РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 5 1995

УДК 539.19 : 681.3

В. М. Зацепин, А. Л. Осипов, Р. Д. Семенов
(Новосибирск)

СИСТЕМА КОМПЬЮТЕРНОГО ПРЕДСКАЗАНИЯ ФИЗИКО-ХИМИЧЕСКИХ И БИОЛОГИЧЕСКИХ СВОЙСТВ ВЕЩЕСТВ

Рассматривается компьютерная система предсказания свойств химических веществ по их структуре на основе фактографических баз данных, в которую заложены оригинальные математические модели, позволяющие обнаруживать скрытые закономерности, объясняющие связь химической структуры с их физико-химическими и биологическими действиями.

Введение. Компьютеризация научной и учебной деятельности в областях химии, биологии, медицины и экологии характеризуется введением в научные исследования новых информационных технологий: систем поддержки профессиональных химико-структурно-биологических баз данных (БД) и знаний; интеллектуальных систем, позволяющих предсказывать и оценивать степень воздействия структурных и других характеристик химических соединений на их токсикологические, физико-химические, фармакологические и другие свойства.

Создание и использование компьютерных технологий, позволяющих отсеивать заведомо неактивные соединения, а также миновать некоторые утомительные стадии исследований и значительно сократить время разработки физиологически активных препаратов, имеет первостепенное значение для фармацевтической промышленности и др. Поэтому неудивительно, что в настоящее время для достижения такой весьма заманчивой цели прилагаются значительные усилия. Например, расходы фирмы "Pfiser" на научные исследования по поиску фармацевтических препаратов в 1990 г. возросли более чем на 20 % и составили 640 млн долларов. По оценке "International Resource Development Inc." особенно быстро растет объем финансирования работ по созданию систем искусственного интеллекта — от 66 млн долларов в 1983 г. до 8,5 млрд долларов в 1993 г. [1]. Это обусловлено тем, что многие фирмы связывают успех в разработке новых препаратов с внедрением компьютерного моделирования, резко сокращающего сроки создания физиологически активных веществ. Поэтому создание и использование отечественных банков данных и знаний по биологической активности веществ необходимо еще и потому, что доступ к информационным ресурсам развитых стран всегда будет в той или иной мере ограничен и, не располагая независимыми информационными ресурсами, невозможно оценить даже степень этой ограниченности, не говоря уже об их эффективной эксплуатации с целью создания конкурентоспособных продуктов [1].

Для целенаправленного синтеза биологически активных соединений (БАС) весьма важно установить взаимосвязи строения химических соединений с их действием. Выявление связи биологической активности химических соединений с их структурой основано на знании молекулярных механизмов взаимодействия биологических систем с химическими препаратами. Теоретические построения, описывающие механизмы действия химических соединений на живые организмы, ввиду сложности этих взаимодействий и самих биологических систем пока не привели к заметным результатам. Детальный

механизм воздействия БАС остается неизвестным, и, по-видимому, должен развиваться прагматический подход — обобщение эмпирического опыта исследований для получения практических результатов. Такой подход не дает прямых сведений о механизмах взаимодействия, обусловливающих активность, но он необходим для решения практических задач и построения теорий, объясняющих связь химической структуры соединений с их биологическим действием. В процессе эмпирического поиска накапливается информация, где отражена связь элементов строения соединений с их свойствами. Используя эту информацию на стадии планирования синтеза, химики ведут выбор или конструирование соединений с заданными свойствами. Привлечение математических методов и компьютерных технологий позволяет обнаружить скрытые закономерности, формализовать некоторые решения, более направленно и обоснованно вести поиск и синтез препаратов с заданными свойствами.

Описание компьютерной системы. Компьютерная информационно-вычислительная система состоит из:

— оригинальной СУБД CHANCE (CHemical ANalysis in Computer Environment) для IBM PC, которая поддерживает обработку таких сложно структурированных объектов, как молекулярные химические графы. СУБД имеет встроенный гипертекстовый HELP, дружественный интерфейс с пользователем, редактор входных и выходных форм. Ввод, вывод и манипулирование структурными формулами (СФ) молекул осуществляется графическим экранным редактором, основные режимы которого — рисование, удаление, редактирование, сборка из фрагментов, создание ароматических комплексов, манипулирование СФ и другие. Поиск информации в БД осуществляется по любому полю или совокупности полей, включая подструктурный поиск, который происходит путем рисования структурного фрагмента и/или фрагментов графическим редактором и внесения их в поисковые предписания на специально разработанном языке запросов. Идентификация химических веществ происходит по каноническому коду, программно порождаемому системой;

— инструментальной системы прогнозирования биологических свойств химических веществ и конструирования новых биологически активных соединений с заданными свойствами по их СФ с учетом или без учета физико-химических параметров молекул. Система позволяет создавать (с помощью системы запросов) обучающие и экзаменационные выборки из БД, задавать или выбирать из меню различные описания химической структуры или иных признаков, выбирать различные модели статистической обработки данных (байесовские алгоритмы теории статистических решений, марковские зависимости и другие) для принятия решений о принадлежности химического соединения к тому или иному типу биологической активности, оценивать их адекватность, предсказывать биологические, фармакологические, токсические, мутагенные и канцерогенные свойства органических веществ по их СФ с учетом или без учета физико-химических параметров. Точность прогнозирования (процент правильных решений) в разработанной системе при предсказании различных биологических свойств составила 85—90 % [2—5];

— инструментальной системы прогнозирования физико-химических и токсикологических свойств, основанной на оригинальных математических моделях (структурно-аддитивных и неаддитивных), применяющихся при нахождении количественных корреляций «структура—свойство», что дает возможность предсказывать такие важные параметры химических веществ, как молекулярная рефракция и липофильность молекул, которые используются в дальнейшем при прогнозировании биологической активности химических веществ. Система позволяет прогнозировать токсикологические параметры химических соединений с использованием моделей теории распознавания образов и кусочно-линейных регрессионных моделей, где интервалами линейности регрессии являются классы опасности химических веществ [6].

Математические модели прогнозирования физико-химических свойств веществ. Развитие и внедрение математических методов анализа связи «структура—активность» с прогнозированием новых активных соединений должно привести к существенному сокращению времени и объема поисковых

работ, а следовательно, и затрат на разработку. Естественно, что используемые математические модели должны по возможности учитывать совокупность накопленных теоретических и эмпирических знаний о характере процессов в биологических системах при воздействии на них активными соединениями, о связи биологического действия со структурой и физико-химическими свойствами активных соединений. Математический аппарат КССА (количественная связь «структура—активность») включает главным образом методы многомерного статистического анализа: линейный и нелинейный регрессионный анализ, дисперсионный анализ, различные методы классификации и распознавания. В настоящее время существование связи структуры и физико-химических свойств БАС с активностью является общепринятым.

В общем случае искомая величина физико-химического параметра f молекулы представляется как линейная функция многих переменных, каждая из которых есть количество подграфов СФ, принадлежащих определенному типу. Подграфами, состоящими из одной вершины и прилегающих к ней связей с другими вершинами, являются символы X^i, $i = 1, 2, ..., n$, например, CH≡, $-CH_2-$, O=, $-CH=$. Этот тип подграфов состоит из атомов или микрофрагментов с учетом валентного состояния и того, в цепи или в кольце стоят эти атомы или микрофрагменты. В качестве двухвершинных подграфов берутся различные пары связанных в структурной формуле атомов или микрофрагментов с учетом того, в кольце или цепи стоят вершины молекулярного графа (аналог «атом—связь—атом»), например, CH = CH, S = O, C – Cl или $(X^i \,\&\, X^j)$, где символ & означает тип связи между вершинами X^i и X^j. Подграфы, состоящие из трех вершин, соответствуют цепям вида $(X^i \,\&\, X^j \,\&\, X^k)$ (это аналог атомов с первым окружением), например, Cl – CH – CH_2, O = C – CH_3, O = P – O. Подграфы $(X^i \,\&\, X^j \,\&\, X^k)$ и $(X^k \,\&\, X^j \,\&\, X^i)$ считаются неразличимыми, т. е. изоморфными.

Следуя вышеуказанному, рассмотрим класс соединений, каждое из которых включает в себя не более n заданных типов атомов $X^1, X^2, ..., X^n$. Расчет физико-химического свойства f структурного соединения G из нашего класса может быть проведен по формуле

$$f(G) = \sum_{i=1}^{n} a_i\{X^i\} + \sum_{(i,\,j)} a_{ij}\{X^i \,\&\, X^j\} + ... + \sum_{(i_1,\,...,\,i_n)} a_{i_1,\,...,\,i_n}\{X^{i_1} \,\&\, X^{i_2} \,\&\, ... \,\&\, X^{i_n}\},$$

где $\{X^{i_1} \,\&\, ... \,\&\, X^{i_n}\}$ — количество подграфов типа $(X^{i_1} \,\&\, ... \,\&\, X^{i_n})$. Здесь каждый из индексов $i_1, i_2, ..., i_n$ пробегает подмножество из $(1, 2, ..., n)$, которое не включает в суммирование изоморфные подграфы. Величины $a_{i_1,\,...,\,i_\alpha}$, где $\alpha = 1, 2, ..., n$, требуется определить с помощью статистических методов, исходя из экспериментальных данных. Эта формула позволяет проводить исследование эффективности различных типов подграфов в отдельности и в связи с другими типами при расчете того или иного физико-химического свойства. В данной работе используется упрощенный вариант используемых типов подграфов (атомы с валентным состоянием «атом—связь—атом» и цепочки атомов произвольной длины без указания промежуточных вершин, т. е. типы подграфов X^i, $(X^i \,\&\, X^j)$ и $(X^{i_1} \,\&\, X^{i_2} \,\&\, ... \,\&\, X^{i_n})$, где в $(X^{i_1} \,\&\, X^{i_2} \,\&\, ... \,\&\, X^{i_n})$ промежуточных вершин $X^{i_2}, ..., X^{i_n-1}$ нет, а есть только конечные вершины X^{i_1} и X^{i_n}, связанные между собой кратчайшим путем, состоящим из типов связей между вершинами, входящими в этот путь, например, CH – – = CH_2).

Данный подход позволяет прогнозировать различные физико-химические параметры, в частности, липофильность органических молекул. В работе [7] приведены среднеквадратические погрешности предсказания этого показателя на выборке в 217 соединений двумя программами (авторов публикации и программы Med-Chem), которые составили $\sigma = 0{,}275$ и $\sigma = 0{,}258$. При нашем расчете по вышеприведенной модели с использованием скользящего контроля среднеквадратическая погрешность составила $\sigma = 0{,}157$, что дает заметно более высокую точность, чем в [7]. Следует отметить, что в [8] была определена

среднеквадратическая погрешность $\sigma = 0{,}119$ на 172 производных бензола, что хорошо согласуется и с нашими расчетами по этой выборке данных. Хотелось бы отметить, что эти эксперименты строились на ограниченной выборке химических веществ в рамках определенного структурно-химического класса, что не позволяет переносить полученные результаты на другие классы.

На основе компьютерной системы CHANCE была создана база данных по экспериментальным значениям липофильности объемом в 2778 соединений из различных химических классов, которая использовалась для проведения различных вычислительных экспериментов по прогнозированию этого параметра, результаты одного из которых, основанные на таблице дисперсионного анализа [9], приведены ниже:

остаточная сумма квадратов	262,54
сумма квадратов регрессии	12305,15
полная сумма квадратов	12567,69
средний квадрат регрессии	58,6
дисперсия ошибок	0,157
стандартное отклонение	0,396
коэффициент детерминации	0,979
коэффициент корреляции	0,989
критерий Фишера	373,25
табличное значение Фишера F (210, 1688, 99,9 %)	1,76
процент необъясненного стандартного отклонения параметра липофильности	14,5 %

Другой подход компьютерного моделирования связи молекулярного строения с биологической активностью состоит в порождении иных дескрипторов, которые отражают пространственные корреляции локальных физико-химических свойств на молекулярной структуре. В данной работе использованы спектры распределения физических свойств на молекулярной структуре и по отношению к вершинам молекулярной структуры [10], дающие богатое описание молекулярных структур в терминах дескрипторов с ясной физической интерпретацией. Обобщенные спектры плотностей физических свойств по отношению к вершинам (атомам) молекулярной структуры определяются соотношениями:

$$C_{i\alpha}(l) = \sum_j \delta(r_{ij} - l)\omega_{j\alpha}, \tag{1}$$

$$C^p_{i\alpha}(l) = \left\{ \sum_j \delta(r_{ij} - l) | \omega_{j\alpha} |^p \right\}^{1/p}, \tag{2}$$

$$D_{i\alpha}(l) = \sum_j \delta(r_{ij} - l)(\omega_{j\alpha} - \omega_{i\alpha}), \tag{3}$$

$$D^p_{i\alpha}(l) = \left\{ \sum_j \delta(r_{ij} - l) | \omega_{j\alpha} - \omega_{i\alpha} |^p \right\}^{1/p}, \tag{4}$$

где $i, j = 1, \ldots, n$ — номера вершин молекулярной структуры; r_{ij} — расстояние между вершинами i и j; $\omega_{j\alpha}$ — вклад в физическое свойство, нумеруемое индексом α, связанный с вершиной j; $l = 0, \ldots, L$, $\delta(r - l)$ — символ Кронекера; p полагается равным двум.

Введенные величины (1)—(4) характеризуют распределения физических свойств на молекулярной структуре: спектры (1) и (2) — плотности физических свойств и их моделей в окружениях (слоях) центральных атомов соответ-

ственно на расстояниях $l = 0, 1, ..., L$, спектры (3) и (4) — аналогичные характеристики неоднородностей распределений физических свойств на молекулярной структуре.

Структурно-аддитивные модели расчета физико-химических параметров (неизвестных коэффициентов корреляционных уравнений «структура—свойство», в том числе и нелинейных) имеют вид

$$\Pi = \Pi_0 + \sum_t \Pi_t n_t, \tag{5}$$

где n_t — число структурных элементов (молекулярных фрагментов) t-го типа в молекуле; Π_t — инкремент t-го структурного элемента в параметр Π.

Описанные в литературе [11, 12] модели для расчета липофильности и гидрофобности характеризуются на разнородных выборках химических соединений абсолютными среднеквадратическими ошибками 0,5, что намного больше погрешности экспериментального определения липофильности (0,1—0,2). Традиционный путь уточнения структурно-аддитивных моделей связан с усложнением структурных фрагментов (например, посредством более детального учета первого и второго окружения центральных атомов фрагментов), что приводит к увеличению числа подлежащих определению параметров.

Альтернативой является уточнение структурно-аддитивных схем посредством учета влияния окружения исходных структурных фрагментов (в качестве которых в данной работе взяты атомы с учетом валентного состояния) через локальные физико-химические свойства, в первую очередь электронные и стерические. В качестве исходных факторов при конструировании моделей неаддитивных вкладов использованы спектры физических свойств (локальных зарядов и ван-дерваальсовых радиусов) по отношению к вершинам (атомам) молекулярной структуры, определенные соотношениями (1)—(4).

Рассмотрим более подробно построение моделей для расчета физико-химических параметров, использованных в комплексе CHANCE, в частности липофильности. В соотношениях (1)—(4) вместо индекса i введем нумерацию из двух индексов: $i \to \{t, i_t\}$, где $t = 1, ..., T$ — тип вершины молекулярной структуры (атом с учетом валентного состояния), $i_t = 1, ..., n_t$ — номера вершин t-го типа. Пусть $S_{\{t, i_t\}\alpha}(l)$ — любой из спектров (1)—(4). Усредненные по вершинам t-го типа спектры

$$S_{t,\alpha}(l) = \frac{1}{n_t} \sum_{i_t} S_{\{t, i_t\}\alpha}(l). \tag{6}$$

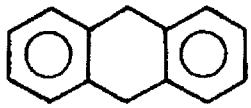
Обобщенная модель (5) имеет тот же вид, но с величинами

$$\Pi_t = \pi_t + \delta\pi_t, \tag{7}$$

где $\delta\pi_t$ — линейная или нелинейная функция компонент спектров (6) с подлежащими определению коэффициентами. Отбор существенных компонент (предикторов) осуществляется процедурами пошаговой регрессии и группового учета аргументов.

Достигнутая в рамках описанного подхода погрешность расчета липофильности составляет 0,1—0,2 в зависимости от химического класса, а на разнородных выборках ошибка около 0,35, что улучшает прогнозирование данного показателя по сравнению с работами [11, 12].

Приведем прогноз липофильности для 9, 10-dihydroanthracene с экспериментальным значением, равным 4,25 [7], который не входил в обучающие выборки и имеет следующую структурную формулу:



Метод	Значение	Остаток
Прогноз по [7]	4,55	−0,30
Прогноз по Med-Chem	4,67	−0,42
Прогноз по I модели настоящей статьи	4,20	0,05
Прогноз по II модели настоящей статьи	3,98	0,27

Заключение. Разрабатываемые базы данных и программные средства позволяют проводить интерпретацию различных структурных признаков и их совокупностей в предсказании того или иного вида биологической активности, а также в возможности построения компактных математических моделей связи «структура—активность» с использованием таких интегральных свойств химической структуры, как липофильность, молекулярная рефракция и т. д., которые часто количественно коррелируют с самыми разнообразными биологическими свойствами.

СПИСОК ЛИТЕРАТУРЫ

1. **Забежайло М. И.** Новые информационные технологии в научных исследованиях и технологических разработках // НТИ. Сер. 2. Информационные процессы и системы. 1992. № 6.
2. **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 // Хим.-фарм. журн. 1985. № 2.
3. **Нигматуллин Р. С., Осипов А. Л., Лазуткин Е. Ю.** Многотерминальная система для конструирования оригинальных химических структур с заданными свойствами // Телекоммуникационные средства использования банков данных: Сб. научн. тр. ГПНТБ СО АН СССР. Новосибирск, 1990.
4. **Осипов А. Л., Семенов Р. Д., Нигматуллин Р. С.** Использование химических баз данных для построения корреляций типа «структура—свойство» с помощью АРМ исследователя // Тез. докл. Междунар. конф. «Автоматизированные библиотечно-информационные системы». Новосибирск, 1993.
5. **Осипов А. Л., Нигматуллин Р. С., Семенов Р. Д.** Компьютерная система поиска и анализа данных о структурных, биологических и физико-химических свойствах веществ (CHANCE) // Тез. докл. II Всерос. конф. по мат. проблемам экологии. Новосибирск, 1994.
6. **Осипов А. Л., Нигматуллин Р. С., Семенов Р. Д.** Создание компьютерной системы предварительной оценки мутагенных, токсикологических и канцерогенных свойств химических соединений // Там же.
7. **Camilleri P., Watts S. A., Boraston J. A.** A surface area approach to determination of partition coefficients // J. Chem. Soc. Perkin Trans. II. 1988. N 9. P. 1699.
8. **Смоленский Е. А., Пономарева Л. А., Зефиров Н. С.** Новый подход к расчету липофильности органических соединений // ДАН СССР. 1990. **312**, № 1.
9. **Афифи А., Эйзен С.** Статистический анализ: Подход с использованием ЭВМ: Пер. с англ. М.: Мир, 1982.
10. **Зацепин В. М.** Представление химических соединений как распределений физических свойств на молекулярных структурах // Вопросы алгоритмического анализа структурной информации (Вычислительные системы. Вып. 119): Сб. науч. тр. Новосибирск: ИМ СО АН СССР, 1987.
11. **Lyman W. I., Rehl W. F., Rosenblatt D. H.** Handbook of Chemical Property Estimation Methods. N. Y.: Mc Graw-Hill, 1982.
12. **Hansh C., Leo A.** Substituent Constants for Correlation Analysis in Chemistry and Biology. N. Y.: Wiley, 1979.

Поступила в редакцию 13 февраля 1995 г.