УДК 539.3

МОДЕЛИРОВАНИЕ СВЯЗИ МЕЖДУ СТРУКТУРОЙ И ФИЗИКО-ХИМИЧЕСКИМИ СВОЙСТВАМИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ НА ОСНОВЕ ОПТИМАЛЬНЫХ АТОМНЫХ ПАРАМЕТРОВ

Ю.Ю. Яковенко, студент, М.И. Скворцова, профессор,
Н.А. Михайлова, старший преподаватель

кафедра Высшей и прикладной математики МИТХТ им. М.В. Ломоносова
e-mail: enf@mail.ru, st.m@list.ru

Предложен новый метод установления количественных соотношений, связывающих структуру и свойства органических соединений, представленных взвешенными молекулярными графами. Метод основан на подборе оптимальных весов вершин этих графов. Приведены примеры применения метода.

New method for establishing the quantitative structure-property relationships of organic compounds represented as weighted molecular graphs is suggested. It is based on the selection of optimal weights of vertices of these graphs. Examples of application of this method are given.

Ключевые слова: *модели связи «структура–свойство», взвешенные молекулярные графы, сульфиды, алкилфенолы, спирты, растворимость в воде, температура кипения, индексы удерживания.*

Key words: *structure-property models, weighted molecular graphs, sulfides, alkylphenols, alcohols, aqueous solubility, boiling point, retention indices.*

Проблема моделирования связи между структурой и свойствами органических соединений – важнейшая математическая задача современной теоретической и компьютерной химии [1, 2]. Найденные закономерности позволяют прогнозировать свойства химических соединений непосредственно по их структуре, минуя эксперимент, и могут быть использованы для целенаправленного поиска соединений с заданными свойствами.

Одним из наиболее распространенных подходов к моделированию связи «структура–свойство» является так называемый статистический подход. В качестве исходных данных для реализации этого подхода используется некоторая выборка соединений, представленных структурными формулами, для которых известны численные значения y рассматриваемого свойства (эта выборка обычно называется обучающей). Структура этих соединений описывается при помощи некоторого набора молекулярных дескрипторов $m_1,\ldots,m_n$, в качестве которых могут быть использованы топологические, электронные, геометрические характеристики молекул или значения каких-либо их физико-химических свойств. Математическая модель связи «структура–свойство» в рамках этого подхода имеет вид уравнения $y=f(m_1,\ldots,m_n)$, связывающего y и $m_1,\ldots,m_n$ при помощи некоторой функции f. Функция f задается заранее посредством некоторого аналитического выражения, зависящего, однако, от ряда подгоночных параметров. Эти параметры подбираются по обучающей выборке соединений так, чтобы вышеуказанное соотношение выполнялось бы на соединениях этой выборки как можно более точно. Обычно в качестве f используется линейная или квадратичная функция.

Важное место в этих исследованиях занимают способы описания структуры молекул. При этом любой способ молекулярного представления базируется на классической структурной формуле, которая задается изначально. Как правило, структуры органических соединений представляют в виде взвешенных (или меченых) графов, вершины и рёбра которых соответствуют атомам и связям молекулы, а веса (метки) вершин и рёбер кодируют атомы и связи различной химической природы. В качестве топологических молекулярных дескрипторов $m_1,\ldots,m_n$ используются инварианты этих графов [1, 2].

Очевидно, выбор весов графа существенно влияет на результат моделирования, так как значения вышеуказанных инвариантов зависят от этих весов. Как правило, выбираемые веса не зависят ни от класса соединений, ни от рассматриваемого свойства и в каждом конкретном случае фиксируются. Например, для насыщенных углеводородов принято рассматривать молекулярные графы без учета атомов водорода, входящих в молекулу, и полагать веса всех вершин равными нулю, а веса всех ребер – равными единице. Примерами весов вершин w_i для графов гетероатомных молекул могут служить следующие величины:

1) $w_i = Z_i^v - h_i$,

2) $w_i = \dfrac{Z_i^v - h_i}{Z_i - Z_i^v - 1}$,

3) $w_i = 1 - \dfrac{6}{Z_i}$,

где Z_i и Z_i^v – общее число и число валентных электронов *i*-ого атома соответственно, h_i – число атомов водорода, присоединенных к *i*-ому атому [1, 2].

Следует отметить, что в процессе такого моделирования возникают проблемы выбора инвариантов $m_1,\ldots,m_n$, функции f и весов вершин и ребер взвешенных молекулярных графов, представляющих химические структуры. Это связано с тем, что заранее неизвестно, от каких структурных особенностей и каким образом зависит изучаемое свойство для данного класса соединений, а для выбора инвариантов, функции и весов имеется бесконечно много вариантов. Заметим также, что результаты, полученные для одного конкретного случая, вообще говоря, не могут быть перенесены на другой.

В связи с этим разработка, обоснование и тестирование общих методов моделирования связи «структура–свойство», имеющих алгоритмический характер и допускающих компьютерную реализацию, является актуальной задачей.

В настоящей работе предложен новый общий метод моделирования связи «структура–свойство», основанный на подборе оптимальных весов вершин молекулярных графов, представляющих химические структуры. Приведены примеры его применения для построения моделей связи «структура–свойство» для различных свойств и классов соединений, показывающие его эффективность.

Описание метода

На первом этапе процесса построения модели проводится некоторая классификация атомов, входящих в структуры изучаемых соединений. Способ классификации выбирается исследователем. Например, атомы могут быть классифицированы только по химическим символам (C, H, N, O и т.д.) или по химическим символам с учетом распределения типов связей, или по картинам окружения 1-ого (или более высокого) порядка. Атомам *k*-ого класса приписывается некоторый неопределенный вес z_k, k = 1, 2,…, m. Далее предполагается, что модель связи «структура–свойство» имеет следующий вид:

$$y = \sum x_i x_j + x_0 \ , \qquad (1)$$

где x_i, x_j – веса *i*-ого и *j*-ого атомов в молекуле в соответствии с их классификацией, т.е. $x_i = z_k$, если *i*-й атом принадлежит *k*-ому классу; суммирование в формуле (1) распространено на все связи (*i*,*j*), x_0 – некоторая константа.

На следующем этапе эти веса подбираются оптимальным образом так, чтобы соотношение (1) выполнялось бы как можно более точно на обучающей выборке соединений. Для этого рассматривается нелинейная функция многих переменных

$$F(z_1,\ldots,z_k) = \sum\nolimits_p (y_p^{эксп.} - y_p^{расч.})^2$$

и ищется ее минимум и соответствующие значения переменных $z_1,\ldots,z_k$; здесь $y_p^{эксп.}$ – известное экспериментальное значение свойства *p*-го соединения, а $y_p^{расч.}$ – выражение для расчета свойства этого соединения, полученное при помощи формулы (1) и зависящее от неопределенных параметров $z_1,\ldots,z_k$. Начальные значения параметров для поиска минимума функции выбираются произвольным образом.

Итак, полученная модель, связывающая структуру и свойство соединения, представляет собой уравнение (1) с подобранными оптимальным образом (в вышеуказанном смысле) значениями параметров $z_1,\ldots,z_k$. Уравнение (1) может быть использовано для расчета свойств других соединений того же класса, не входящих в обучающую выборку.

Выбор функции, описывающей зависимость свойства от структуры молекулы, в виде (1), основан на следующих фактах. Одним из наиболее популярных инвариантов графов, используемых при моделировании связи «структура-свойство», является так называемый индекс молекулярной связности (или индекс Рандича) χ:

$$\chi = \sum (v_i v_j)^{-1/2} \ , \qquad (2)$$

где v_i, v_j – степени *i*-ой и *j*-ой вершин молекулярного графа, а суммирование распространено на все ребра (*i, j*) этого графа [3]. Индекс Рандича обладает следующим свойством: для графов-деревьев с фиксированным числом вершин он принимает свои экстремальные значения на наиболее и наименее разветвленных деревьях – цепи и графе-звезде. В силу этого свойства он может служить количественной мерой степени ветвления ациклической молекулы. Индекс χ был обобщен на случай, когда суммирование в формуле (2) происходит по заданным подграфам более общего вида, а также на случай, когда в формуле (2) вместо степеней вершин стоят некоторые фиксированные веса вершин, зависящие от характеристик соответствующих атомов (например, указанные выше веса w_i в п.1) и 2)). Эти модификации индекса χ также

нашли широкое применение в корреляциях «структура–свойство» [4, 5].

Естественно рассмотреть еще одно обобщение индекса Рандича, а именно, выражение, получающееся из формулы (2) путем замены фиксированных величин $(v_i)^{0.5}$ и $(v_j)^{0.5}$ на подбираемые атомные параметры x_i и x_j. Очевидно, что самая простая функция *f*, используемая при моделировании связи «структура–свойство» – линейная; при описании структуры молекулы при помощи только одного дескриптора *x* получаем корреляционное уравнение вида $y=ax+b$, в котором подбираются коэффициенты *a* и *b*. Используя в модели единственный молекулярный дескриптор – обобщенный индекс Рандича с подбираемыми атомными параметрами и линейную функцию *f*, получаем уравнение вида (1).

Тестирование метода

Нами проведено тестирование предлагаемого метода построения моделей связи «структура–свойство» для некоторых баз данных по структурам и физико-химическим свойствам органических соединений. Для каждого конкретного случая выбирался свой способ представления химических структур в виде молекулярных графов. Предварительно базы данных разбивались случайным образом на обучающую и тестовую выборки; по первой выборке строилась модель, вторая выборка использовалась для проверки прогнозирующей способности этой модели. Для оценки точности модели на обучающей выборке соединений строилась корреляция между расчетными и экспериментальными значениями изучаемого свойства, и для нее определялся коэффициент корреляции R_1, а также средняя относительная ошибка $\delta_{ср}(\%)_1$. Для оценки точности прогноза на тестовой выборке аналогичным образом определялись соответствующие величины R_2 и $\delta_{ср}(\%)_2$.

В подобных исследованиях обычно считается, что для физико-химических свойств модель «хорошая», если $R_1, R_2 \geq 0.95$ или если $\delta_{ср}(\%)_1, \delta_{ср}(\%)_2 \leq 5\%$.

Однако критерии качества модели могут быть и менее жесткими.

Рассматривались следующие свойства и классы соединений (N_1 и N_2 – число соединений в обучающей и тестовой выборке соответственно).

I. *Индексы удерживания алкилфенолов.* Исходные данные взяты из [6]; $N_1 = 45$, $N_2 = 5$.

Графы, представляющие эти соединения, соответствуют их структурным формулам с удаленными атомами водорода и группой OH. Атомы углерода C были классифицированы следующим образом: 1) C в ароматическом кольце, связанный с OH; 2) C в ароматическом кольце, связанный с C в алкильной группе; 3) C в алкильной группе, связанный с C в ароматическом кольце; 4) C в алкильной группе, связанный с C в ароматическом кольце и с C в алкильной группе; 5) C в алкильной группе, связанный с C в ароматическом кольце и двумя C в алкильной группе; 6) C в алкильной группе, связанный с C в ароматическом кольце и тремя C в алкильной группе; 7) C в алкильной группе, связанный с C в алкильной группе; 8) C в ароматическом кольце, но не связанный с OH; 9) C в алкильной группе, связанный с двумя C в алкильной группе.

Оптимальные значения соответствующих атомных параметров и константы x_0:

$x_1 = 9.2,\ x_2 = 19.2,\ x_3 = -0.09,\ x_4 = 4.5,\ x_5 = 7.6,$

$x_6 = 10,\ x_7 = 1.6,\ x_8 = 15.1,\ x_9 = 10.8,\ x_0 = 48.$

Статистические характеристики модели для обучающей и тестовой выборки:

$R_1 = 0.98$, $R_2 = 0.99$, $\delta_{ср}(\%)_1 = 1.31$,

$\delta_{ср}(\%)_2 = 1.53$.

II. *Температура кипения сульфидов.* Исходные данные взяты из [7]; $N_1 = 37$, $N_2 = 5$.

Графы, представляющие эти соединения, соответствуют их структурным формулам с удаленными атомами водорода. Атомы углерода C и серы S были классифицированы следующим образом: 1) C, связанный только с S; 2) C, связанный только с C и S; 3) C, связанный только с двумя C и S; 4) C, связанный с тремя C и S; 5) C, связанный только с C; 6) C, связанный только с двумя атомами C; 7) C, связанный только с тремя C; 8) атом S.

Оптимальные значения соответствующих атомных параметров и константы x_0:

$x_1 = 4.27,\ x_2 = 4.74,\ x_3 = 4.56,\ x_4 = 4.30,\ x_5 = 3.80,$

$x_6 = 5.10,\ x_7 = 4.80,\ x_8 = 4.61,\ x_0 = 10.$

Статистические характеристики модели для тестовой и обучающей выборки:

$R_1 = 0.98$, $R_2 = 0.96$, $\delta_{ср}(\%)_1 = 3.45$,

$\delta_{ср}(\%)_2 = 3.47$,

III. *Растворимость спиртов в воде.* Исходные данные взяты из [8]; $N_1 = 45$, $N_2 = 5$.

Графы, представляющие эти соединения, соответствуют их структурным формулам без атомов водорода и группы OH. Атомы углерода C были классифицированы следующим образом: 1) C, связанный только с OH и одним C; 2) C, связанный только с OH и двумя C; 3) C, связанный с OH и тремя C; 4) C, связанный только с двумя C; 5) C, связанный только с

тремя С; 6) С, связанный с четырьмя С; 7) С, связанный только с С.

Оптимальные значения соответствующих атомных параметров и константы x_0:

x_1=0.45, x_2=0.68, x_3=0.68, x_4=0.76, x_5=0.95,

x_6=1.20, x_7=0.21, x_0=0.70.

Статистические характеристики модели для обучающей и тестовой выборки:

R_1 = 0.99, R_2 = 1.00, $\delta_{cp}(\%)_1$ = 3.39,

$\delta_{cp}(\%)_2$ = 0.77.

IV. *Температура кипения спиртов.* Исходные данные взяты из [9]; N_1 = 70, N_2 = 30.

Графы, представляющие эти соединения, соответствуют их структурным формулам без атомов водорода и группы ОН. Атомы углерода С были классифицированы следующим образом: 1) С, связанный только с С; 2) С, связанный только с двумя С; 3) С, связанный только с тремя С; 4) С, связанный с четырьмя С; 5) С, связанный только с группой ОН и С; 6) С, связанный только с ОН и двумя С; 7) С, связанный с ОН и тремя атомами С.

Оптимальные значения соответствующих атомных параметров и константы x_0:

x_1=7.89, x_2=4.47, x_3=3.39, x_4=2.92, x_5=12.10,

x_6=4.42, x_7=2.80, x_0=3.50.

Статистические характеристики модели для обучающей и тестовой выборки:

R_1 = 0.97, R_2 = 0.98, $\delta_{cp}(\%)_1$ = 4.11,

$\delta_{cp}(\%)_2$ = 2.91.

Заключение

В работе предложен новый общий метод построения математических моделей связи между структурой и свойствами органических соединений, основанный на нахождении оптимальных весов для взвешенных молекулярных графов, представляющих химические структуры. Метод технически применим к любым свойствам, измеряемым количественно, и любым химическим соединениям, представляемых графами. Разработанный подход обладает определенной гибкостью: проведя более детальную классификацию атомов и используя большее количество подбираемых параметров, можно улучшить качество построенной модели. Приведены примеры применения предложенного метода для построения моделей связи «структура–свойство» для различных свойств и классов соединений, показывающие его эффективность.

ЛИТЕРАТУРА:

1. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Топологические индексы в органической химии // Успехи химии. 1988. Т. 57. С. 337–366.
2. Raevsky O.A. Molecular structure descriptors in the computer-aided design of biologically active compounds // Russ. Chem. Rev. 1999. V. 68. P. 505–524.
3. Randic M. On characterization of molecular branching // J. Am. Chem. Soc. 1975. V. 97. P. 6609–6615.
4. Kier L.B., Hall L.H. Molecular connectivity in structure-activity analisis. – N.Y.: Research Studies Press Ltd., John Wiley and Sons Inc., 1986. 262 p.
5. Kier L.B., Hall L.H. molecular connectivity in chemistry and drug research. – N.Y.: Academic Press, 1976. 257 p.
6. Zefirov N.S., Palyulin V.A. Fragmental approach in QSPR // J. Chem. Inf. Comput. Sci. 2002. V. 42. P. 1112–1122.
7. Zefirov N.S., Palyulin V.A. QSAR for boiling points of «small» sulfides. Are the «high-quality structure-property-activity regressions» the real high quality QSAR models? // J. Chem. Inf. Comput. Sci. 2001. V. 41. P. 1022–1027.
8. Skvortsova M.I., Fedyaev K.S., Palyulin V.A., Zefirov N.S. Molecular design of chemical compounds with prescribed properties from QSAR models containing the Hosoya index // Internet Electronic J. Mol. Design. 2003. № 2. P. 70–85.
9. Randić M., Pompe M. The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: A comparative study of $^1\chi^f$ against descriptors of CODESSA // J. Chem. Inform. Comput. Sci. 2001. V. 41. P. 631–638.