РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
АВТОМЕТРИЯ
№ 3 1998

УДК 539.19 : 681.3

А. Л. Осипов, А. А. Башелханов, М. В. Борисов
(Новосибирск)

СИСТЕМА МОДЕЛИРОВАНИЯ ПАРАМЕТРОВ, ПРЕДСТАВЛЯЮЩИХ ЭФФЕКТЫ БИОЛОГИЧЕСКОЙ СРЕДЫ

Описана компьютерная система для проведения научных исследований по моделированию связи между строением вещества и его свойствами с использованием фактографических банков данных. Приводятся результаты моделирования физико-химических параметров, представляющих эффекты биологической среды, с использованием компьютерной системы в сравнении с отечественными и зарубежными подходами.

Введение. Центральной проблемой теоретической химии является нахождение зависимостей между структурой органических соединений и проявляемыми ими свойствами (физико-химическими, биологическими и др.). Решение данной проблемы невозможно в настоящее время без привлечения средств вычислительной техники и методов математического моделирования, так как они позволяют отказаться от традиционного метода поиска химических веществ с заданными свойствами путем экспериментов, которые являются чрезвычайно сложными, длительными и дорогостоящими. Так, при исследованиях химического канцерогенеза обычными лабораторными средствами характерное время эксперимента – около трех лет, а стоимость изучения одного соединения – около 12 млн долларов США. Если учесть, что к настоящему времени известно более 13 млн химических веществ, число которых ежегодно увеличивается на 500 тысяч вновь созданных, то утверждение о затрудненности экспериментального анализа их обычным путем становится просто очевидным [1]. Аналогично эффективность работ по поиску новых химических средств защиты растений обычными средствами (путем синтеза и массовых испытаний большого числа самых различных химических соединений) чрезвычайно мала, так как для создания конкурентоспособного нового действующего вещества необходим синтез и биологические испытания от 30 до 100 тысяч химических соединений, причем стоимость разработки нового продукта без затрат на создание промышленного производства составляет 10–35 млн долларов, а время, затрачиваемое на разработку, – около 10 лет [2]. Поэтому применение математических методов и компьютерных технологий для установления количественных соотношений структура – активность и их использование для прогнозирования потенциально активных соединений становятся одними из важных и основных путей изыскания препаратов с заданными

свойствами, так как они приведут к существенному сокращению времени и объема поисковых работ, а следовательно, и затрат на разработку.

Модели предсказания свойств химических веществ. Сложность проблемы связи молекулярная структура – биологическая активность обусловлена особенностями биологических систем, характеризующихся взаимодействием с окружающей средой и сложной совокупностью отдельных взаимосвязанных химических реакций и процессов транспорта, обмена веществ и энергий, протекающих в разнородных биологических структурах. Для достижения биологического эффекта существенны следующие фазы воздействия химического вещества на биологические системы: кинетическая, динамическая, а также воздействия, которые по своей природе являются физико-химическими. Все эти фазы характеризуются различными физико-химическими параметрами и делятся на электронные, стерические и гидрофобные, которые представляют эффекты среды. Отсюда вытекает принципиальная возможность установления связи биологической активности с физико-химическими характеристиками вещества и биологической системы. Электронные параметры, используемые в КССА (количественная связь структура – активность), делятся на корреляционные константы (типа Гаммета) и квантово-механические параметры. Большинство из этих параметров (энергия молекулярных орбиталей, электронная плотность и т. д.) характеризуют способность молекул к донорно-акцепторному связыванию с рецептором. Одним из широко применяемых и полезных параметров для КССА является молекулярная рефракция, отражающая объем заместителей и их способность к дисперсионным взаимодействиям. Гидрофобные параметры описываются логарифмом коэффициента распределения октанол/вода (липофильность) и сильно влияют на проникновение, транспорт и связывание с биологической макромолекулой.

Компьютерный расчет вышеперечисленных физико-химических параметров и их использование в дальнейшем для прогноза биологической активности молекул – одна из важных задач современной химии. При моделировании этих параметров широкое применение нашли эмпирические модели, базирующиеся на принципе структурной аддитивности [3–6]. Большинство из этих моделей, используемых для целей прогнозирования физико-химических свойств веществ, обладают следующими недостатками: низкой точностью, узостью области применения по классу химических веществ и диапазону изменения параметров. Создание в настоящее время компьютерных систем ввода, хранения и обработки физико-химической и структурной информации [3, 8, 9] позволяет поставить на качественно новый уровень решение следующих задач:

– полная автоматизация расчетов по описанным в литературе моделям;

– ревизия структурно-аддитивных моделей с использованием новых структурных элементов;

– разработка новых структурно-неаддитивных моделей.

Далее на ряде конкретных примеров моделирования связи между строением вещества и его физико-химическими свойствами показана возможность разработки новых структурно-неаддитивных моделей. Предлагается подход, который позволяет прогнозировать различные физико-химические параметры, в частности липофильность органических молекул. Именно липофильные свойства определяют способность молекул проникать через липидные слои мембран и гидрофобное взаимодействие их с отдельными участками рецептора. Предлагаемый подход компьютерного

моделирования связи молекулярного строения с физико-химическими свойствами [3] состоит в порождении дескрипторов, которые отражают пространственные корреляции локальных физико-химических свойств на молекулярной структуре. В данной работе использовались лишь некоторые из них, а именно спектры распределений физических свойств на молекулярной структуре и спектры плотностей физических свойств по отношению к вершинам молекулярной структуры [3], дающие богатое описание молекулярных структур в терминах дескрипторов с ясной физической интерпретацией, которые определяются соотношениями:

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

$$D_{i\alpha}(l) = \sum_j \delta(r_{ij} - l)(\omega_{j\alpha} - \omega_{i\alpha}), \qquad (2)$$

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, который связан с вершиной j; $l = 0, \ldots, L$, $\delta(r - l)$ – символ Кронекера.

Введенные величины (1), (2) характеризуют распределения физических свойств на молекулярной структуре: спектры (1) – плотности физических свойств и их моделей в окружениях (слоях) центральных атомов соответственно на расстояниях $l = 0, 1, \ldots, L$ спектры (2) – аналогичные характеристики неоднородностей распределений физических свойств на молекулярной структуре.

Описанные в литературе [5, 10] модели для расчета липофильности характеризуются на разнородных выборках химических соединений абсолютными среднеквадратичными ошибками 0,4, что намного больше погрешности экспериментального определения параметра липофильности $\log P$ (0,1 – 0,2). Традиционный путь уточнения структурно-аддитивных моделей связан с усложнением структурных фрагментов (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 плотностей физических свойств (локальных зарядов и ван-дер-ваальсовых радиусов) по отношению к вершинам (атомам) молекулярной структуры, определенные соотношениями (1), (2).

Более подробное построение структурно-неаддитивных моделей, механизм которого заложен в компьютерной системе, для расчета параметра липофильности приведено в [3].

Реализация компьютерно-моделирующей системы. Для моделирования связи химическая структура – физико-химические свойства – биологическая активность разработана высокоэффективная компьютерная система MP (Make Prognosis), реализованная на персональных компьютерах под управлением операционной системы MS DOS. MP позволяет строить

модели предсказания физико-химических и биологических свойств химических веществ и предусматривает реализацию следующих основных этапов моделирования:

– выявление дескрипторов структуры и молекулярных параметров, описывающих существенные для проявляемой активности особенности строения соединений обучающей выборки;

– формирование модели описания молекулярной структуры и построение на ее основе математической модели предсказания физико-химических и/или биологических свойств химических веществ;

– проверка модели на адекватность физико-химическим и биологическим данным.

Интерфейс пользователя компьютерной системы состоит из четырех частей: 1) область системных сообщений; 2) блок общей информации; 3) блок описания классов, включающих обучающие и экзаменационные выборки; 4) область функциональной клавиатуры.

В область системных сообщений помещается информация о текущем состоянии системы и все сообщения об ошибках, которые дублируются звуковым сигналом. В области функциональной клавиатуры приведен список исполняемых системой следующих команд:

Выбор (*F4*) – составление выборки по логическим условиям над полями базы данных (БД).

Итоги (*F5*) – просмотр на экране разметки БД с результатами анализа, изменение состава выборки.

Анализ (*F6*) – анализ помеченных документов выборки.

Прогон (*F7*) – полный анализ БД.

Печать (*F8*) – печать и сохранение результатов анализа.

DOS (*F9*) – выход в операционную систему без выгрузки системы МР.

Конец (*F10*) – завершение работы с системой.

Блок общей информации состоит из семи полей:

имя прогноза – это имя будет использоваться для идентификации прогноза в системе;

имя базы данных – БД, по которой будет производиться прогноз;

имя ведущего поля – используется для идентификации документа для пользователя в итоговой сводке;

имя поля структуры – идентифицирует поле для генерации дескрипторов и проведения анализа типа структура – физико-химические и биологические свойства;

общий заголовок прогноза – используется для идентификации прогноза пользователем;

число независимых классов – применяется в прогнозе (любое число от двух до лимита системы по памяти);

типы используемых дескрипторов: атом (микрофрагмент) с валентным состоянием; атом (микрофрагмент) – связь – атом (микрофрагмент); атом (микрофрагмент) с первым окружением; атом (микрофрагмент) со вторым окружением; все кратчайшие пути в графе структурной формулы; вышеперечисленные дескрипторы, но с разметкой – в цепи, кольце или мостике находятся вершины графа; различные информационно-топологические индексы [7]; дескрипторы, отражающие прост-

ранственные корреляции локальных физико-химических свойств на молекулярной структуре (электронные и стерические) [3].

Блок описания классов состоит из заголовка, который используется для идентификации класса пользователем, и логических условий, которые характеризуют принадлежность документа к данному классу. Условия классов представляют собой логические операции, скомбинированные знаками & (и), | (или), ! (логическое отрицание) и скобками () произвольной вложенности. Логические операции: сравнение целых и вещественных чисел (<, >, =, !=, <=, >=, в качестве операндов могут выступать числовые константы и имена полей целого и вещественного типов, определенные в БД); операция поиска включения подстроки в строку <: (в качестве операндов используются имена текстовых полей в БД или текстовые константы); операция существования * *op*1 дает истину, если поле *op*1 в данном документе БД не пустое, и ложь в противном случае. Пример контроля гипотетического документа на принадлежность классу:

$$* \textit{тип} \,\&\, * \textit{индекс} \,\&\, (\text{"}\textit{деф}\text{"} <: \textit{тип} \mid \textit{индекс} > 55{,}6),$$

где предполагается наличие в БД текстового поля *тип* и целого или вещественного поля *индекс*. Документ принадлежит классу, если поля *тип* и *индекс* определены, в поле *тип* есть подстрока *деф* или индекс данного документа превышает число 55,6.

Эксперименты по проверке эффективности предсказания параметра липофильности. Для проверки эффективности предсказания коэффициента распределения была создана база данных по экспериментальным

Таблица 1

Характеристика модели	Экзамен
Остаточная сумма квадратов	37,03
Дисперсия ошибок	0,0365
Стандартная ошибка	0,1911
Полная сумма квадратов	1201,46
Сумма квадратов регрессии	1164,43
Коэффициент детерминации	0,969
Коэффициент корреляции	0,984
Необъясненное стандартное отклонение параметра липофильности, %	17,55
Средний квадрат регрессии	23,29
Критерий Фишера	638
Средняя относительная ошибка, %	10
Количество соединений	1064

значениям липофильности объемом в 3500 соединений из различных химических классов, которая и послужила основой для проведения различных вычислительных экспериментов по прогнозированию этого параметра.

Достигнутая в рамках описанного подхода погрешность расчета липофильности составляет 0,1–0,2 в зависимости от химического класса. При проведении вычислительного эксперимента с использованием скользящего контроля по предсказанию параметра липофильности для таких рядов соединений, как углеводороды ароматические, алифатические, спирты, фенолы, простые эфиры, фенолята, погрешность расчета составила 0,19, что видно из табл. 1 дисперсионного анализа.

При выполнении предсказания на более разнородных выборках ошибка на контрольной выборке составляет 0,31, что улучшает прогнозирование данного показателя по сравнению с работами [5, 10–12].

Дальнейшие эксперименты заключались в сравнении экстраполяционной способности предложенного метода расчета липофильности с известным методом гидрофобных фрагментарных констант Реккера [11] и методом на основе топологических индексов (данный метод был опробован на обучающей выборке объемом 195 химических веществ и экзаменационной выборке объемом 83 соединения) [12]. Получены следующие соотношения между стандартными ошибками, основанными на экзаменационной и обучающей выборках: для метода на основе топологических индексов [12] $\sigma_э/\sigma_о$ = =1,369 для метода Реккера [11, 12] $\sigma_э/\sigma_о$ = 6,081, для метода, предложенного

Таблица 2

Характеристика модели	Обучение	Экзамен
Остаточная сумма квадратов	144,827380	168,039450
Дисперсия ошибок	0,087668	0,098211
Стандартная ошибка	0,296088	0,313387
Полная сумма квадратов	2298,700394	2393,854508
Сумма квадратов регрессии	2153,873014	2225,815058
Коэффициент детерминации	0,936996	0,929804
Коэффициент корреляции	0,967986	0,964263
Необъясненное стандартное отклонение параметра липофильности, %	25,1	26,5
Средний квадрат регрессии	43,077460	44,516301
Критерий Фишера	492	454
Средняя относительная ошибка, %	15	16
Количество соединений	1702	1761

в данной работе, $\sigma_{эк}/\sigma_{0} = 1{,}057$, что следует из полученных результатов моделирования (табл. 2).

Заключение. Созданная интегрированная компьютерная система, включающая базы фактографических данных, математические модели, методы и программные средства, представляет собой мощный инструментарий, который дает возможность в режиме диалога вести оперативный прогноз физико-химических и биологических свойств, проверять на больших выборках гипотезы о связи структуры веществ с их биологическим действием, а также анализировать информативную ценность различных групп факторов при изучении механизмов взаимодействия веществ с биологическими системами.

СПИСОК ЛИТЕРАТУРЫ

1. **Забежайло М. И.** К проблеме расширения ДСМ-метода автоматического порождения гипотез на данные с числовыми параметрами // НТИ. Сер. 2. Информационные процессы и системы. 1993. № 2.
2. **Мельников Н. Н.** Пестициды. Химия, технология и применение. М.: Химия, 1987.
3. **Зацепин В. М., Осипов А. Л., Семенов Р. Д.** Система компьютерного предсказания физико-химических и биологических свойств веществ // Автометрия. 1995. № 5. С. 86.
4. **Рид Р., Прауснии Дж., Шервуд Т.** Свойства газов и жидкостей. Л.: Химия, 1982.
5. **Hansh C., Leo A.** Substituent Constants for Correlation Analysis in Chemistry and Biology. N. Y.: Wiley. 1979.
6. **Татевский В. М.** Теория физико-химических свойств молекул и веществ. М.: МГУ, 1987.
7. **Химические** приложения топологии и теории графов /Под ред. Р. Кинга. М.: Мир, 1987.
8. **Strokov I. I., Lebedev K. S.** A new modular architecture lor computer systems in chemistry // J. Chem. Inf. Comput. Sci. 1996. N 36.
9. **Хазановский К. П.** Графический редактор химических структур CHAR как средство создания новых программных систем по химии // Итоги науки и техники. Сер. Информатика. Т. 15. М.: ВИНИТИ. 1991.
10. **Lyman W. I., Rehl W. F., Rosenblatt D. H.** Handbook of Chemical Property Estimation Methods. N. Y.: Mc Graw-Hill, 1988.
11. **Rekker R. F.** The Hydrophobic Fragmental Constant. Amsterdam: Elsevier Scientific Publishing Company, 1977.
12. **Дрбоглав В. В.** Инварианты графов и их использование для обработки структурной информации: Автореф. дис. ... канд. техн. наук /СО АН СССР. ИОХ. Новосибирск, 1987.

Поступила в редакцию 16 июля 1997 г.