и композиций она позволяет осуществлять целенаправленный отбор менее токсичных и опасных соединений, используя для этого целый набор качественных и количественных критериев. Широкое использование при таком отборе математических методов, компьютерных технологий и фактографических банков данных позволяет отсеивать заведомо неактивные или высокотоксичные вещества, тем самым значительно сокращая сроки создания физиологически активных соединений с заданными токсикологическими свойствами.

Математические модели прогноза токсичности. Теоретической базой для построения моделей и развития расчетных методов определения токсичности является объективно существующая связь между токсическим действием вещества, его физическими свойствами и химической структурой [1]. Из-за отсутствия в большинстве случаев адекватных теоретических представлений о механизмах биологического действия, сложности процессов, происходящих с веществом в живых системах, широкое применение находят эмпирические закономерности, устанавливающие связь между строением молекул и их физико-химическими и токсикологическими характеристиками. В данной работе исследуются эмпирические обобщения в форме современных методов и моделей многомерной регрессии, а также теории распознавания образов. В качестве

информационной поддержки исследуемых моделей использовался фактографический банк данных по токсичности органических молекул объемом в 4624 соединения различных структурно-химических классов. Предсказание LD_{50} осуществлялось в два этапа. На первом этапе осуществлялся качественный прогноз, позволяющий определить класс токсичности или опасности вещества, что является весьма актуальной задачей, так как во многих химических исследованиях нет необходимости в строгой оценке параметров токсичности и достаточно знать классы опасности веществ [2, 3]. На втором этапе в каждом из классов токсичности строились оптимальные регрессионные зависимости и по ним осуществлялся количественный прогноз.

Модель качественного прогноза. Прогноз класса токсичности осуществлялся с использованием моделей и алгоритмов распознавания образов и теории статистических решений. Рассматривалась задача распознавания образов применительно к случаю двух классов. Отметим, что при любом другом числе классов последовательным разбиением на два класса можно построить разделение и на произвольное число k классов. Для этого достаточно провести k разбиений по принципу отделения элементов первого класса от смеси остальных, затем элементов второго класса от остальных и т. д.

Обозначим через H_1 соответствующий класс токсичности. Будем рассматривать объекты обучающей выборки, входящие в H_1, как положительные примеры класса H_1, а объекты, не входящие в H_1, — как контрпримеры или отрицательные объекты класса H_1, множество которых обозначим через H_2. Запишем бинарный вектор наблюдений X в виде $(d_1, d_2, \ldots, d_n)$, где $d_i = 1$ или 0 в зависимости от того, присутствует или отсутствует i-й фрагмент структуры в описании соединения. Обозначим через $p_i = P(d_i = 1/H_1)$ и $q_i = P(d_i = 1/H_2)$ вероятности появления i-го дескриптора в классах H_1 и H_2 соответственно.

В предположении условной независимости можно записать условные плотности распределения вероятностей в каждом классе в виде произведения вероятностей для компонент вектора наблюдений:

$$P(X/H_1) = \prod_{i=1}^{n} p_i^{d_i}(1 - p_i)^{1 - d_i},$$

$$P(X/H_2) = \prod_{i=1}^{n} q_i^{d_i}(1 - q_i)^{1 - d_i}.$$

Отношение правдоподобия при этом определяется выражением

$$\frac{P(X/H_1)}{P(X/H_2)} = \prod_{i=1}^{n} \left(\frac{p_i}{q_i}\right)^{d_i} \left(\frac{1 - p_i}{1 - q_i}\right)^{1 - d_i}.$$

Прологарифмировав это отношение, получим байесовскую решающую функцию

$$l(X) = \sum_{k=1}^{n} d_k l_k + l_0,$$

где

$$l_k = \log \frac{p_k(1 - q_k)}{q_k(1 - p_k)}$$

— информационный вес k-го дескриптора, а

$$l_0 = \sum_{k=1}^{n} \log \frac{1 - p_k}{1 - q_k}$$

— константа. Байесовское решающее правило, минимизирующее среднюю вероятность ошибки, запишется следующим образом [4, 5]:

$$\text{если } l(X) > \log \frac{p(H_2)}{p(H_1)}, \quad \text{то } X \in H_1, \quad \text{иначе } X \in H_2.$$

При выводе решающего правила мы исходили из того, что потери при правильной классификации равны нулю, а при ошибочной — единице. При построении систем распознавания возможны такие ситуации, когда априорные вероятности появления объектов соответствующих классов $p(H_1)$ и $p(H_2)$ неизвестны. Применительно к этой ситуации рационально использовать минимаксный критерий, который минимизирует максимально возможное значение среднего риска. Показано [6], что минимаксное правило представляет собой специальное правило Байеса для наименее благоприятных априорных вероятностей. В этом случае решающая граница выбирается так, чтобы обеспечить равенство ошибок первого и второго рода, которые соответственно вычисляются по формулам:

$$\varepsilon_1 = \int_{H_2} P(X/H_1)dX \quad \text{и} \quad \varepsilon_2 = \int_{H_1} P(X/H_2)dX.$$

Оценка величин p_i и q_i осуществляется по конечному числу выборочных представителей образов в соответствующих классах:

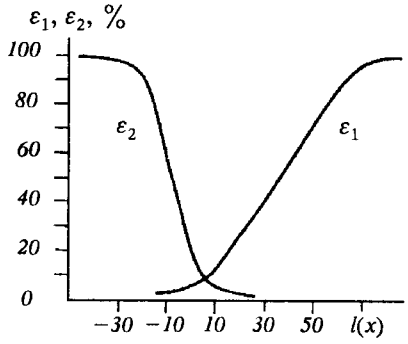
$$p_i = \frac{h_{i1} + 1}{N_1 + 2}, \qquad q_i = \frac{h_{i2} + 1}{N_2 + 2},$$

где h_{i1}, h_{i2} — числа наличия i-го дескриптора в первом и втором классах; N_1, N_2 — объемы выборок в этих классах.

Проверка работоспособности и эффективности решающего правила исследовалась на обучающих выборках по определению класса токсичности, указание которых вместе с выбором информативных подструктурных фрагментов осуществлялось автоматически при помощи оригинальной СУБД и системы запросов к базе данных. Вся выборка разбивалась на четыре класса опасности. Первый класс содержал 479 соединений, и показатель токсичности располагался в интервале $0 < LD_{50} \leq 50$, второй — 654 соединения и $50 < LD_{50} \leq \leq 200$, третий — 1402 соединения и $200 < LD_{50} \leq 1000$, а четвертый — 2089 соединений и $LD_{50} > 1000$. В качестве признаков распознавания использовались подструктурные дескрипторы, порождаемые автоматически и описанные на простом языке. Это язык описания атомов и функциональных групп с учетом валентного состояния, а также их цепочки произвольной длины с указанием нахождения атома или группы в цепи, кольце или мостике. Информативность дескрипторов оценивалась по критерию дивергенции Кульбака

$$D_i = \frac{N_1 N_2}{N_1 + N_2} (p_i - q_i) l_i,$$

которая является мерой различимости двух выборок по i-му признаку [7], при этом выбирались те из них, у которых критерий превышал пороговое значение. Отнесение химического соединения к соответствующему классу токсичности проводилось по значениям $1 - \varepsilon_2^k$, где ε_2^k — ошибка второго рода для k-го класса



Экспериментальные оценки ошибок первого и второго рода для четвертого класса опасности

в зависимости от отношения правдоподобия l, а значение k, на котором достигается $\max_u(1 - \varepsilon_2^u)$, и является номером класса опасности. Точность предсказания (процент правильных решений) при скользящем контроле по выбранным классам токсичности колебалась в пределах от 89 до 95 %, что хорошо видно из полученных экспериментальных оценок ошибок ε_1 и ε_2. На рисунке приведен график зависимости этих ошибок для четвертого класса опасности.

Модель количественного прогноза. Количественный прогноз осуществлялся на основе неаддитивных моделей с использованием понятия о парциальных вкладах структурных элементов [8]. Используемые модели параметров, входящих в структурно-неаддитивные модели, имеют вид

$$f = f_0 + \sum_{k=1}^{m} f_k d_k,$$

где f_k — парциальный вклад k-х структурных элементов в параметр f; d_k — доля k-х структурных элементов в молекуле:

$$d_k = \frac{n_k}{\sum_{i=1}^{m} n_i}.$$

В нашем случае в качестве параметра f использовался нормированный показатель токсичности

$$\ln \frac{LD_{50}}{M},$$

где M — масса молекулы. В каждом классе опасности строились оптимальные регрессионные уравнения, в которых величины f_k определялись, исходя из экспериментальных данных, устойчивым методом наименьших квадратов. В качестве подструктурных элементов использовались следующие типы подграфов:

— атомы (фрагменты) с валентным состоянием, например, −O−;

— атомы (фрагменты) с учетом первого окружения, например, CH=C=O;

— цепочки атомов (фрагментов) произвольной длины без указания промежуточных вершин, но с указанием промежуточных связей, например, CH−−=CH_2.

Результаты одного из вычислительных экспериментов с использованием скользящего контроля и дисперсионного анализа приведены в табл. 1.

Известно [9], что основной параметр токсичности $lgLD_{50}$ определяется в экспериментах на животных, причем обычно стандартное отклонение, связанное с погрешностью эксперимента, находится в пределах 0,3—0,5. Там же отмечено, что естественная биологическая вариабельность этой величины еще больше, поскольку известно, что она зависит от возраста животных, времени года и еще от многих факторов, определяющих ее резистентность. Анализ ошибок наблюдателя для соединений, неправильно классифицированных по байесовскому алгоритму, показал, что при прогнозе количественных значений LD_{50} относительные ошибки их предсказания не превосходят 94 %. Из приве-

Таблица 1

Классы токсичности	I	II	III	IV
Остаточная сумма квадратов	86,32	74,79	160,33	364,42
Сумма квадратов регрессии	1814,52	568,60	1275,80	11760,44
Полная сумма квадратов	1899,84	643,39	1436,13	12124,86
Средний квадрат регрессии	36,29	11,37	25,52	235,21
Дисперсия ошибок	0,289	0,164	0,163	0,251
Стандартная ошибка	0,538	0,405	0,404	0,501
Коэффициент детерминации	0,955	0,884	0,888	0,970
Коэффициент корреляции	0,977	0,940	0,943	0,985
Критерий Фишера	126	70	157	936
Процент необъясненного стандартного отклонения LD_{50}	21,2	34,1	33,4	17,3
Средняя относительная ошибка, %	23	20	19	17

денных результатов и вышеизложенного следует вывод о высокой эффективности системы при компьютерном расчете LD_{50}, сравнимой с экспериментальным определением этой величины. Хотелось бы отметить, что коммерческий пакет TopKat [10, 11] имеет стандартное отклонение, связанное с погрешностью прогнозирования, равное 0,62, и коэффициент корреляции, равный 0,721. Все это позволяет сделать заключение о том, что прогноз по разработанным моделям дает более высокую точность, чем использование пакета TopKat.

Система компьютерной поддержки. Разработана автоматизированная информационно-поисковая система [12], оснащенная программами математических процедур статистического моделирования токсикологических свойств химических веществ, состоящая из: 1) подсистемы поддержки профессиональных структурно-химических баз данных и знаний; 2) подсистемы прогнозирования токсических свойств органических молекул с учетом или без учета их физико-химических параметров, позволяющей создавать обучающие и экзаменационные выборки из баз данных, задавать или выбирать из меню различные описания химической структуры или иных признаков, выбирать различные модели статистической обработки данных для построения решений о принадлежности молекул к тому или иному классу токсичности, а также структурно-аддитивные и неаддитивные математические модели, которые используются для нахождения корреляции между структурами и свойствами.

Эта компьютерная система позволяет осуществлять прогноз токсикологических параметров веществ с использованием моделей теории распознавания образов и кусочно-линейных регрессионных моделей, где интервалами линейности являются классы опасности химических соединений.

В качестве примера приведем машинный прогноз параметра токсичности LD_{50} для соединения с химическим названием malononitrile, o-chlorobenzylidene, которое в нашу выборку не входило и имеет экспериментальное значение токсичности 178 мг/кг. При качественном прогнозе система отнесла данное соединение ко второму классу опасности. Результаты количественных прогнозов авторов и по программе TopKat приведены в табл. 2.

Таблица 2

Модель прогноза	Прогноз, мг/кг	Абсолютная ошибка	Относительная ошибка, %
Авторы статьи	146,9	31,1	17,5
Пакет TopKat	285	107,0	60,1

Заключение. Созданная компьютерная информационно-поисковая система представляет собой мощный инструмент для оперативного прогноза в режиме диалога токсикологических

показателей для проверки на больших выборках гипотезы о связи структуры веществ с их биологическим действием, а также для анализа сравнительной информативной ценности различных групп факторов при изучении механизмов взаимодействия веществ с живым организмом.

СПИСОК ЛИТЕРАТУРЫ

1. **Курляндский Б. А., Шитиков В. К., Тихонов В. Н.** Прогнозирование значений ПДК и других нормативов методом регрессионного анализа с использованием информационно-поисковой системы // Гигиена и санитария. 1986. № 8.
2. **Белик А. В., Гусева В. В., Зайцев Ю. А., Тужилкова Т. Н.** Оценка класса токсичности производных тиазолидина методом потенциальных функций // Химико-фармацевтический журнал. 1993. № 12.
3. **Нигматуллин Р. С., Осипов А. Л., Пузаткин А. П.** Способ предсказания токсичности химических веществ по их молекулярным спектрам // Использование вычислительных машин в спектроскопии молекул и химических исследованиях. Новосибирск, 1989.
4. **Нигматуллин Р. С., Осипов А. Л., Пузаткин А. П., Коптюг В. А.** Статистический метод предсказания биологической активности многоатомных молекул на основе дескрипторов графов структурных формул // Химико-фармацевтический журнал. 1985. № 2.
5. **Дуда Р., Харт П.** Распознавание образов и анализ сцен. М.: Мир, 1976.
6. **Фукунага К.** Введение в статистическую теорию распознавания образов. М.: Наука, 1979.
7. **Кульбак С.** Теория информации и статистика. М.: Наука, 1967.
8. **Зацепин В. М., Нигматуллин Р. С., Осипов А. Л.** Структурно-аддитивные и структурно-неаддитивные модели расчета свойств органических соединений // Синтез и применение пестицидов и кормовых добавок в сельскохозяйственном производстве. Волгоград, 1988.
9. **Павленко Ю. С., Никонов А. М.** Прогнозирование распространения загрязнений в приземном слое атмосферы // Гигиена и санитария. 1993. № 6.
10. **Meyer D. E., Warr W. A., Love R. A.** Chemical structure software for personal computers // Amer. Chem. Soc. Washington, DC, 1988.
11. **Venable H. L.** The developments and application of algorithms for generating estimates of toxicity for the NOHS data base // NIOSH Technical Report. DHHS (NIOSH), Pub. N87—101, 1986.
12. **Осипов А. Л., Нигматуллин Р. С., Семенов Р. Д.** Создание компьютерной системы предварительной оценки мутагенных, токсикологических и канцерогенных свойств химических соединений // Математические проблемы экологии. Новосибирск, 1994.

Поступило в редакцию 16 августа 1995 г.