Оптимизация обработки информации в химических базах данных

Т.В. Авдеева[1], А.П. Сергеев[2]

[1] *Кафедра математической физики, НТТУ «КПИ», Победы пр. 37, Киев, Украина*
[2] *ООО «Диалектика», Академика Глушкова пр. 2, корп. 6, к. 214, Киев, Украина*

avdeeva_t1@rembler.ru, a_p_sergeyev@mail.ru

Аннотация. *Предлагается быстрый алгоритм поиска и анализа структурных компонентов химических соединений. Рассматривается его применение при работе с химическими базами данных. Выполнена оценка вычислительной сложности алгоритма и приведены рекомендации по его практическому использованию.*

Ключевые слова

HPC-UA, параллельные вычисления, молекулярный граф, изоморфизм, фармакофорный поиск, молекулярное подобие, виртуальный скрининг, оптический изомер

1 Введение

Химические базы данных применяются не только для хранения и выборки информации, но и позволяют выполнять ряд дополнительных операций по обработке данных:

- выполнять фармакофорный поиск;
- осуществлять поиск соединений из базы данных по молекулярному подобию;
- производить виртуальный скрининг;
- конструировать химические соединения с заранее заданными свойствами.

Все эти операции предъявляют высокие требования к вычислительным ресурсам. Ключ к повышению эффективности обработки информации, находящейся в химических базах данных, - применение универсальных быстрых оптимизирующих алгоритмов, допускающих распараллеливание.

2 Похожие публикации

Быстрый алгоритм поиска и анализа структурных компонентов химических соединений создан на основе быстрого алгоритма идентификации изоморфных XSD-схем [6]. Первая версия этого алгоритма, предназначенная для идентификации подобных химических соединений, рассмотрена в работе [7]. В текущую реализацию алгоритма включен настраиваемый модуль поиска и идентификации компонентов химических соединений. Сформулировано и доказано утверждение об идентичности класса стереоизомеров химических соединений и изоморфных молекулярных графов, описано применение предлагаемого алгоритма для выполнения обработки информации в химических базах данных, приведена оценка вычислительной сложности алгоритма и рекомендации по его применению.

3 Основной раздел

Предлагаемый алгоритм поиска и анализа структурных компонентов химических соединений может применяться для выполнения следующих операций в химических базах данных.

Поиск фармакофоров. Согласно определению ИЮПАК [1], фармакофор – это набор пространственных и электронных признаков, которые являются определяющими совокупность оптимальных супрамолекулярных операций, выполняемых по отношению к той или иной биологической модели. Эти операции могут вызывать или блокировать биологический ответ. Другими словами, фармакофорные признаки – это фармакофорные

центры и определенные диапазоны интервалов между ними, задающие биологическую активность соединения. Обычно в качестве фармакофорных центров выступают анионные и катионные центры, ароматические кольца, акцепторы и доноры водородных связей. Соединения, обладающие одними и теми же фармакофорными признаками, обладают сходной биологической активностью. Благодаря этому открывается путь к синтезу дешевых аналогов (дженериков) оригинальных лекарственных препаратов, которые зачастую являются дорогостоящими.

Для кодирования фармакофорных признаков используются молекулярные графы, представленные матрицами смежности (рис. 1).

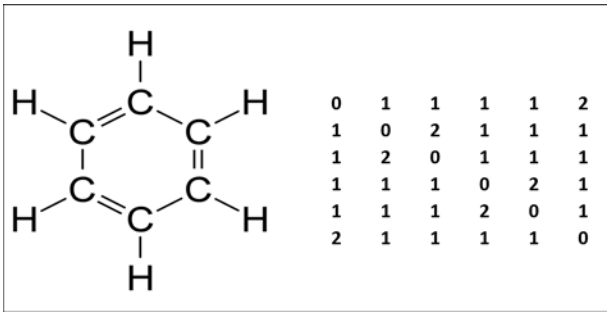

Рис. 1. Молекула бензола и соответствующий молекулярный граф

В результате анализа матрицы смежности химического соединения, находящегося в базе данных (или добавляемого в базу данных), идентифицируются фармакофорные центры. На основе идентифицированных фармакофорных центров строятся классы химических соединений с идентичными фармакофорными центрами, которые обладают сходной биологической активностью. Благодаря оптимизированному быстрому алгоритму поиска с возвращением (с распараллеливанием) существенно сокращается время, затрачиваемое на поиск и выборку информации из химической базы данных.

Поиск химических соединений по молекулярному подобию. Согласно [4], подобные химические соединения обладают подобными свойствами. И чем больше степень подобия, тем более близкими (по химическим и биологическим характеристикам) будут химические соединения.

Свойством молекулярного подобия обладают *пространственные изомеры (стереоизомеры)*, которым присущи одинаковая структура и разное расположение атомов в пространстве. Стереоизомеры относятся к одной из следующих трех категорий:

- оптические изомеры;
- диастереомеры;
- геометрические изомеры.

Наибольшей степенью химического подобия обладают *оптические изомеры* (либо *энантиомеры*) – стереоизомеры, которые являются зеркальным отражением друг друга и не совмещаются в пространстве (свойство *хиральности*). На рис. 2 показана молекула офлоксацина, которая представляет собой комбинацию двух оптических изомеров (L- и R-формы).

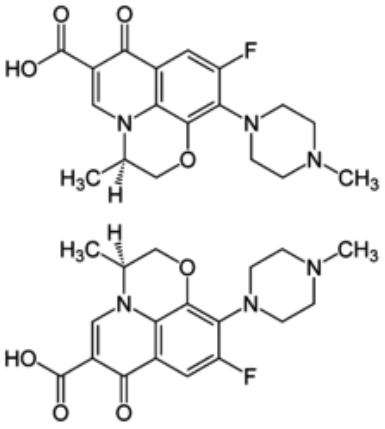

Рис. 2 Молекула офлоксацина представляет собой комбинацию двух оптических изомеров (L- R-энантиомеры)

Утверждение 1. *Оптическим изомерам химических соединений соответствуют изоморфные молекулярные графы*.

В соответствии с определением изоморфизма два молекулярных графа G и H называются изоморфными, если можно построить биективное отношение между множествами вершин графов f: $V_G \rightarrow V_H$, причем если две вершины u и v графа G смежные, будут смежными вершины f(u) и f(v) графа H. Очевидно, что для оптических изомеров также можно построить биективное отношение между атомами, сохраняющее химические связи. Следовательно, оптическим изомерам соответствуют изоморфные молекулярные графы.

Из справедливости этого утверждения следует возможность применения быстрого алгоритма фильтрации изоморфных XSD-схем ([6]) для поиска и фильтрации оптических изомеров химических соединений.

Виртуальный скрининг. В процессе выполнения виртуального скрининга [8] осуществляется автоматизированный просмотр баз данных химических соединений. В результате скрининга выбираются химические соединения, обладающие требуемыми свойствами. Подобная процедура обычно применяется при разработке новых лекарственных препаратов, обладающих требуемой биологической активностью.

Перед выполнением процедуры виртуального скрининга производится настройка блока распознавания химических компонентов. Эта настройка требуется для выборки искомых соединений, молекулярные графы которых обладают заданными свойствами. В процессе настройки конструируется образцовый молекулярный граф, моделирующий свойства искомого химического соединения. Затем выполняется поиск по образцу, в процессе которого создается массив молекулярных графов, идентичных образцовому молекулярному графу.

Конструирование химических соединений с заранее заданными свойствами. Одна из важнейших задач, возникающих в области хемоинформатики, заключается в синтезе химических соединений, обладающих набором заданных свойств [5]. Для выполнения этой задачи можно воспользоваться модулем генератора молекулярных графов, встроенным в алгоритм поиска и анализа компонентов химических соединений.

Описание алгоритма. На рис. 3 приведена блок-схема алгоритма поиска и анализа структурных компонентов химических соединений.

Рис. 3. Блок-схема алгоритма поиска и идентификации структурных компонентов химических соединений

Поступление на вход молекулярных графов. В химических базах данных информация представлена в виде молекулярных графов – связанных неориентированных графов. Количество вершин соответствует числу атомов (либо повторяющихся групп атомов) химического соединения, ребер – числу химических связей. Молекулярные графы представлены в виде матриц смежности.

Определение характеристик молекулярного графа (количество вершин и ребер, сильная связность и т.д.). Этот модуль применяется для вычисления характеристик молекулярных графов, применяемых при выполнении разных операций в химической базе данных. Все расчеты выполняются на основе анализа матриц смежности молекулярных графов. Если планируется поиск стереоизмеров органических соединений, на основе матриц смежности вычисляется система инвариантов изоморфизма [6], применяемых в дальнейшем для быстрого поиска изоморфных молекулярных графов.

Быстрый поиск с возвращением молекулярного графа с заданными свойствами (применяется распараллеливание). С помощью этого модуля выполняется поиск молекулярного графа с заданными характеристиками (свойство изоморфизма по отношению к исходному графу либо набор определенных свойств матрицы смежности). В процессе поиска с возвращением осуществляется распараллеливание, позволяющее ускорить в химической базе данных поиск молекулярных графов с заданными характеристиками.

Генератор молекулярных графов. Назначение этого модуля — генерирование матриц смежности молекулярных графов с заранее заданными свойствами. Используется как автоматическое генерирование матриц смежности, так и генерирование, управляемое пользователем с помощью заданного набора критериев.

Создание массива графов с заданными свойствами. Этот модуль создает массив, в котором сохраняются молекулярные графы, сгенерированные с помощью генератора молекулярных графов.

Вывод результатов. Отображение результатов выполнения алгоритма на экране, сохранение в файле либо вывод на печать.

Практическое применение. Предлагаемый алгоритм может применяться для выполнения следующий операций с информацией, хранящейся в химических базах данных:

- фармакофорный поиск;
- поиск соединений из базы данных по молекулярному подобию;
- выполнение виртуального скрининга;
- конструирование химических соединений с заранее заданными свойствами.

Алгоритм поиска и анализа компонентов химических соединений может применяться для поиска и идентификации соединений с заданными свойствами, выполнения синтеза новых соединений, поиска новых лекарственных форм и выполнения ряда других задач, возникающих при работе с химическими базами данных. При работе с онлайновыми химическими данными потребуется реализация веб-интерфейса, обеспечивающего связь между химической базой данных и алгоритмом.

Оценка вычислительной сложности. Вычислительная сложность комбинированного алгоритма подсчитывается по формуле $E=E_1+\ldots+E_n$, где E_n – вычислительная сложность n-го блока. Вычислительная сложность каждого блока алгоритма приведена в следующем перечне:

- ввод молекулярных графов – $m(n/2)^2$, где n – размерность молекулярного графа, m – количество графов;
- генератор молекулярных графов – $2m(n/2)^2$, где n – размерность молекулярного графа, m – количество графов;
- определение характеристик молекулярного графа (количество вершин и ребер, сильная связность и т.д.) – mn^2, где n – размерность молекулярного графа, m – количество графов;
- быстрый поиск с возвращением молекулярного графа с заданными свойствами (применяется распараллеливание) – $(2mn^n)/p$, где n – размерность молекулярного графа, m – количество графов, p – количество параллельных конвейеров;
- создание массива графов с заданными свойствами – $m(n/2)^2$, где n – размерность молекулярного графа, m – количество графов;
- вывод результатов – $m(n/2)^2$, где n – размерность молекулярного графа, m – количество графов.

Вычислительная сложность алгоритма $E=O(mn^2) + (2mn^n)/p$, где n – максимальное количество вершин молекулярных графов, m – количество молекулярных графов, p – количество параллельных конвейеров в вычислительной системе.

4 Выводы

Предложен быстрый алгоритм поиска и анализа структурных компонентов химических соединений, предназначенный для работы с химическими базами данных. Получена оценка вычислительной сложности алгоритма и приведены рекомендации по практическому применению.

Список литературы

[1] *http://www.iupac.org/*.
[2] Р. Кинг. Химические приложения топологии и теории графов. *М.: Мир*, 1987..
[3] Р. Бейдер. Атомы в молекулах. Квантовая теория. *М.: Мир*, 2001.
[4] A.M. Johnson, G.M. Maggiora. Concepts and Applications of Molecular Similarity. *New York: John Willey & Sons*, 1990.

[5] И.И. Баскин, Е.В. Гордеева, Р.О. Девдариани, Н.С. Зефирова, В.А. Палюлин, М.И. Станкевич. Методология решения обратной задачи в проблеме связи «структура-свойство» для случая топологических индексов». *ДАН СССР 307(3)*: 613-616.
[6] А.П. Сергеев. Быстрый алгоритм фильтрации изоморфных XSD-схем. *Материалы конференции HPC-UA'2011, Киев*, 2011, с. 122-126.
[7] А.П. Сергеев. Быстрый алгоритм идентификации подобных химических соединений. *Материалы конференции PDCS* 2013, Харьков, 13-14 марта, 2013, с. 299-300.
[8] Varnek A., Tropsha A. Chemoinformatics Approaches to Virtual Screening. – *RSCPublishing, 2008*. – ISBN 978-0-85404-144-2.

Optimization of processing of information in chemical databases

Abstract. *The fast algorithm of search and the analysis of structural components of chemical compounds is offered. Its application during the work with chemical databases is considered. The estimation of computing complexity of algorithm is executed and recommendations about its practical use are provided.*