«Вестник МИТХТ», 2008, т. 3, № 6

Математические методы и информационные технологии в химии и химической технологии

УДК 541.6

МОДЕЛИРОВАНИЕ СВЯЗИ «СТРУКТУРА-СВОЙСТВО» ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ НА ОСНОВЕ БАЗИСНЫХ ПОДГРАФОВ МОЛЕКУЛЯРНЫХ ГРАФОВ

М.И. Скворцова, доцент
кафедра Высшей и прикладной математики
МИТХТ им. М.В. Ломоносова
e-mail: skvorivan@mail.ru

Предложен новый метод построения количественных соотношений, связывающих структуру и свойства органических соединений, представленных мечеными графами. Метод основан на некоторых результатах спектральной теории графов. Приведены примеры применения метода.

Ключевые слова: *модели связи «структура-свойство», молекулярные графы, базисные подграфы*

Введение

Одно из важнейших направлений современной теоретической химии связано с построением количественных соотношений, связывающих структуру и свойства химических соединений. Полученные соотношения могут быть использованы для оценки свойств соединений, для которых отсутствуют экспериментальные данные, для разработки математических моделей каких-либо физико-химических процессов или механизмов действия биологически активных веществ.

Важное место в этих исследованиях занимают способы количественного описания структуры молекул. Один из наиболее распространенных подходов к этой проблеме основан на представлении органических соединений в виде меченых (взвешенных) графов, вершины и ребра которых соответствуют атомам и связям молекулы, а метки вершин и ребер кодируют атомы и связи различной химической природы. Для количественного описания структуры таких графов, называемых молекулярными, используются их инварианты – числа, определяемые по матрице весов графа, не зависящие от способа нумерации его вершин.

Для получения модели связи «структура-свойство» обычно используется так называемый статистический подход, суть которого заключается в следующем. Имеется выборка соединений с известными численными значениями некоторого свойства (физико-химического или биологической активности). Структура каждого из этих соединений описывается при помощи набора инвариантов $x_1,\ldots,x_n$ соответствующих молекулярных графов. Как правило, математическая модель связи «структура-свойство» в рамках этого подхода имеет вид уравнения, связывающего исследуемое свойство y и параметры $x_1,\ldots,x_n$ при помощи некоторой функции f от n переменных:

$$y=f(x_1,\ldots,x_n) \qquad (1)$$

Тип функции f предполагается известным (например, линейная или квадратичная функция); однако f зависит от ряда подгоночных параметров. Эти параметры подбираются по известным численным значениям рассматриваемого свойства соединений заданной выборки так, чтобы соотношение (1) выполнялось бы как можно более точно на этой выборке.

В процессе построения моделей описанным выше способом возникают проблемы выбора инвариантов графов и вида функции f из бесконечно большого числа возможных вариантов. Это связано с тем, что заранее неизвестно, от каких факторов зависит изучаемое свойство, и каким образом.

В настоящей работе предложен новый общий метод построения моделей типа (1), основанный на некоторых результатах спектральной теории графов, а также приведены примеры его применения.

Метод построения моделей

Изложим кратко результаты теории графов, которые лежат в основе разработанного метода.

Известно, что взвешенный граф восстанавливается однозначно по набору его собственных чисел и соответствующих линейно независимых собственных векторов: имеется формула, связывающая матрицу графа и эти данные. Однако в общем случе

граф не определяется однозначно по набору собственных чисел: имеется много примеров неизоморфных графов с одинаковыми собственными числами. Известно также, что коэффициенты характеристического полинома взвешенного графа с n вершинами (т. е., по сути, его собственные числа) однозначно определяются по набору его подграфов на k=1,2,..., n вершинах, состоящих из объединения изолированных вершин, ребер и циклов [1]. В связи с этим представляется целесообразным поиск подграфов, определяющих однозначно и собственные вектора графа. Нами было предложено решение вышеуказанной задачи: найдены подграфы, определяющие собственные вектора графа, и выведены соответствующие формулы для вычисления компонент этих векторов [2, 3]. Так как собственные вектора графа связаны с нумерацией его вершин, то и найденные подграфы, очевидно, зависят некоторым образом от нумерации вершин графа. Однако эти подграфы имеют вполне определенную структуру, а зависимость от нумерации вершин выражается в виде некоторых ограничений на номера вершин, принадлежащих тому или иному подграфу.

Используя эти результаты и игнорируя ограничения на номера вершин, можно выделить набор подграфов (не зависящих от нумерации вершин), которые участвуют в восстановлении как собственных значений, так и собственных векторов графа, т.е. графа в целом. Эти подграфы имеют k=1,2,..., n вершин и состоят из объединения следующих несвязных компонент: изолированных вершин, ребер, циклов, цепей длины более двух (при этом подграф может содержать не более одной такой цепи). Описанный выше набор подграфов несет в себе полную информацию о структуре графа, и любой инвариант графа есть некоторая функция, зависящая от чисел вхождения в граф этих подграфов, а также от их некоторых структурных характеристик. Назовем выделенные подграфы *базисными*.

На основе полученных теоретико-графовых результатов можно предложить следующий общий метод построения моделей связи «структура-свойство»: для описания структуры молекулярных графов использовать инварианты, равные числам вхождения в граф базисных подграфов (называемых далее также базисными), а в качестве аппроксимирующей функции использовать многочлен нескольких переменных от этих параметров. Этот многочлен можно построить, например, следующими способами.

Способ 1. Из набора базисных инвариантов выберем k параметров ($k<N$ – фиксированное число, N – число соединений в базе данных), дающих наилучшую линейную модель. Обозначим множество выбранных параметров через M_1. Затем рассмотрим квадраты и попарные произведения этих инвариантов. Обозначим это множество инвариантов через M_2. Построим наилучшую линейную модель, содержащую k параметров, выбирая их из объединения M_1 и M_2. Затем процедуру повторяем; при этом в качестве основных инвариантов используются те, что были отобраны на втором шаге, и т. д. Процедура заканчивается, когда ее очередной шаг не приводит к улучшению модели или когда достигнута требуемая точность модели. Остается открытым вопрос об оптимальном выборе числа k. Однако в общем случае нельзя указать теоретически обоснованное правило выбора k. Наши исследования показали, что при $k\approx N/3$ можно получить достаточно хорошие результаты.

Способ 2. Этот подход построения многочлена использует некоторую модификацию стандартного метода пошаговой линейной регрессии. Напомним, что этот метод заключается в следующем: для получения линейной модели с k параметрами, выбранными из некоторого набора параметров, сначала отбирают один параметр, дающий наилучшую модель; затем к нему добавляют второй, дающий наилучшую двупараметровую модель, и т. д. В случае модификации этого метода для наилучшего параметра x, отобранного на i-ом шаге, тестируются последовательно его степени x^m (m=2,3,…) с целью найти x^m (m >1) лучший, чем x. Если оказалось, что x^{m+1} хуже, чем x^m при некотором $m\geq 1$, то эта процедура прекращается и в модель добавляется параметр x^m. Процесс построения модели заканчивается, если она содержит требуемое число параметров или имеет заданную точность.

Отметим, что параметры для построения модели можно выбирать не из всего полного набора базисных инвариантов, а лишь из какой-то его части, вводя ограничения на число вершин рассматриваемых подграфов.

Тестирование метода. Проведено тестирование предлагаемого метода

построения моделей связи «структура-свойство» для некоторых баз данных (БД) по структурам и свойствам органических соединений, а также его сравнение с другими методами для этих же БД. Рассмотрены следующие БД (*N* – число соединений в базе):

I. Галоидпроизводные метана и этана с известными значениями их наркотической активности $lnAD_{50}$ (AD_{50} – концентрация вещества, вызывающая анестезию у половины подопытных животных; *N*=15; данные взяты из [4]). Построены молекулярные графы этих соединений, полностью соответствующие классическим структурным формулам. Для построения модели связи «структура-активность» рассмотрены базисные подграфы следующего вида: 1) изолированные вершины с метками *C, H, Cl, F*; 2) цепи длины 1, 2, 3 со всеми возможными расположениями меток *H, C, Cl, F*; 3) подграфы на четырех вершинах, состоящие из двух несмежных ребер со всевозможными расположениями вышеуказанных меток на их вершинах.

Получена модель, линейная относительно параметров g_1, g_2^2, g_3, g_1g_2, g_4 (коэффициент корреляции *R=0.990,* среднеквадратичное отклонение *s=0.25)*. Здесь параметры g_1, g_2, g_3, g_4 соответствуют фрагментам вида *Cl, HCF, FCCl, ClCCl.* В работе [4] сообщается о линейной корреляции между $lnAD_{50}$ и параметром *V*(ат. ед) – электростатическим потенциалом в точке, расположенной на расстоянии 2 Å от атома *H* в направлении связи *C-H (R=0.606)*. Отметим, что при использовании предложенных нами параметров наилучшая корреляция с одним параметром (g_1) имеет *R=0.857*. Таким образом, однопараметрическая модель, получаемая в рамках предложенного подхода, является более точной, чем модель, основанная на использовании квантовохимического параметра.

II. Нитробензолы и нитротолуолы с известными значениями мутагенной активности $ln\mu$ (на *Salmonella typhimurium,* μ - количество ревертантов на наномоль; *N*=14; данные взяты из [4]). Соединения представлены графами, представляющими собой шестичленные циклы (с однократными ребрами), на вершинах которых расставлены метки *C, B, A*, соответствующие атомам углерода в бензольном кольце с заместителями *H,* NO_2, CH_3, соответственно. Рассмотрены следующие базисные подграфы этих графов: 1) изолированные вершины с метками *A, B, C*; 2) ребра с метками *BB, BC, CC*; 3) цепи длины 2: *CCC, BCC, BCB, BBC, CBC, BBB*; 4) цепи длины 3: *BBCC, BBCB, CBBC, BBBC, CCCB, BCCB, CBCC*; 5) подграфы, состоящие из двух изолированных ребер типа: *CC/CC, CB/CB, CB/CC, CB/BB, BB/CC*.

Получена модель, линейная относительно параметров g_1, $(g_2g_3)^2$, $(g_1g_3)^2$, g_4, g_1g_5 (*R*=0.984, *s*=0.466). Здесь параметры g_1, g_2, g_3, g_4, g_5 соответствуют фрагментам вида: *B, BB/CC, BBC, A, C.* В [4] сообщается о корреляции $ln\mu$ с энергией нижней свободной молекулярной орбитали (*R*=0.940). При использовании предлагаемого подхода для однопараметровой корреляции *R*=0.914, а для двупараметровой - *R*=0.940.

III. Хлорзамещенные анилины с известными значениями токсичности $logEC_{50}$, где EC_{50} – концентрация вещества (миллимоль/литр), вызывающая уменьшение интенсивности люминесценции в два раза у морских бактерий *Photobacterium phosphoreum* (так называемый Microtox™ test); N=17; данные взяты из [5]. Структуры этих соединений представлены в виде вершинно-меченых графов с метками *A, B, C* следующим образом: бензольному кольцу соответствует шестичленный цикл с однократными ребрами, метки *A, B, C* вершин этого цикла соответствуют атомам углерода в бензольном кольце с заместителями NH_2, *Cl, H*, соответственно. Рассмотрены следующие базисные подграфы этих графов: 1) вершины с меткой *B*; 2) цепи длины 1 и 2 со всеми возможными расстановками меток *A, B, C*; 3) цепи длины 3 со всеми возможными комбинациями меток *B* и *C*; 4) подграфы, состоящие из двух изолированных ребер со всеми возможными комбинациями меток *B* и *C*. Полученное уравнение является линейным относительно следующих параметров: g_1 , g_2, g_3^6, g_4^3, g_5^3, g_6 (*R*=0.981, *s*=0.14). Параметры g_1, g_2, g_3, g_4, g_5, g_6 соответствуют фрагментам следующего вида: *CCC, CBC, BB/BB, CB/BB, BBBC, BBBB.*

Сравним полученный результат с аналогичным результатом, приведенным в [5]. В этой работе для моделирования связи «структура-свойство» был использован широко известный метод TLSER (Theoretical Linear Solvation Energy Relationship), согласно которому рассматриваемое свойство зависит линейно только от шести параметров, один из

которых – молекулярный ван-дер-ваальсов объема V, а пять других – некоторые квантовохимические параметры. В [5] установлено, что из этих шести параметров только один параметр квантовохимического типа является существенным, и для соответствующей корреляции $R=0.658$, $s=0.45$. Отметим также, что наилучшая модель при нашем подходе для одного параметра имеет $R=0.772$.

Заключение

Таким образом, в работе описан и обоснован новый общий теоретико-графовой метод построения математических моделей связи «структура-свойство». Метод имеет алгоритмический характер и не требует «угадывания» молекулярных параметров для модели. Он формально применим к любым БД химических соединений, представленных произвольно мечеными графами, и любым свойствам, измеряемым количественно. Из приведенных выше примеров следует эффективность предложенного метода, а также его преимущество по сравнению с другими подходами, основанными на использовании некоторых физико-химических теорий. Так как метод оперирует с подграфами меченых графов, то, очевидно, результат его применения зависит от способа построения этих графов.

ЛИТЕРАТУРА:

1. Цветкович, Д. Спектры графов. Теория и применение / Д.Цветкович, М. Дуб, Х. Захс – Киев: Наукова Думка,1984. – 384 с.

2. Скворцова, М. И. О связи между собственными векторами взвешенных графов и их подграфами / М. И. Скворцова, И.В. Станкевич // Дискретная математика. – 2004. – Т.16, вып. 4. – С. 32-40.

3. Skvortsova, M. I. Eigenvectors of weighted graphs: Supplement to Sachs' Theorem / M.I. Skvortsova, I.V. Stankevich // J. Mol. Struct. (THEOCHEM). – 2005. – V. 719. – P. 213-223.

4. Дьячков, П.Н. Квантовохимические расчеты в изучении механизма действия и токсичности чужеродных веществ / П.Н. Дьячков // Итоги науки и техники. ВИНИТИ. Сер. Токсикология. – 1990. – Т. 16. – 280 с.

5. Sixt, S. Quantitative structure-toxicity relationships for 80 chlorinated compounds using quantum chemical descriptors / S. Sixt, J. Altschuh, R. Bruggemann // Chemosphere. – 1995. – V. 30, №12. – P. 2397-2414.