Информационные процессы, Том 13, № 4, 2013, стр. 290–294

MolDescript - программный пакет вычисления молекулярных дескрипторов[1]

А.А. Белушкин*, О.О. Бочкарева, М.Д. Казанов**, Е.В. Лопатина**, В.С. Мухина**, Г.В. Пономарев**, И.А. Суворова**, Г.Г.Федонин****

**Московский Государственный Университет им М.В.Ломоносова, Москва, Россия*
***Институт Проблем Передачи Информации им А.А.Харкевича, Москва, Россия*

Поступила в редколлегию 03.10.2013

Аннотация—Молекулярные дескрипторы химических соединений представляют собой численные характеристики особенностей их молекулярной структуры. Молекулярные дескрипторы применяют для предсказания свойств химических соединений, что широко используется при разработке лекарственных препаратов. К сожалению, большинство программных продуктов, предназначенных для вычисления молекулярных дескрипторов, являются коммерческими. В тоже время большинство экспертов связывают возможный выход из сложившегося кризиса в области разработки лекарств с открытой коллаборацией между исследовательскими группами, предоставлению в открытый доступ данных и инструментов. В данной работе представлен разработанный нами программный комплекс MolDescript с открытыми исходными кодами, предназначенный для вычисления основных молекулярных дескрипторов химических соединений.

КЛЮЧЕВЫЕ СЛОВА: молекулярные дескрипторы, QSAR, вычислительные методы разработки лекарств.

1. ВВЕДЕНИЕ

На сегодняшний день многими экспертами признается кризисное состояние дел в области исследований по разработке лекарственных препаратов во всем мире [1]. Среди основных причин сложившегося кризиса отмечают высокую стоимость циклов разработки новых лекарств, их большую продолжительность (7–10 лет), а главное, высокий процент неудач на заключительных этапах разработки лекарств. Возможным выходом из кризиса исследователи называют переход на так называемую открытую модель исследований в области разработки лекарственных препаратов, подразумевающую открытый доступ к экспериментальным данным, создание коллаборативных сетей, позволяющих осуществлять тесное сотрудничество между исследовательскими центрами и фармацептическими фирмами, а также создание интернет-приложений и программных комплексов с открытым исходным кодом, применяемых для решения вычислительных задач в данной области [2]. Вычислительные исследования в области поиска и разработки лекарств приобретают все большее значение, однако, большинство существующих программных продуктов являются коммерческими продуктами с закрытыми программными кодами. Последнее относится и к программным продуктам реализующим методы QSAR (Quantitative Structure-Activity Relation) – методы количественных соотношений структура-свойство, имеющих многолетнюю успешную историю применения при разработки лекарств [3]. Несмотря на появление в последнее время общедоступных интернет-приложений, реализующих некоторые из QSAR методов [4], данные приложения невозможно устанавливать

[1] Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение №8135

локально и интегрировать с другими свободно распространяемыми программными пакетами в программные комплексы конвеерного типа (pipelines). Такая задача является по-прежнему актуальной на вычислительном этапе цикла разработки новых потенциальных лекарственных препаратов. В данной работе мы представляем разработанный нами программный комплекс MolDescript с открытыми исходными кодами, предназначенный для решения одной из основных задач в области QSAR исследований – вычисления молекулярных дескрипторов химических соединений.

Методы QSAR имеют широкое применение при разработке лекарственных препаратов и, как правило, используются для прогнозирования свойств и активности химических соединений с известной молекулярной структурой [5]. Известно, что молекулярная структура химического соединения определяет его свойства. Однако, на данный момент не существует методов однозначного определения свойств химических соединений по его структуре. Методы QSAR используют подходы теории машинного обучения для предсказания свойств и активности конкретного химического соединения на основе известных наборов данных для химических соединений со схожей молекулярной структурой. Как правило, такая прогностическая QSAR модель разрабатывается в два этапа. На первом этапе вычисляются так называемые молекулярные дескрипторы химического соединения, выражающие численным образом различные особенности его молекулярной структуры. На втором этапе к полученным молекулярным дескрипторам применяется регрессионная модель, заранее построенная на основе данных о свойствах схожих молекулярных соединений. Большинство существующих программных пакетов вычисления молекулярных дескрипторов химических соединений являются коммерческими программными продуктами [6–8]. Некоторые из данных коммерческих продуктов имеют бесплатные версии в виде интернет-приложений, однако это не позволяет интегрировать данные приложения и программы, реализующие методы машинного обучения, в единый программный комплекс. Существующие свободно распространяемые программные пакеты вычисления молекулярных дескрипторов как правило обладают ограниченной функциональностью. Таким образом, разработка свободно распространяемого программного обеспечения с открытыми исходными кодами для вычисления молекулярных дескрипторов химических соединений на данный момент является актуальной задачей в области вычислительных методов разработки лекарственных препаратов.

Данная статья организована следующим образом. Глава 2 содержит краткое описание групп молекулярных дескрипторов, вычисление которых реализовано в программном комплексе MolDescript. Глава 3 содержит описание полученных результатов и их обсуждение.

2. МОЛЕКУЛЯРНЫЕ ДЕСКРИПТОРЫ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

Молекулярные дескрипторы реализованные в пакете MolDescript можно логически разделить на несколько групп, кратко описываемых далее в этой главе.

2.1. Конституциональные дескрипторы

Данная группа дескрипторов включает в себя основные характеристики химического соединения, отражающие молекулярную композицию соединения. В данную группу не входят дескрипторы, описывающие топологию и молекулярную геометрию структуры соединения. Основными молекулярными дескрипторами данной группы являются: число атомов рассматриваемого химического соединения, число ковалентных связей, число атомов определенного типа, число простых, двойных, тройных и ароматических связей, молекулярный вес, количество функциональных групп и т.д.

2.2. Топологические индексы

Данную группу молекулярных дескрипторов составляют численные величины, описывающие топологические характеристики графа, вершины которого соответствуют атомам рассматриваемого химического соединения, а ребра – химическим связям. Топологические индексы описывает размер, форму, степени симметрии, разветвленности и цикличности графа, соответствующего молекулярной структуре химического соединения. В данную группа молекулярных декскрипторов входят следующие топологические индексы: индекс Винера [9], индексы Загреба [10], Z-индексы Хосои [11] и другие топологические индексы.

2.3. Количество обходов и маршрутов

В отдельную группу молекулярных дескрипторов обычно выделят топологические индексы, соответствующие количествам обходов и маршрутов определенного порядка (длины) в графе, соответствующем молекулярной структуре соединения. Обходом по графу называется определенная последовательность вершин и связывающих их ребер. Маршрутом (путем) в графе называется такой обход, вершины которого не повторяются больше одного раза.

2.4. Индексы связности

Данная группа молекулярных дескрипторов описывает степень связности вершин графа, соответствующего молекулярной структуре соединения. При вычисления данных дескрипторов используются как простые значения степеней вершин графов, так и их модификации – валентные степени вершин или степени вершин, взвешенные согласно типам ковалентных связей. Широко используемыми индексами связности являются индекс связности Рэндика [12] и Киер-Хола [13].

2.5. Индексы информационного содержания

В данной группе молекулярных дескрипторов граф, соответствующий молекулярной структуре соединения, рассматривается в качестве источника вероятностных распределений, к которым могут быть применены элементы теории информации. Молекулярные дескрипторы данной группы представляют собой численные оценки степени структурной однородности молекулярного графа, которое связано с наличием элементов симметрии в графе.

2.6. Дескрипторы автокорреляций

Молекулярные дескрипторы данной группы вычисляются с использованием автокорреляционной функции, применяемой к различным свойствам атомов, таким как, атомный вес, ван дер Ваальсовский объем атома, значений электроотрицательности и поляризуемости. Расстояния между атомами, определяемые структурой химического соединения, могут вычисляться как для двухмерного представления структуры молекулы, так и для трехмерной структуры. Широко используемыми дескрипторами данного типа являются индексы Моро-Брото [14], Морана [15] и Джири [16].

2.7. Спектральные индексы

Молекулярные дескрипторы данной группы основаны на вычислении собственных чисел матриц смежности, расстояний, смежности ребер, валентных расстояний, и других специально определенных матриц молекулярного графа. Собственные числа каждой из перечисленных матриц имеют различный физико-химический смысл. Так, например, сумма собствен-

ных чисел матрицы смежности молекулярного графа описывает степень стабильности молекулы. Примерами часто используемых спектральных индексов являются индексы Ловаж-Пеликана [17] и индексы Балабана [18].

2.8. Геометрические дескрипторы

Данная группа молекулярных дескрипторов описывает характеристики пространственного расположения атомов рассматриваемой молекулярной структуры. Для вычисления геометрических дескрипторов, кроме информации о химической структуре соединения, необходимо иметь информацию о расположении атомов в пространстве. Такая информация может быть получена с помощью методов определения трехмерной структуры химического соединения. Примерами геометрических дескрипторов являются молекулярные степени геометрического расстояния, эксцентричности, геометрические радиус и диаметр, а также 3D-индекс Винера [19].

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В представленном пакете MolDescript на момент написания статьи (версия 0.9) реализовано вычисление 53-х молекулярных дескрипторов химических соединений. Программа доступна для загрузки по адресу http://mg.uncb.iitp.ru/MolDescript/MolDescript.html. По данному адресу также доступны исходные тексты программы. Программа предназначена для локальной установки на операционных системах OS X, Windows версии XP и выше, OS Linux и Unix-подобных системах. На вход программы подается файл описания химической структуры соединения в одном из стандартных форматов – Smiles, SDF, MOL, и др. Вычисление молекулярных дескрипторов может быть выполнено как для полного списка дескрипторов, так и выборочно для отдельных групп или выбранных дескрипторов. Численные значения молекулярных дескрипторов выводятся в стандартный вывод в виде списка с разделителями и могут легко перенаправлены на вход сторонней программы. Таким образом MolDescript может быть легко интегрирован в единый программный комплекс с общедоступными программами, реализующими методы машинного обучения. Такой программный комплекс позволил бы выполнять предсказание различных свойств интересующих химических соединений, включая ингибирование терапевтических белков-мишеней, что является одной из часто встречающихся и актуальных задач в области разработки лекарственных препаратов.

СПИСОК ЛИТЕРАТУРЫ

1. Bunin B.A., Ekins S. Alternative business models for drug discovery. *Drug Discovery Today*, 2011, vol.16, no.15, pp.643-645.

2. Munos B. Can open-source R&D reinvigorate drug research? *Nature Reviews Drug Discovery*, 2006, vol.5, no.9, pp.723-729.

3. Cramer R.D. The inevitable QSAR renaissance. *Journal of Computer-Aided Molecular Design*, 2012, vol.26, no.1, pp.35-38.

4. Sushko I., Novotarskyi S., Korner R., Pandey A.K., Rupp M., Teetz W., Brandmaier S., Abdelaziz A., Prokopenko V.V., Tanchuk V.Y., Todeschini R., Varnek A., Marcou G., Ertl P., Potemkin V., Grishina M., Gasteiger J., Schwab C., Baskin I.I., Palyulin V.A., Radchenko E.V., Welsh W.J., Kholodovych V., Chekmarev D., Cherkasov A., Aires-de-Sousa J., Zhang Q.Y., Bender A., Nigsch F., Patiny L., Williams A., Tkachenko V., Tetko I.V.. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design*, 2011, vol.25, no.6, pp.533-554.

5. Gasteiger J., Engel T. *Chemoinformatics: A Textbook.* WILEY-VCH, 2003.

6. Adriana Code web-page. http://www.molecular-networks.com/products/adrianacode/, 2011.

7. Chemaxon-toolkits and desktop applications for chemoinformatics: calculator Plugins. http://www.chemaxon.com/library/scientific-presentations/calculator-plugins/, 2011.

8. Todeschini R., Consonni V. *Molecular descriptors for chemoinformatics.* Wiley-VCH, New York, 2009.

9. Wiener H. Structural Determination of Paraffin Boiling Points. *Journal of the American Chemical Society*, 1947, vol.69, no.1, pp.17-20.

10. Gutman I., Trinajstic N. Graph theory and molecular orbitals. Total p-electron energy of alternant hydrocarbons. *Chemical Physics Letters*, 1971, vol.17, pp.535-538.

11. Hosoya H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bulletin of the Chemical Society of Japan*, 1971, vol.44, pp.2332-2339.

12. Randic M. On characterization of molecular branching. *Journal of the American Chemical Society*, 1975, vol.97, pp.6609-6615

13. Kier L.B., Hall L.H. An electrotopological-state index for atoms in molecules. *Pharmaceutical Research*, 1990, vol.7, pp.801-807.

14. Moreau G., Broto P. Autocorrelation of molecular structures. Application to SAR studies. *Nouveau Journal de Chimie*, 1980, vol.4, pp.757-764.

15. Moran P.A.P. Notes on continuous stochastic phenomena. *Biometrika*, 1950, vol.37, pp.17-23.

16. Geary R.C. The contiguity ratio and statistical mapping. *Incorp. Statist.*, 1954, vol.5, pp.115-145.

17. Lovasz L., Pelikan J. On the eigenvalue of trees. *Periodica Mathematica Hungarica*, 1973, vol.3, pp.175-182.

18. Balaban A.T. Using real numbers as vertex invariants for third-generation topological indexes. *Journal of Chemical Information and Computer Sciences*, 1992, vo.32, pp.23-28.

19. Mekenyan O., Peitchev D., Bonchev D., Trinajstic N., Bangov I.P. Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. *Arzneimittel Forschung*, 1986, vol.36, pp.176-183.

MolDescript – open-source molecular descriptors calculation software package

A.A. Belushkin, O.O. Bochkareva, M.D. Kazanov, E.V. Lopatina, V.S. Mukhina, G.V. Ponomarev, I.A. Suvorova, G.G. Fedonin

Molecular descriptors of chemical compunds are numerical characteristics of its molecular structure. Molecular descriptors are widely applicable for prediction of compound properties in drug discovery field. Unfortunately, the majority of existing molecular descriptor calculations software are proprietary, while the field itself is moving toward the open source model. In this study we present MolDescript – the open-source software package for calculation of molecular descriptors.

KEYWORDS: molecular descriptors, QSAR, computational methods in drug discovery.