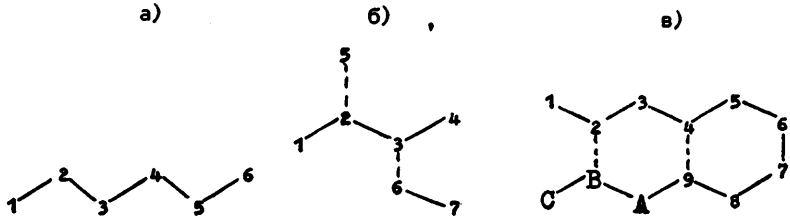
МАТЕМАТИЧЕСКИЕ ИССЛЕДОВАНИЯ В ХИМИЧЕСКОЙ ИНФОРМАТИКЕ
(Вычислительные системы)
1990 год Выпуск 136

УДК 519.17

ОДИН ИЗ СПОСОБОВ ОПИСАНИЯ ХИМИЧЕСКИХ ГРАФОВ С ПОМОЩЬЮ ЦЕПЕЙ

И.И.Строков

Детальное описание структур органических соединений на уровне отдельных атомов или малых групп атомов (мелкоблочное кодирование) является универсальным и удобным для математической обработки. С другой стороны, крупноблочные коды (такие, как линейная запись Висвессера [1]), как правило, более компактны и понятны химику. По ним легче выяснить, из каких частей состоит молекула, и восстановить ее полный вид. Объединить достоинства обоих принципов кодирования позволил бы программно-реализуемый переход между ними. Сложность такого перехода, обусловленная многообразием используемых при кодировании структурных блоков и правил, может быть уменьшена в рамках предлагаемого в данной работе варианта крупноблочного кода, в котором в качестве крупных блоков выступают только линейные участки (цепи) молекул. Использование цепей для описания графов предлагалось и ранее [2] и, с точки зрения химика, является вполне естественным. В самом деле, скелет молекулы можно рассматривать как цепочку вершин (рис.1а,в), либо несколько цепочек (рис.1б), связанных между собой в определенных местах. Для полного описания цепочки достаточно задать ее длину (число входящих в нее вершин, рис.1а). В более сложных случаях следует



код покрытия цепями = 6 412.2536 C.2B49

Рис. 1*)

указать длины цепей, составляющих молекулу, а также те узлы, которые связывают их между собой (рис.1б,в). Таким образом,код химического графа может состоять из двух компонент: описания цепей и описания связей между ними.

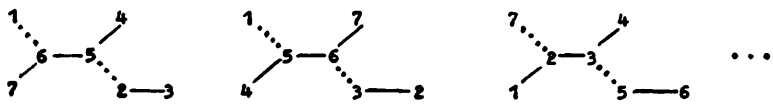
Назовем такой способ представления графа "покрытием цепями" и определим его более строго с помощью терминов "цепь" и "связка". Цепь - это неповторяющаяся последовательность инцидентных друг другу вершин и ребер. Цепь должна начинаться и кончаться вершиной, самая короткая цепь состоит из одной вершины. Покрытие цепями заключается в выборе непересекающихся (т.е. не имеющих общих вершин) цепей таким образом, чтобы в них вошли все вершины графа. Ребра, не вошедшие в цепи, назовем "связками". В такой терминологии покрытие цепями графа представляет собой множество цепей и множество связок (в теории графов под покрытием цепями понимается набор цепей, который включает все ребра и вершины графа, причем вершины,возможно,более одного раза).

---

*) На этом рисунке и далее цепи показаны сплошными линиями, связки - пунктиром. Номера вершин обозначаются одним символом: цифровая нумерация от 1 до 9, затем последовательно буквы латинского алфавита. В коде покрытия цепями числовое значение каждого символа слева от точки определяет длину цепи, справа - пары вершин, связывающие цепи.

Покрытие цепями в принятом определении можно записать коротким и просто дешифруемым кодом, в котором цепи описаны как их длины, а связки - как пары номеров связных вершин. При этом вершины должны быть пронумерованы по порядку вначале в одной цепи с одного до другого конца, затем в другой и т.д. При этом длины цепей в коде должны быть записаны в той же последовательности, в которой цепи выбираются для нумерации в них вершин. Связки, напротив, могут быть перечислены в любом порядке, однако для единообразия мы будем располагать их по старшинству, так чтобы число, отвечающее списку связок, было минимальным.

Можно видеть, что указанные условия допускают запись одного покрытия цепями разными кодами в зависимости от того, с какой цепи и с какого ее конца начать (и продолжить) нумерацию вершин графа. Так, для покрытия цепями на рис.1б можно составить 24 разных кода, часть которых показана на рис.2. Иными словами, код покрытия цепями вырожден. Более того, и само покрытие цепями вырождено, т.е. граф (за исключением отдельных



124.1625 124.1536 421.2735 ...

Рис. 2

симметричных графов) можно представить несколькими разными покрытиями (будем считать, что два покрытия цепями совпадают, если их можно записать одинаковым кодом). Так, кроме покрытия графа, приведенного на рис. 1б, есть еще два (не считая другие, производные от них разбиения, когда две цепи и связка составляют цепь) (рис.3). Нетрудно убедиться, что с учетом трех допу-

стимых покрытий число разных кодов, описывающих в данном случае граф 2,3-диметилпентана, достигает 76.

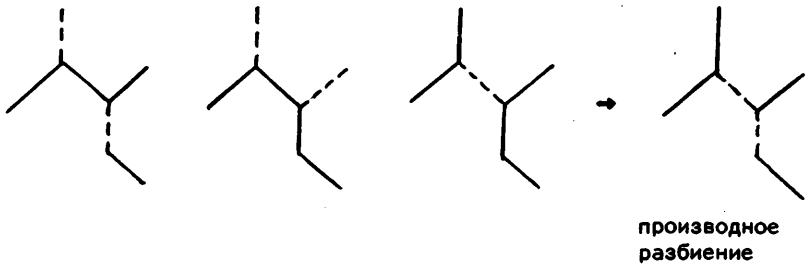


Рис. 3

Таким образом, существует вырожденность покрытий цепями и вырожденность их кодов. С другой стороны, очевидно, что выбор единственного (канонического) кода мог бы облегчить решение ряда задач обработки структурной информации (в том числе поиск заданного графа в базе данных, например, при регистрации химических соединений). Рассмотрим два приема канонизации кодов покрытия цепями.

Первый назовем "условно-каноническим". Суть его состоит в разбиении молекулярного графа на цепи и связки любым, но единообразным способом при условии, что его вершины уже получили каноническую нумерацию по какому-либо алгоритму. При этом изоморфные графы действительно получат одинаковые коды,однако каноничными они будут только в рамках конкретных программ нумерации вершин и разборки графов на цепи. Если воспользоваться другими программами для этой же цели, то код графа может измениться. Простота реализации "условно-канонического" приема обеспечивается,с одной стороны,распространенностью удобных и эффективных программ канонической нумерации, а с другой - возможностью применить несложный алгоритм разбиения на цепи, который может заключаться, например, в выборе цепей таким образом,что-

бы первая и последующие вершины цепи имели по возможности меньшую кратность. Последнее условие, как нетрудно убедиться, для большинства графов органических молекул гарантирует минимальную длину кода, а его идентичность для одинаковых графов обеспечивает предварительная каноническая нумерация вершин.

Более привлекателен, на наш взгляд, другой подход, когда каноническим кодом покрытия цепями считается лексикографически самый младший среди всех возможных для данного графа. Иными словами, если рассматривать последовательность символов кода как число, то канонический код обязан быть минимальным. Такой способ канонизации (назовем его "строго каноническим") задает свойства кода, не ограничивая пути его достижения. Надежный (но длительный) путь поиска строго канонического кода - получение всех кодов для данного графа и выбор наименьшего из них. Можно показать, что минимальный (строго канонический) код имеет минимальное число цепей и связок. Действительно, количество символов в коде, т.е. его длина L, равна n+2b (где n - число цепей а b - число связок). С другой стороны, через n и b выражается число ребер B графа: $B = \sum_{i=1}^{n} (l_i - 1) + b = \sum_{i=1}^{n} l_i - n + b$, где l_i - длина i-й цепи. Поскольку сумма длин цепей l_i равна числу узлов N графа, то $n = b + (N-B)$. Отсюда $L = 3n + 2(B-N) = 3b + (N-B)$. Так как N и B - константы, то минимальному L должны отвечать минимальные n и b.

Опишем кратко используемый нами переборный алгоритм поиска канонического кода покрытия цепями.

Шаг 1: удаление части ребер графа таким образом, чтобы у любой вершины осталось не более двух инцидентных ей ребер. В результате должен получиться граф, состоящий только из цепей или простых циклов. При каждом возвращении на этот шаг из исходного графа удаляется другой, не встретившийся ранее набор ребер, а при исчерпании всех вариантов алгоритм оканчивается.

Шаг 2: подсчет количества цепей, на которые распадается граф после удаления части ребер. Если, кроме цепей, обнаружены циклы либо число цепей больше, чем в одном из прежних разбиений, то следует сразу же вернуться на первый шаг.

Шаг 3: нумерация вершин в цепях в таком порядке, при котором код был бы младшим. Несложно показать, что для разбитого на n цепей графа в общем случае есть $(2^n)*n!$ способов нумерации. В действительности, конечно, нет необходимости получать все коды, так как понятно, что для самых младших нумерация начинается с коротких цепей. Для того, чтобы номера вершин в связках были как можно меньше, для каждой связки требуется свой порядок нумерации цепей. Для выбора оптимальной нумерации каждая связка "голосует" за свой порядок, причем приоритет отдается связкам, которые обеспечивают более младший код. Таким образом, на этом шаге перебор можно исключить. Новый код запоминается, если он младше всех, полученных ранее, затем следует возврат на первый шаг.

Описанный выше алгоритм позволяет кратко и однозначно закодировать граф без учета химической природы вершин. Известно, однако, что структуру одного и того же соединения можно представить несколькими графами. Например, граф бифенила имеет 22 вершины, если ими являются атомы, и только две, если за вершину принять фенил. Подобных неоднозначностей мы избежали, приняв, что вершинами могут быть обычные для органической химии элементы вместе со связанными с ними атомами водорода, например, C, CH, CH_2, CH_3, O, OH и т.д. При этом сам атом водорода не должен выступать в качестве отдельной вершины и ни одна вершина не может быть составлена из других, т.е. вершины типа CO, CS, C_2H_5 и т.д. недопустимы.

Ребра химических графов также можно описать неоднозначно. Например, допустимы структурные формулы нафталина, приведенные на рис.4. Формально эти графы различаются кратностью ре-

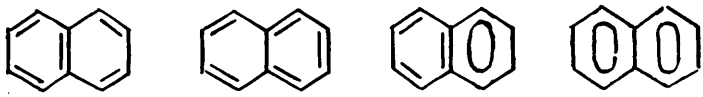


Рис. 4

бер, но отвечают одной молекуле, кратности связей которой определяются гибридизацией атомов углерода. Поэтому мы посчитали целесообразным вместо кратности связей указывать гибридизацию связных вершин. Количество негибридизованных р-электронов, способных образовывать двойные (тройные) связи, задается при описании химической природы вершины вместе с кодом химического элемента и его координационным числом в данном валентном состоянии (рис.5).

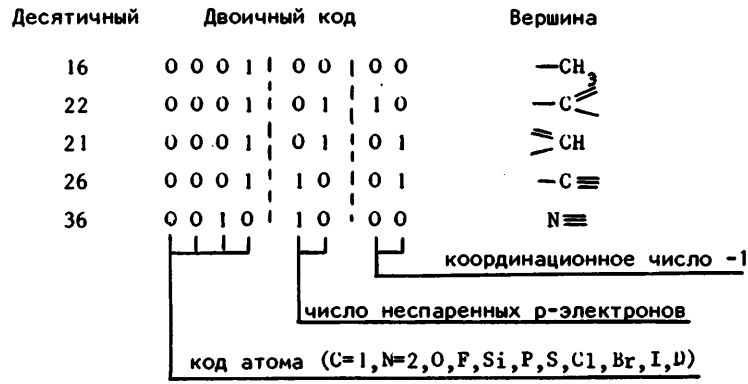


Рис. 5

Количество атомов водорода не указывается явно, но его легко вычислить, если от текущей валентности элемента вычесть его координационное число и число р-электронов. Отметим, что координационное число отражает химическое свойство вершины и может отличаться от ее степени в конкретном графе. В этом случае считается, что вершина имеет свободные связи - столько, насколько отличие.

Предложенные выше правила позволяют закодировать химическую природу вершин и связей молекулы лишь единственным, каноническим образом, правда, за счет потери части информации. Так, зная лишь гибридизацию вершин, нельзя различить резонансные формы в антиароматических π-системах. Но учитывая, что даже очень подробные сведения не исчерпывают всех особенностей строения молекулы, мы остановились на данном способе кодирования, учитывая его компактность и однозначность и оставляя возможность любую дополнительную информацию (например, распреде - ление зарядов, геометрию молекулы и т.д.) в необходимых случаях хранить отдельно.

Структура данных. Оценка применимости предлагаемого метода покрытия цепями проводилась на специально сконструированном макете базы данных. Основу базы составляет файл прямого доступа, в котором хранится вся необходимая информация в виде записей переменной длины. Основной объем файла занимают 10 500 записей об органических молекулах в виде канонического кода по - крытия цепями и описания химической природы вершин (на рис.6

9	21 16 22 21 21 21 22 25 36	1 -8 1 3 1 7
число вершин	типы вершин	цепи связки
		канонический код покрытия цепями

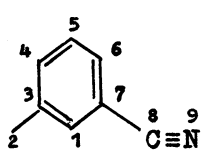


Рис. 6

приведена подробная структура записи). Данные о молекулах по - лучены из машинного каталога системы КОМПАС-МС [3] и зашифро - ваны в виде канонического кода по описанному алгоритму. Дополнительно в базу данных включено около 400 записей совершенно иного характера. К ним относятся: записи органических реакций на языке команд "структурного редактора" (см.ниже), списки номеров записей, обладающих заданным свойством и т.д. При этом общий размер файла пока сравнительно мал - 200 килобайт.

Основной файл позволяет находить требуемую запись по ее текущему номеру. Однако есть возможность приписать любой записи одно или несколько произвольных имен (ключей). Эти имена вместе с отвечающими им номерами записей организованно хранятся в отдельных файлах меньшего размера, что обеспечивает не только поиск записи по имени, но и произвольную организацию этих имен с возможностью их удаления и вставки. В этом смысле основной файл - бесструктурное хранилище информации, а файлы имен задают каждый свою структуру данных.

Программы поиска и обработки информации. К ним относятся программные средства для решения задач, которые наиболее часто возникают при работе с базой данных по структурам и свойствам органических молекул:

1) изображение молекул на экране или на бумаге в привыч - ном для химика виде (визуализация);

2) обратный процесс - ввод человеком в ЭВМ информации о молекулах органических соединений (ввод структур);

3) поиск химических графов с заданным фрагментом (подст - руктурный поиск);

4) поиск пересечения химических графов (максимальный об- щий подграф).

Для этих целей мы используем разработанные нами програм - мы. Отметим кратко их особенности.

Задача 1. Чтобы получить близкое к правдивому изображение, программа визуализации работает с простой математической моделью молекулы, где вершины ведут себя как шарики заданного диаметра, причем связные притягиваются, а остальные отталкиваются. По желанию моделирование может происходить как в плоскости,так и в пространстве. Во избежание ошибок модель строится постепенно, начиная с любой вершины, причем после добавления каждого нового "шарика" модель приводится в равновесие. При этом учитывается взаимодействие всех пар связных вершин и пар несвязных, если они находятся не очень далеко друг от друга.

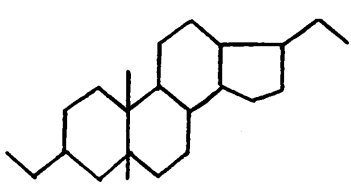
Задача 2. Для ввода в ЭВМ структурных формул обычно применяют программы, позволяющие скопировать рисунок молекулы на экран дисплея, используя вместо карандаша и резинки клавиатуру, световое перо или другие технические средства. Однако известен и другой путь, когда оператору предоставляется набор команд для изменения связности и химической природы вершин,а рисунок молекулы создается автоматически по ее топологии с помощью программ визуализации (см.,например,[4]). Этот подход и реализован в нашей программе. В ней использован принцип "химической клавиатуры", когда каждой клавише отвечает определенный химический фрагмент, который при нажатии клавиши либо образует новую молекулу, либо взаимодействует с введенной ранее по типу реакции замещения, присоединения или циклизации. Причем можно выбрать как тип реакции, так и вершины, по которым произойдет взаимодействие (реакционные центры). Нам кажется, что этот подход психологически близок к восприятию химиком и может быть развит далее с учетом многообразия органических реакций.

Задачи 3,4. В каждой из этих задач требуется проверка на идентичность подграфов в сопоставляемых графах. Упрощение этой процедуры мы нашли в итеративном переходе к реберным графам,по-

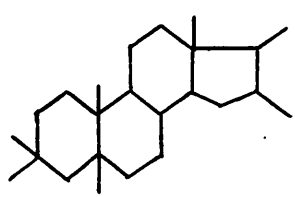
скольку вершине n-реберного графа в исходном графе отвечает подграф из n вершин.

Обсуждение результатов. Накопленный экспериментальный опыт позволяет отметить следующие особенности предложенного способа кодирования молекулярных структур.

1. Небольшие размеры кода в сочетании с простотой его расшифровки. Во многих задачах, связанных с просмотром больших массивов данных, значительное время, пропорциональное размеру записей, тратится на их чтение с внешних носителей и исполь - зование более лаконичных записей позволяет ускорить этот процесс. Приведем статистические данные по размерам канонического кода. Среди 1000 соединений с молекулярными весами от 160 до 200 максимальная длина кода составила 22 байта, минимальная 6, а средняя 12 при среднем количестве вершин 18, т.е. по 0,6 байта на вершину. Заметим, что длина кода напрямую не зависит от числа вершин и ребер графа. Например, для изомеров октана она варьирует от 1 (н-октан $CH_3-(CH_2)_6-CH_3$) до 13 (2,2,3,3-тетраметилбутан, $(CH_3)_3C-C(CH_3)_3$). В общем случае для ациклических молекул отношение длины кода к количеству вершин свидетельствует о разветвленности молекулы.



Канонический код покрытия цепями: 112K.162I3K7BAFENIN
время = 7 секунд



Канонический код покрытия цепями: 11111J.182B3I4K5K7BAFENIN
время = 20 секунд

Рис. 7

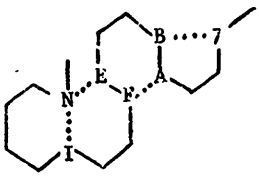


Рис. 8

2. Для соединений,близких по структуре,в кодах покрытия цепями встречаются совпадающие или незначительно различающиеся участки. Например, на рис.7 нетрудно заметить в кодах повтор ...7BAFENIN, который отвечает крупному участку скелета, общему для обоих соединений, приведенному на рис.8.Есть основания полагать, что подобные совпадения неслучайны и в таком случае могут оказаться полезными в решении задач по нахождению общих частей графов. Этот вопрос, однако,нуждается в дальнейшей проработке.

Остается также невыясненным, нельзя ли предложить неперебор-ный алгоритм поиска строго канонического кода покрытия цепями. Время получения этого кода по описанному в данной работе алгоритму приблизительно пропорционально корню третьей степени из числа всех возможных для заданного графа разбиений на цепи. Нетрудно видеть, что каждая вершина степени 3 увеличивает число разбиений в 3 или 2 раза, поэтому длительность поиска канонического кода для крупных и разветвленных молекул может оказаться значительной.

Литература

1. WISWESSER W.J. A line-formula chemical notation - New York: Thomas Y. Crowell Co., 1954.

2. КОЧЕТОВА А.А., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Язык описания структурной информации ОГРА-30 //Машинные методы обнаружения закономерностей,анализа структур и проектирования. - Новосибирск, 1982. - Вып. 92: Вычислительные системы. -С. 70-79.

3. КИРШАНСКИЙ С.П., ЛЕБЕДЕВ К.С., ДЕРЕНДЯЕВ Б.Г. Извлечение структурной информации из масс-спектров с помощью ЭВМ. Система КОМПАС-МС, база данных и принципы организации //Журн. анал. химии. - 1987. - Т. XLII, вып. 6. - С. 1092-1098.

4. FELDMANN R.J., HELLER S.R. An application of interactive graphics - the nested retrieval of chemical structures //J.Chem. Doc. - 1972. -Vol. 12. - P. 48-54.

Поступила в ред.-изд.отд.
2 ноября 1990 года