

В. В. СЕРОВ, М. Е. ЭЛЯШБЕРГ, Л. А. ГРИБОВ

СИСТЕМА МАТЕМАТИЧЕСКОГО СИНТЕЗА И АНАЛИЗА ВСЕХ СТРУКТУРНЫХ ФОРМУЛ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ, СООТВЕТСТВУЮЩИХ ЗАДАННОЙ БРУТТО-ФОРМУЛЕ

(Представлено академиком И. П. Алимариным 6 I 1975)

Известно, что структура огромного числа органических соединений и многие их свойства могут быть объяснены как следствия из системы аксиом, лежащих в основе классической теории строения молекул. Применение этих аксиом к конечной совокупности различных атомов с заданным распределением валентностей создает принципиальную возможность построения всех структурных изомеров, соответствующих определенной брутто-формуле (б.ф.) ^(1, 2).

Нами была предпринята попытка разработать достаточно гибкую систему алгоритмов и программ, позволяющих синтезировать и всесторонне анализировать полный набор структурных формул изомеров, отвечающих данной б.ф. В настоящем сообщении изложены основные принципы построения этой системы.

Исходная б.ф. может включать в себя любые атомы при условии, что их валентность или координационное число фиксированы для каждой конкретной задачи, причем допустимо введение атомов одинаковых элементов, но с разными валентностями. Наряду с отдельными атомами в качестве дискретных единиц структуры могут использоваться также молекулярные фрагменты и функциональные группы. На количество разнородных элементов, входящих в б.ф., никаких ограничений не накладывается за исключением требования, чтобы задача решалась в пределах разумных затрат машинного времени (при быстройдействии ЭВМ 20 тыс. оп./сек одна структура синтезируется за 1—4 сек.).

Для генерации и анализа структур используется аппарат теории графов ^(3, 4). Структурная формула соединения рассматривается как связный конечный мультиграф, представляемый его матрицей смежности. Атомы (или фрагменты) соответствуют вершинам мультиграфа, а связи — его ребрам, кратность которых совпадает с кратностью связей. Программа генерирует все матрицы смежности, удовлетворяющие б.ф. и заданному распределению валентностей единиц структуры. Полученные матрицы затем проверяются на связность. Из всех связных матриц смежности, отвечающих данному изомеру, программа выделяет одну, которая называется канонической.

В системе предусмотрена возможность управления генерацией со стороны специалиста. Это достигается благодаря наличию подпрограмм, которые позволяют в зависимости от поставленной задачи генерировать структуры, принадлежащие к некоторому классу органических соединений или к комбинации классов, а также включать или исключать из рассмотрения изомеры, содержащие определенное число наперед заданных функциональных групп или фрагментов. В частности, можно запретить выдачу изомеров, существование которых в природе представляется маловероятным. Информация о строении запрещенных изомеров предварительно сообщается программе в виде машинного каталога соответствующих фрагментов

(например,



и т. п.). Удовлетворяющая всем условиям

матрица выдается на печать. Необходимо заметить, что в процессе генерации поворотные, оптические и стереоизомеры не различаются, и геометрия молекул в расчет не принимается. Поэтому, например, одна и та же матрица смежности будет соответствовать как цис-, так и транс-изомеру.

В систему включены также программы, позволяющие детально анализировать топологические особенности генерируемых структур, такие как наличие или отсутствие циклов в структурной формуле, размеры цикла, количество ординарных, двойных и тройных связей и последовательность вершин в нем и т. п.

Следует ожидать, что описываемая система может быть достаточно эффективно применена для решения широкого круга исследовательских задач. К их числу относятся, например, задачи следующего типа: 1) установить, какие классы или комбинации классов соединений можно построить из данного набора атомов, а какие — нельзя; 2) «синтезировать» все изомеры из заданного множества фрагментов для фиксированной б.ф. при условии, что суммарный атомный состав всех фрагментов вместе взятых меньше или равен б.ф.; 3) оценить соотношение числа ациклических и циклических изомеров и выяснить зависимость этого соотношения от числа атомов скелета для соединений различного элементного состава и разных классов; 4) осуществить математический синтез структур, обладающих наперед заданными свойствами при определенном атомном составе; 5) установить структуру соединения по молекулярным спектрам и результатам элементного анализа (для некоторых классов соединений подобная задача решалась в работе ⁽²⁾); 6) провести сравнение информационной емкости ⁽³⁾ молекул, принадлежащих к разным классам соединений, но имеющих общую б.ф. и т. п. По-видимому, система может оказаться полезным инструментом для построения гипотетических структур при изучении механизмов реакций, особенно в химии низких температур и космохимии.

Приведем краткое описание основных принципов, положенных в основу алгоритмов программ. Матрица смежности A_r , соответствующая структурной формуле, симметрична, и ее диагональные элементы равны нулю. Выделим в A_r квадратную подматрицу B_r , которая включает в себя только элементы смежности атомов скелета. Подматрице B_r поставим в соответствие число $K_r = \sum_i C_i \sum_j b_{ij} C_j$, $i, j = 1, 2, \dots, n$, где i — номер строки, j — номер столбца, n — число атомов скелета, b_{ij} — значение элемента матрицы B_r (кратность связей), C_i, C_j — весовые коэффициенты строки и столбца. Для весовых коэффициентов должно соблюдаться условие $C_i > C_{i+1} \cdot V$, где V — максимальная наблюдаемая в данных соединениях кратность связей. Величины C_i вычисляются программой по заданному значению C_1 . «Большей» считается та матрица, которой соответствует большее число. Процесс генерации состоит в построении «максимальной» для данной брутто-формулы матрицы B_m и последовательном «вычислении» из нее минимально возможных чисел. Для получения B_m последовательно сверху вниз просматриваются строки матрицы, начиная с элемента $(i, i+1)$. Каждый элемент заполняется максимально возможным по правилам валентности числом, пока сумма всех элементов строки не сравняется с валентностью атома, представляющего строку, или строка не кончится (в этом случае оставшиеся единицы валентности заполняются атомами водорода). «Вычитание» заключается в обратном просмотре строк B_r до первого ненулевого элемента (i, j) , уменьшении (i, j) на 1 и построении «максимальной» матрицы, начиная с элемента $(i, j+1)$.

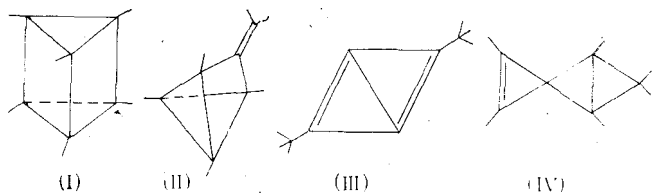
Полученные в процессе генерации матрицы проходят обязательные проверки на связность и каноничность. Алгоритм проверки на связность основан на следующей теореме ⁽³⁾. Пусть L — граф с матрицей смежности A , тогда (i, j) элемент матрицы A^m равен числу маршрутов длины m из вершины i в вершину j . Если при последовательном воздействии матрицы в

степени 2, 3, ..., n (n — размерность матрицы) обнаружится, что имеются хотя бы две вершины, для которых не существует соединяющих их маршрутов, то матрица несвязна.

В качестве канонической выбрана максимальная из возможных матриц смежности, соответствующих заданному графу. Алгоритм анализа на каноничность заключается в проверке, возможна ли такая перенумерация строк матрицы B_r , которая приводит к матрице B_p такой, что $K_p > K_r$. Если это невозможно, то B_r — канонична.

Наличие или отсутствие циклов определяется путем многократного последовательного просмотра (сверху вниз) строк B_r . Если в строке окажется один ненулевой элемент, то его значение и значение симметричного ему элемента приравнивается нулю. Если же число ненулевых элементов больше единицы, то строка остается без изменений. Просмотры строк повторяются до тех пор, пока или B_r станет нулевой (циклов нет), или не окажется строк с одним ненулевым элементом (циклы есть, причем все оставшиеся элементы входят в тот или иной цикл). Алгоритм выделения цикла аналогичен алгоритму выделения цепей данной длины ⁽⁴⁾. По матрице B_r строится «усовершенствованная матрица смежности» B_{rv} , которая указывает не только количество, но и сами ребра, соединяющие данную пару вершин. Матрица B_{rv} последовательно возводится в степени 2, 3, ..., l , ..., n и каждой матрице B_{rv}^l ставится в соответствие матрица D_{rv} такая, что элементу (i, j) B_{rv}^l соответствует подматрица $D_{rv}(i, j)$, которая в мнемонической форме содержит суммы произведений ребер (каждое произведение — маршрут из i в j длины l). Маршруты с повторяющимися ребрами исключаются. Диагональные подматрицы $D_{rv}(i, j)$ содержат циклы длины l . Последовательность вершин в циклах легко восстанавливается по матрицам B_{rv} и B_r .

Система была испытана при решении ряда конкретных задач. Так, машине предлагалось построить все изомеры, относящиеся к классу олефинов, ациклических парафинов и циклопарафинов в отдельности для углеводородов различного состава. Программа выделила требуемые классы, причем интересно отметить, что для состава C_6H_6 наряду с формулами бензола Кекуле, Дьюара и Клауса было выдано множество полициклических структур, например, такие:



Некоторые из этих структур (например, I) уже синтезированы химиками, а возможность существования других представляется вполне реальной ^(6, 7).

Для $C_2H_5NO_2$ обнаружено 84 структурных изомера, из них 48 ациклических. Последнее число в работе ⁽²⁾ было равным 38.

Институт геохимии и аналитической химии
им. В. И. Вернадского
Москва

Поступило
3 XII 1974

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

- ¹ В. В. Разников, В. Л. Тальрозе, Журн. структ. хим., т. 11, 357 (1970). ² Е. А. Feigenbaum, В. G. Buchanan, J. Lederberg, Machine Intelligence, v. 6. Edinburgh. 1971. p. 165. ³ Ф. Харари, Теория графов, М., 1973. ⁴ А. А. Зыков, Теория конечных графов, Новосибирск, 1969. ⁵ Ю. А. Жданов, Теория строения органических соединений, М., 1971. ⁶ L. N. Ferguson, J. Chem. Education, v. 46, № 7 (1969). ⁷ М. Дьюар, Теория молекулярных орбиталей в органической химии, М., 1972.