Разработка универсальной базы данных структуры и свойств химических соединений для построения моделей «структура-свойство» на основе эвристических алгоритмов

Н.В. Звягинцев, Р.Н. Гордеев, А.В. Бурилин

Введение

Интенсивное развитие химии привело к значительному увеличению количества информации о структуре и свойствах химических соединений. На основе данной информации активно разрабатываются модели зависимости между различными свойствами, а также модели «структура-свойство», упрощающие поиск соединений с заранее заданными свойствами. При построении подобных моделей активно применяются эвристические алгоритмы (нейронные сети, генетические алгоритмы, нечеткая логика) [1-6]. Эффективность подобных методов исследования существенно зависит от объема и полноты информации, поэтому наличие универсальной БД (базе данных) существенно упростит разработку подобных моделей.

Сложность разработки универсальной БД структуры и свойств химических соединений в том, что данные поступают от различных источников и актуальны только при конкретных условиях. Сохраняя данные о структуре и свойствах химических соединений необходимо указывать единицы измерений, точность прибора, условия, при которых данные получены и т.д.

Универсальность БД подразумевает следующие характеристики:

1. Возможность наполнять БД данными различных экспериментов (в том числе расчетными данными), не дорабатывая ее структуру. Для этого требуется разработка словарей с типами экспериментов и отдельные таблицы для хранения условий проведения экспериментов (в случае расчетных исследований условиями экспериментов являются физическое приближение и параметры модели).

2. Возможность заполнять БД данными о химической структуре с различной детализацией.
3. Возможность хранить в БД как информацию индивидуальных веществ, так и информацию о свойствах смесей веществ.

Построение моделей «структура-свойство» на основе нейронных сетей

При построении моделей «структура-свойство» чаще всего используются топологическим дескрипторами. Топологический дескриптор – это некоторая характеристика химической структуры, которая по замыслу исследователя должна влиять на наличие определенного свойства у соединения. Так, в качестве топологического дескриптора можно использовать количество определенных функциональных групп в молекуле.

Кроме топологических дескрипторов, в качестве характеристик химического строения можно использовать данные квантовохимических исследований (энергии граничных молекулярных орбиталей, частичные электрические заряды на атомах, дипольные/мультипольные моменты), физико-химические дескрипторы (например, липофильность органических соединений), информацию о молекулярных полях.

При изучении зависимостей «структура-свойство» нейросетевыми алгоритмами на первом этапе выбирается набор структурных дескрипторов, которые по гипотезе могут влиять на появление определенных свойств химических соединений.

Далее для обучения сети передается вектор со значениями структурных дескрипторов, сформированный для каждого соединения из обучающей выборки.

После обучения нейронной сети точность прогноза проявляется на контрольной выборке, данные о которой не участвовали в обучении нейронной сети.

Хранение в БД информации о структуре химического соединения

Структура химических соединений является наиболее значимой характеристикой при построении моделей «структура-свойство». Информация о структуре химического соединения имеет различную степень детализации:

1. Наименее описательной является брутто-формула, отражающая только элементный состав химического соединения. Таким образом, брутто-формула задает состав соединения в виде пар $\{A_i, n_i\}$, где A_i – конкретный тип атома, n_i –количество атомов в молекуле.
2. Информация о топологии химического соединения, которая отражает последовательность атомов в химическом соединении. ПО сути, топология химического соединения описывается графом [7-9], в котором атомы являются вершинами, а химические связи – ребрами. Также часто топологию химических соединений кодируют посредством матриц смежности, которые задаются в обычном виде:

$$A_{i,j,i\neq j} = \begin{cases} 0, \textit{если атомы связана} \\ 1, \textit{иначе} \end{cases} \quad (1)$$

3. Пространственная структура, учитывающая элементы симметрии структуры химического соединения.
4. Информация о геометрических параметрах соединений (длины связей, валентные и торсионные углы), задаваемые векторами $\{A_i, j, r_i, \alpha_i, \theta_i\}$, где A_i – конкретный тип атома, j – номер атома, относительно которого указываются координаты данного атома, r_i, α_i, θ_i - соответственно расстояние между атомами i и j, валентный и торсионный угол. Данная запись координат называется *Z*-матрицей. В ряде случаев используют декартовы координаты всех атомов. Информацию о геометрическом строении можно получить только в результате рентгеноструктурных исследований или компьютерного моделирования.

Также в исследованиях часто используется информация об электронной структуре химического соединения.

В УБДССХС (универсальной базе данных структуры и свойств химических соединений) ключевое значение имеет таблица «структура» (STRUCT), которая содержит информацию о количестве атомов, суммарном электрическом заряде, систематическом наименовании, приписываемом данной структуре и уникальный идентификатор данной структуры в БД. Именно он является внешним ключом для таблиц, отображающих топологию (STRUCT_TOPOLOGY) и геометрическое строение (STRUCT_COORD – декартовы координаты, и STRUCT_Z_MATR – *Z*-матрица структуры).

Таблица STRUCT_TOPOLOGY хранит данные в виде векторов $\{A_i, A_j, t\}$, где A_i, A_j – типы атомов *i* и *j*, связанных взаимодействием типа *t*. Чаще всего, подразумевается химическое взаимодействие. Типы взаимодействий определяются в словаре BOND_TYPE.

Информация об элементном составе химических соединений может быть получена из таблиц STRUCT_TOPOLOGY, STRUCT_COORD и STRUCT_Z_MATR, поэтому отдельно в БД не кодируется.

Информация о типах атомов храниться в словаре ATOM_TYPE, отражает заряд *Z*-ядра и массу ядра *m*, что позволяет различать информацию об изотопах.

Хранение в БД информации о свойствах отдельных атомов

Информация о свойствах отдельных атомов, входящих в состав химических структур, может быть разнообразной. В качестве свойств отдельных атомов могут быть указаны различные параметры: предполагаемый электрический заряд, валентность, спиновое состояние и т.д. Данная информация часто используется при формировании структурных дескрипторов и имеет большое значение при построении моделей «структура-свойство».

Как показано на рис. 1, для свойств атомов организован словарь atom_property_list, в котором указаны доступные данные и их размерность. Сами значения хранятся в таблице atom_data, которая по внешнему ключу связаны с конкретным атомов, входящим в состав химической структуры.

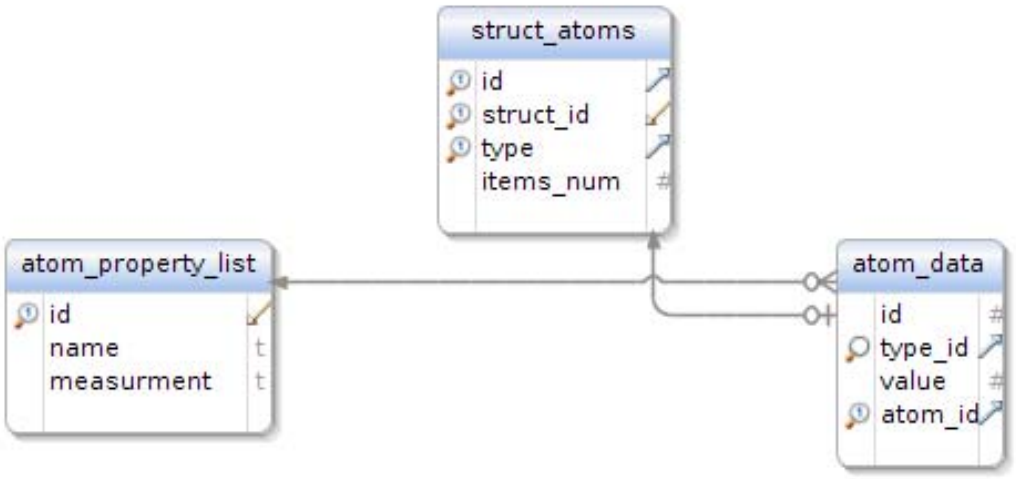

Рис. 1. - Хранение информации о свойствах атомов в БД

Хранение в БД информации о свойствах химических соединений

Каждая химическая структура обладает определенным свойством P, задаваемым в БД векторами $\{S_i, P_i, sp_i\}$, где S_i – идентификатор структуры, P_i – значение свойства, и sp_i – тип хранимого свойства, который содержит информацию о размерности, методе и точности получения данных, а также об условиях, при которых свойство актуально. На рис. 2 показано, как реализовано хранение информации свойствах химических структур в БД.

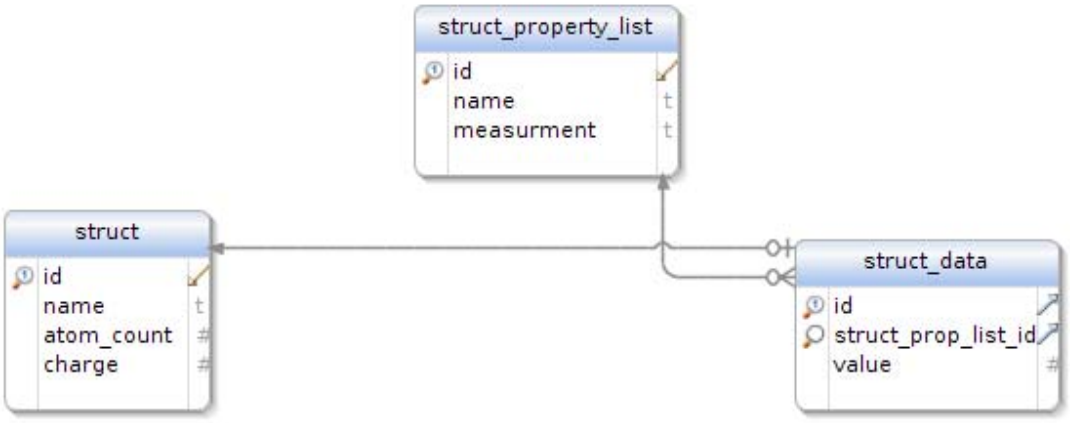

Рис. 2. – Хранение информации о свойствах химических соединений в БД

Так, типы данных, которые могут присваиваться структурам, хранятся в словаре struct_proprty_list. Свойство P структуры S хранится в таблице struct_data и связано со struct по внешнему ключу.

Большинство экспериментальных данных по свойствам химических соединений относятся к смесям различных структур S_i. Как правила, свойства структур определяются в результате компьютерного моделирования. Поэтому для накопления экспериментальных данных следует ввести понятие состава вещества как набора пар $C_i = \{S_i, c_i\}$, где S_i – структура в составе, c_i – доля структуры в составе ($0 < c_i \leq 1$). В БД состав храниться в таблице composition.

Различные химические изомеры также фиксируются в БД в виде вектора из уникальных идентификаторов химических структур: $I_j = \{S_1, \ldots, S_i, \ldots\}$ и кода типа изомерии.

Хранение в БД спектроскопических данных

Спектроскопические данные также могут быть использованы в при построении моделей «структура-свойство» [10-13]. Спектроскопические данные имеют большое значение при исследовании химических соединений

и могут активно применяться при построении моделей «структура-свойство». Спектроскопические данные хранятся в виде вектора $\{\nu_1,\ldots,\nu_i,\ldots\}$.

Заключение

Предложенная структура БД дает возможность хранить информацию о строении химических соединений с различной степенью детализации. Такой подход позволяет упростить формирование структурных дескрипторов, используемых при построении моделей «структура-свойство» на основе генетических алгоритмов.

Также структура БД позволяет обрабатывать эвристическими алгоритмами (в частности, нейросетевыми) информацию о смесях химических соединений и избегать дублирования данных для различного рода изомеров.

Организация хранения свойств смесей химических соединений, чистых химических соединений и отдельных атомов по принципу «атрибут-значение» придает БД определенную универсальность. Таким образом, не меняя структуры БД, можно вводить новое свойство химического соединения.

Литература

1. Григорьев, В. Ю. Количественные модели «структура – свойство» органических соединений [Текст]: дис. д. х. наук: 02.00.03, 02.00.04: защищена 22.01.10 : утв. 15.07.10 / Григорьев Вениамин Юрьевич – Черноголовка, 2013. – 324 с. – Библиогр.: С. 280–324.

2. Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биологических исследованиях [Текст]: // Вестник Московского университета. Серия 2, Химия, 1999, Т. 40, №5.

3. Попок, Н.И., Пята М.В. использование нейронных сетей и нечеткой логики для прогнозирования физико-химических свойств материалов [Электронный ресурс]: // Ползуновский вестник, 2008, № 1 – Режим доступа:

http://elib.altstu.ru/elib/books/Files/pv2008_0102/pdf/055popok.pdf (доступ свободный) – Загл. с экрана. – Яз. рус.

4. R. D. Cramer, D. E. Patterson, J. D. Bunce (1988). «Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins». J. Am. Chem. Soc. 110 (18): 5959-5967

5. Галушка В.В., Молчанов А.А., Фатхи А.А. Применение многослойных радиально-базисных нейронных сетей для верификации реляционных баз данных [Электронный ресурс] // «Инженерный вестник Дона», 2012, №1. – Режим доступа: http://ivdon.ru/magazine/archive/n1y2012/686 (доступ свободный) – Загл. с экрана. – Яз. рус.

6. Галушка В.В., Фатхи В.А. Формирование обучающей выборки при использовании искусственных нейронных сетей в задачах поиска ошибок баз данных // «Инженерный вестник Дона», 2013, №2. – Режим доступа: http://www.ivdon.ru/magazine/archive/n2y2013/1597 (доступ свободный) – Загл. с экрана. – Яз. рус.

7. Bertoline, G. R., Wiebe, E. N., Miller, C., Mohler, J. L. Technical graphics communications (2nd ed.). New York, NY: McGraw-Hill. 1997

8. Molodtsov S.G. Generation of Molecular Graphs with a Given Set of Nonoverlapping Fragments // MATCH 1994. - v. 30. - P. 203-212.

9. Molodtsov S.G. Computer-Aided Generation of Molecular Graphs // Ibid. - P. 213-224.

10. Литвиненко В.И., Кругленко В.П., Повстяной М.В. применение транспонированной регрессии в задаче предсказания свойств лазерных красителей класса имитринов [Электронный ресурс] // Труды Одесского политехнического университета, 2003, вып. 1(19) – Режим доступа: http://archive.nbuv.gov.ua/portal/natural/popu/2003_1/5/5-7.pdf (доступ свободный) – Загл. с экрана. – Яз. рус.

11. Funatsu K., Nobuyoshi M., Sasaki S.-I. Futher Development of Structure Generation in Automated Structure Elucidation System CHEMICS // J. Chem. Inf. Comput. Sci. 1987. - Vol. 28. - P. 18-28.

12. Funatsu K., Susuta Y., Sasaki S.-I. Application of Infrared Data Analysis Based on Symbolic Logic in Automated Structure Elucidation by SHEMICS //Anal. Chim. Acta. 1989. - Vol. 220. - P. 155-169.

13. Curry B. An Expert System for Organic Structure Determination // ACS Symp. Ser. 1986. - Vol. 306. - P. 350-364.