УДК 547.518

ВЫБОР ОПТИМАЛЬНОГО ОПИСАНИЯ СТРУКТУРЫ МОЛЕКУЛЫ В ЗАДАЧЕ СТРУКТУРА–АКТИВНОСТЬ ДЛЯ ЗАДАННОЙ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ

С.С. Григорьева, В.Т. Чичуа, Д.А. Деветьяров, М.И. Кумсков

(Механико-математический факультет, кафедра вычислительной математики; e-mail: ss_grigoreva@mail.ru)

В работе рассмотрено применение специального алгоритма решения задачи «структура–активность» для молекул амбровых одорантов.

Задача поиска зависимости между структурами химических соединений и их свойствами продолжает оставаться весьма актуальной. В настоящей работе описано применение специального алгоритма к решению этой задачи на примере выборки молекул амбровых одорантов (низкомолекулярных соединений, обладающих амбровым запахом). Особенность использованного в данной работе подхода к решению задачи «структура–активность» состоит в том, что прогнозирование определенных (заданных) свойств молекулы основывается не на исходных молекулярных графах, а на производных из них трехмерных графах, в вершинах которых располагаются не атомы молекул, а так называемые ***особые точки*** (ОТ) [1, 2]. В результате объектом, подлежащим анализу и классификации, становится пространственный граф, вершины которого располагаются в ОТ на молекулярной поверхности. Только затем по полученному трехмерному меченому молекулярному графу вычисляются значения структурных 3D-дескрипторов, по которым в дальнейшем можно делать QSAR-прогнозы.

Указанная молекулярная поверхность создается на основе ван-дер-ваальсовых радиусов атомов. Вокруг каждого атома молекулы строится шар данного радиуса и рассматривается объединение построенных шаров. Полученная область является основой молекулярной поверхности, на которой после «сглаживания» выделяются особые точки. Их определяют как геометрические локальные экстремумы поверхности (в терминах ближайших и наиболее отдаленных от определенных групп атомов точек на поверхности) или как физико-химические экстремумы. Для каждой молекулы вычисляется набор особых точек, представленных своими координатами, геометрическим типом и потенциалом (рис. 1).

Каждой особой точке присваивается символьная метка (маркер). Маркировка определяется параметрами описания, оптимизируемыми в задаче, и мо-

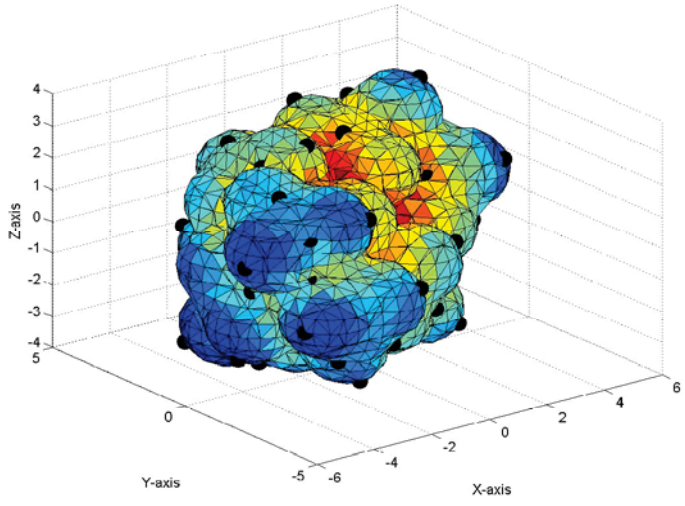
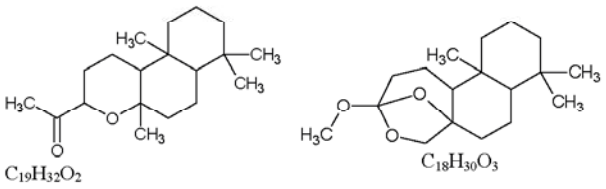


Рис. 1. Особые точки на молекулярной поверхности

Схема



Примеры структурных формул соединений выборки амбровых одорантов

жет вычисляться как по геометрическому типу, так и по значению электростатического потенциала на поверхности. Значение электростатического потенциала вычисляется по формуле

$$\phi = \Sigma(Q_i/(4\ \pi\varepsilon\varepsilon' R_i),$$

где R_i – расстояние от атома до ОТ, Q_i – заряд на атоме. Рассматривается множество всех полученных значений электростатического потенциала, и интервал, содержащий это множество, делится на несколько *интервалов разбиения*. Выделяются три интервала: интервал «близких к нулю» значений (абсолютное значение потенциала меньше заданного *порогового значения*) и интервалы положительных и отрицательных значений (абсолютное значение потенциала больше порогового значения).

В результате каждая ОТ получает две характеристики: 1) локальный максимум или локальный минимум (2 варианта); 2) интервал значения электростатического потенциала в точке (3 варианта). Комбинациями данных характеристик получаются 6 типов особых точек. Согласно данным типам, особым точкам присваиваются символьные метки. В данной работе в качестве меток использовали числа от 1 до 6.

На следующем этапе необходимо построить семейство дескрипторов, адаптированных к данному свойству (активности), а затем сформировать матрицу «структура–дескриптор». Известной моделью, используемой в работах по изучению соотношений «структура–биологическая активность», является пространственный треугольник, у которого вершины имеют заданные локальные физико-химические свойства, а стороны треугольника задаются интервалами расстояний. Если существует 3D-конформация молекулы, «содержащая» такой треугольник, то считается, что она будет обладать заданным биологическим свойством. Более сложным вариантом такой модели является пространственная пирамида с заданными свойствами «вершин» и «ребер». Исходя из такой модели мы попытались построить алфавит дескрипторов таким образом, чтобы представить взаимное расположение пар, троек, четверок особых точек молекулярной поверхности.

Далее была сформирована матрица «молекула–дескриптор», строки которой соответствуют молекулам обучающей выборки, а столбцы – дескрипторам. Следует отметить, что кроме структурных 3D-дескрипторов при формировании матрицы можно также использовать ряд скалярных дескрипторов: общие химико-физические характеристики молекул (например, молекулярный вес, объем, рефракция, поверхностное натяжение, плотность, диэлектрическая постоянная, поляризуемость) и классические топологические индексы.

На четвертом этапе на полученной нами матрице X «молекула–признак» размера $N \cdot M$ (N – количество объектов обучающей выборки, M – количество выявленных признаков объектов) и на столбце свойств (в зависимости от того, обладает или нет i-я молекула данным свойством, в столбце на i-й строке стоит соответственно 1 или –1) был «запущен» алгоритм прогнозирования [3]. В качестве последнего мы использовали метод группового учета аргументов (МГУА), позволяющий отбирать существенные для прогноза свойства молекул столбца матрицы (поскольку построенная матрица часто получается очень «широкой», т.е. $M >> N$) [4].

Из-за «неоднородности» пространственных форм молекул обрабатываемой выборки зависимость «структура–активность» в рамках метода МГУА ищется в виде дерева решений таким образом, что исходная выборка молекул при обучении разбивается на группы (кластеры), а затем для каждого из кластеров находят свой классификатор (рис. 2).

Для идентификации кластеров применяется иерархический метод *кластерного анализа* [3]. Важным преимуществом такого метода является возможность отказа от прогноза, если исследуемое соединение «не похоже» на молекулы обучающей выборки. Чтобы оценить прогностическую устойчивость модели, применяется «скользящий контроль» [3] и вычисляется коэффициент множественной корреляции. Этот коэффициент позволяет оценить качество описания выборки в заданной модели классификации, построенной при выбранных параметрах. Варьирование параметров вычисления дескрипторов позволяет выбрать лучшую модель, а соответствующий ей набор дескрипторов называется «оптимальным» для заданного свойства.

Как указывалось выше, рассмотренный алгоритм был опробован нами на выборке из 50 молекул, 37 из которых являлись амбровыми одорантами (обла-

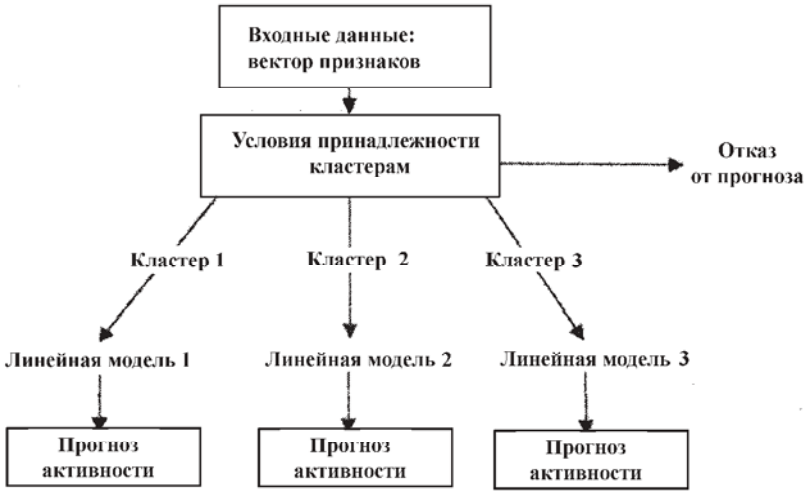


Рис. 2. Схема использования дерева решений при прогнозировании биологической активности

дали запахом) (см. схему), т.е. были биологически активны, а 13 близких по структуре молекул такой активностью не обладали. Для каждого соединения выборки было сделано 3D-описание соответствующего молекулярного графа: перечислены вершины графа (атомы) с дополнительными атрибутами: символом химического элемента, трехмерными координатами в ангстремах и электрическим зарядом. Пример одной из полученных молекулярных поверхностей с ОТ представлен на рис. 1.

После нахождения пар и троек для каждого соединения были получены дескрипторы длинной 703. На сформированной матрице «молекула–признак» был реализован алгоритм МГУА, в результате чего удалось выделить два крупных кластера из 28 и 10 элементов с прогностической оценкой на скользящем контроле 78,6 и 80% соответственно [5]. Это означает, что с помощью предложенного нами алгоритма полученные маркированные особые точки *описывают с довольно высокой точностью активность* в ряду амбровых одорантов.

В дальнейшем при описании молекулярной поверхности в рамках предложенного подхода можно учитывать и другие свойства молекул, например, их липофильность или реакционную способность (способность отдать или принимать электрон или протон), еще более улучшив таким образом прогностическое качество представленной модели.

СПИСОК ЛИТЕРАТУРЫ

1. *Lee B., Richards F.M.* // J. Mol. Biol. 1971. **3**. N 55. P. 379.
2. *Rouvray D.H.* // Computational Chemical Graph Theory. N. Y., 1989.
3. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. М., 1989.
4. *Kumskov M.I., Mityushev D.F.* // Pattern Recognition and Image Analysis. 1996. **6**. P. 497.
5. *Svitanko I.V., Kumskov M.I., Tcheboukov D.E., Dolmat M.S., Zakharov A.M., Ponomareva L.A., Grigor'eva S.S., Chichua V.T.* // Proc. of the 16th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modelling. 2006.

Поступила в редакцию 09.04.07

SELECTION OF OPTIMAL DESCRIPTION OF MOLECULE'S PATTERN FOR GIVEN BIOACTIVITY IN STRUCTURE-PROPERTY PROBLEM

S.S. Grigoryeva, V.T. Chichua, D.A. Devetyarov, M.I. Kumskov

(Department of Mechanics and Mathematics, Division of Computational Mathematics)

Application of the special algorithm QSAR-task solution for the selection of amber odorants is presented in the article.