УДК 681.3.06

Генератор структур системы идентификации органических соединений по их масс-спектрам. Митрофанов Б.П.//Научное приборостроение. Автоматизация научных исследований. Л.: Наука, 1988, с.14

Рассматриваются принципы организации генератора структур, являющегося важной частью системы идентификации структур органических соединений по их масс-спектрам. Быстродействие алгоритма достигается за счет применения метода одновалентных заместителей для построения ациклической части структуры, обладающего большой скоростью и позволяющего проверять вхождение ациклических фрагментов в процессе построения структур. Лит. - 2 назв.,ил. - 1.

Ю.П.Митрофанов (НТО АН СССР)

ГЕНЕРАТОР СТРУКТУР СИСТЕМЫ ИДЕНТИФИКАЦИИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ ПО ИХ МАСС-СПЕКТРАМ

Системы идентификации соединений по спектральной информации, относящиеся к направлению "искусственный интеллект", включают в себя три основных этапа: 1) анализ спектра; 2) генератор структур; 3) отбор наиболее вероятных структур.

Задачей первого этапа является определение списка обязательных и запрещенных структурных фрагментов в анализируемом соединении. В разработанной нами системе идентификации эта задача решается методами распознавания образов на основе статистического анализа специально созданного набора масс-спектральной информации и соответствующих структур органических соединений (обучающей выборки).

Генератор структур строит все возможные с точки зрения правил валентности топологические изомеры, удовлетворяющие брутто-формуле, содержащие обязательные и не содержащие запрещенные структурные фрагменты, определенные на первом этапе.

Третий этап заключается в отборе из построенного генератором структур множества изомеров тех, которые наиболее вероятно соответствуют анализируемому спектру. У нас эта задача решается посредством выявления методами многомерного статистического анализа спектро-структурных корреляций в обучающей выборке и построения на этой основе меры близости "структура - спектр".

Возможности систем идентификации в значительной степени определяются возможностями используемого генератора структур. Основные требования к генератору структур: а) большая скорость построения структур изомеров с заданным элементным составом; б) экономное использование оперативной памяти ЭВМ для того, чтобы иметь возможность строить структуры молекул с большим числом атомов; в) эффективное использование информации о наличии или отсутствии структурных фрагментов в процессе построения структур.

Описываемый генератор структур содержит три органично взаимосвязанных метода построения структурных формул (химических графов): 1) одновалентных заместителей для построения ациклических структур [1] ; 2) перебора связных и каноничных матриц смежности (аналог программы МАИСС в системе РАСТР [2] ; 3) построения структур из суператомов.

Исходными данными для генератора структур являются: брутто-формула анализируемого соединения, набор фрагментов и суператомов, а также диапазон изменения количества π-электронов, который определяет количество кратных связей в строя-

щихся структурах. Суператомами мы называем структурные фрагменты, удовлетворяющие условиям: они не должны пересекаться (иметь общих вершин в строящихся химических графах); их симметрия не должна увеличиваться в процессе построения структур. Структурные фрагменты, не являющиеся суператомами, делятся на два типа. К первому типу мы относим ациклические фрагменты, у которых все свободные валентности сосредоточены на одном атоме. Все остальные фрагменты будем называть фрагментами второго типа. Для каждого фрагмента задается минимальное и максимальное допустимое число его вхождений в структуру. Если максимальное число вхождений равно нулю, то это запрещенный фрагмент.

На первом шаге работы генератора структур матрицы смежности суператомов последовательно заносятся в матрицу смежности строящейся структуры. Таким образом, суператомы хранятся в "раскрытом" виде (в большинстве существующих генераторов структур суператомы считаются в начале одной вершиной и лишь на конечном этапе раскрываются).

Для каждого фрагмента второго типа рассчитывается циклический индекс, т.е. определяется скольким циклам и какой длины принадлежит каждый атом фрагмента. В дальнейшем эта циклическая характеристика используется при проверке построенной структуры на вхождение фрагментов. Использование этой информации позволяет в десятки раз сокращать время работы генератора структур (рисунок).

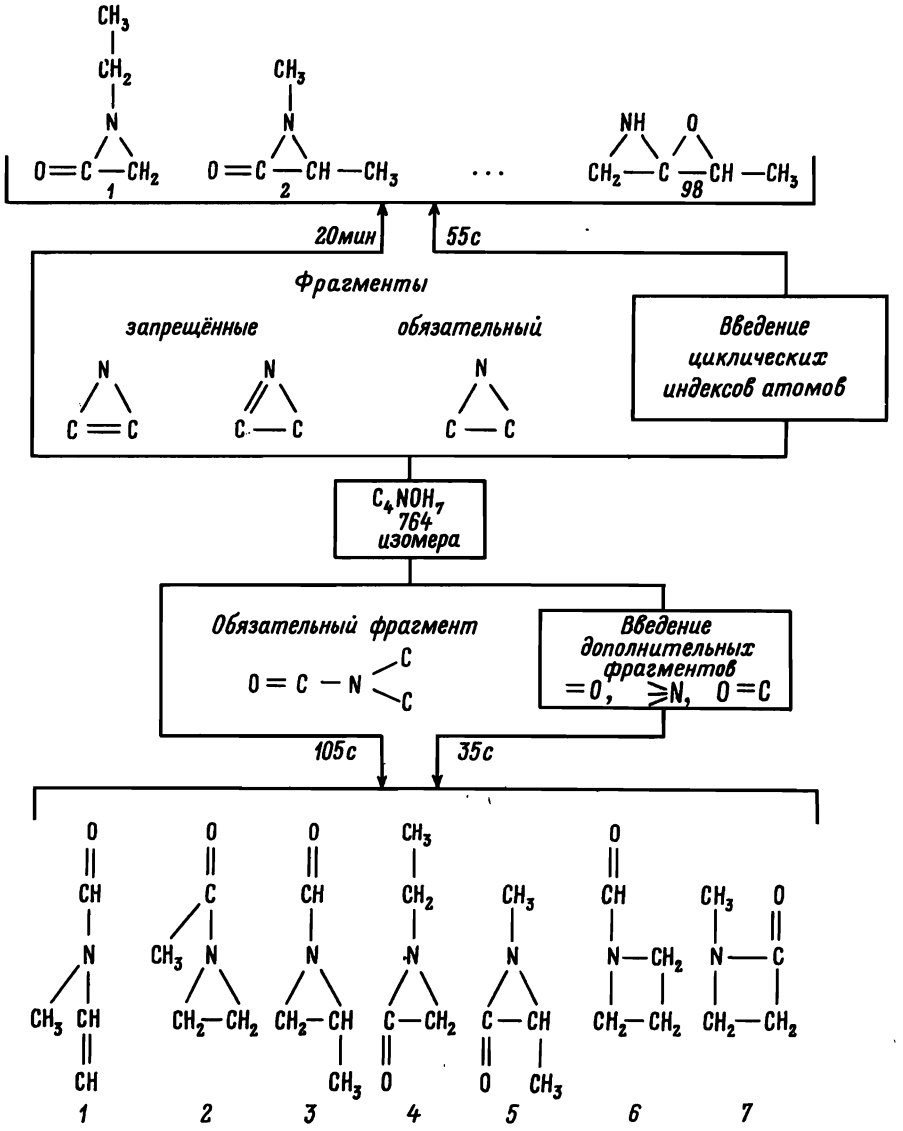
Следующим шагом является определение максимально возможного числа π-электронов по формуле

$$K_M = 2 + \sum_{i=1}^{n} (v_i - 2),$$

где n – общее количество атомов в брутто-формуле, v_i – валентность i-го атома. Следует отметить, что количество π-электронов должно быть четным и неотрицательным. В противном случае брутто-формула не соответствует связной структуре. Если нет информации о диапазоне, в котором может изменяться количество π-электронов, то сначала строятся все структуры с количеством π-электронов, равным K_M. Затем количество π-электронов уменьшается на два и строятся все структуры, имеющие $K_\pi = K_M - 2$ π-электронов. Снова уменьшаем на два количество π-электронов и т.д.до тех пор, пока не построим все структуры с $K_\pi = 0$. На этом генератор структур свою работу заканчивает. Если есть ограничения на количество π-электронов, то строятся структуры только с тем количеством π-электронов, которые удовлетворяют этим ограничениям.

Когда определено количество π-электронов, применяется метод одновалентных заместителей. Этот метод имеет иерархическую многоуровневую структуру. На первом уровне к атомам скелета молекулы присоединяются фиктивные одновалентные заместители – π-электроны, которые служат для образования кратных связей. На втором уровне присоединяются одновалентные атомы из брутто-формулы. На третьем и последующих уровнях присоединяются одновалентные заместители (связные фрагменты с одной свободной валентностью), образовавшиеся на предыдущем уровне. Каждый уровень содержит столько подуровней, сколько различных одновалентных заместителей оказалось на предыдущем уровне.

На каждом уровне присоединения одновалентных заместителей осуществляется проверка частично построенной структуры на вхождение фрагментов первого типа. Это возможно, поскольку для каждого такого фрагмента легко определить уровень, на котором он должен появиться, а суператомы, если они присутствуют, хранятся в раскрытом виде. Например: фрагменты CH_3, NH, OH и т.п.должны появиться на втором уров-



Примеры работы генератора структур при различных способах задания ограничительной информации

не; фрагменты C $(CH_3)_2$, CH(CH_3) и т.п. – на третьем уровне. Введение проверки на вхождение фрагментов первого типа в процессе построения структур позволяет существенно сократить время генерации (рисунок).

При построении ациклических структур на последнем уровне получаются два одновалентных заместителя. Образованием связи между ними заканчивается построение очередной структуры. Для циклических структур на последнем уровне новых одновалентных заместителей не образуется и все остальные связи строятся с помощью алгоритма перебора связных и каноничных матриц смежности.

Алгоритм перебора матриц смежности заключается в построении максимальной матрицы B_M и последовательном "вычитании" из B_M минимально возможных чисел.

Матрицы сравниваются поэлементно слева направо и сверху вниз. Для получения B_M последовательно слева направо и сверху вниз просматриваются строки матрицы, начиная с элемента (i , $i+1$). Каждый элемент заполняется максимально возможным числом, пока сумма элементов строки не станет равной валентности соответствующего атома. Значение элемента (i , j) не должно превышать минимальной из свободных валентностей атомов с номерами i и j . "Вычитание" заключается в обратном просмотре матрицы смежности справа налево и снизу вверх до первого ненулевого элемента b_{ij} , где $i<j$, уменьшении b_{ij} на единицу и построении максимальной матрицы, начиная с элемента (i , $j+1$).

Построенные матрицы проверяются на связность соответствующей структуры и на каноничность. Необходимость проверки матриц смежности на каноничность связана с тем, что соответствие химического графа матрице смежности не является взаимно однозначным. Каждому графу соответствует множество матриц смежности. Любую матрицу из этого множества можно получить из любой другой матрицы, принадлежащей тому же множеству, с помощью некоторой перестановки строк и соответствующих им столбцов. Эта операция соответствует перенумерации вершин химического графа. Для того, чтобы выделить одну матрицу из этого множества, вводится критерий каноничности. В качестве канонической выбирается максимальная матрица из этого множества. Таким образом, проверка построенной матрицы смежности на каноничность заключается в поиске такой перестановки строк и соответствующих столбцов, что в результате получится матрица, которая больше исходной. Если такая перестановка существует, то построенная матрица смежности – не каноническая.

В описываемом генераторе структур с помощью алгоритма перебора матриц смежности строится только та часть матрицы, которая соответствует атомам, у которых после присоединения одновалентных заместителей еще остались свободные валентности или π -электроны. После построения каждой строки подматрицы осуществляется предварительная проверка на каноничность – проверка построенной строки на максимальность. Структуры-дубликаты могут появиться и тогда, когда в процессе построения структур присутствуют суператомы. В этом случае проверка на каноничность заключается в построении только тех перестановок строк и соответствующих столбцов, при которых не изменяются значения элементов, соответствующих связям в суператомах.

Большое значение для быстродействия генератора структур имеет алгоритм проверки наличия в структуре заданного фрагмента, поскольку время, затрачиваемое на эту проверку, может в десятки раз превосходить время, затрачиваемое на построение очередной структуры. Поэтому при проверке химического графа на вхождение фрагмента необходимо эффективно использовать информацию о его структуре.

Пусть имеется мультиграф A , соответствующий проверяемой структуре и представленный его матрицей смежности A . Пусть исходный набор, определяющий матрицу смежности A состоит из N атомов (с присоединенными к ним одновалентными атомами) и $N=\sum_{\kappa=1}^{m} n_\kappa$, где m – число типов атомов (T_1 ,..., T_m), n_κ – число атомов κ -го типа. Предположим, что необходимо установить наличие или отсутствие в A подграфа y , содержащего в вершинах P атомов ($P \le N$), причем $P=\sum_{\kappa=1}^{m} p_\kappa$, где p_κ – число атомов κ -го типа в подграфе y . Пусть подграфу y соответствует матрица смежности G .

Выберем в матрице A те строки и столбцы, которые соответствуют строкам и столбцам матрицы подграфа G . Для этого используем первый критерий соответствия: строка матрицы смежности A соответствует данной строке матрицы G , если для каждого элемента строки матрицы подграфа $g_{T_i T_j}$ в строке матрицы A найдет-

ся элемент $a_{T_iT_j}$ такой, что $a_{T_iT_j} \geq g_{T_iT_j}$. Из выбранных строк и совпадающих с ними по номеру столбцов матрицы A сформируем матрицу A_1. Пусть подматрица A_1 содержит a_i вершин типа T_i. Если хотя бы для одного T_K не выполняется неравенство $a_K \geq P_K$, то подграф $\mathcal{G}$ отсутствует в анализируемом графе. Другими словами, атом структуры соответствует атому фрагмента по первому критерию, если первое окружение атома структуры содержит первое окружение атома фрагмента.

Если отсутствие подграфа $\mathcal{G}$ по этому признаку не обнаружено, то проверяем по первому критерию строки матрицы A_1 на соответствие строкам матрицы G и т.д. до тех пор, пока либо обнаружится отсутствие подграфа, либо на некотором шаге получим подматрицу A_S, у которой каждая строка соответствует какой-либо строке матрицы подграфа. В этом случае для подматрицы A_S рассчитывается циклический индекс, и строки подматрицы A_S проверяются на соответствие строкам матрицы G по второму критерию соответствия, а именно: строка i матрицы A_S соответствует строке j матрицы G, если для каждой длины цикла K, число циклов длины K, которым принадлежит вершина a_i, не меньше числа циклов длины K, которым принадлежит вершина подграфа g_j.

Из выбранных строк и соответствующих им столбцов составляется матрица A_{S+1}. Если отсутствие подграфа по количеству атомов каждого типа не обнаружено и были не соответствующие по второму критерию строки, то матрица A_{S+1} снова проверяется по первому критерию соответствия. Этот процесс продолжается до тех пор, пока либо обнаружится отсутствие подграфа, либо будет построена подматрица A_Z такая, что ее строки соответствуют строкам матрицы G по обоим критериям соответствия.

В этом случае осуществляется последний этап проверки. Пусть матрица A_Z, являющаяся подматрицей матрицы A, содержит d_i вершин типа T_i, d_j вершин типа T_j и т.д. Учитывая, что неравенство $d_K \geq P_K$ выполняется для каждого T_K, составим все сочетания из d_i вершин по ρ_i ($C_i = C_{d_i}^{\rho_i}$), из d_j вершин по ρ_j ($C_j = C_{d_j}^{\rho_j}$) и т.д., после чего из каждого набора C_1, C_2,... будем выбирать по одному сочетанию ρ_1, ρ_2, ... вершин. Каждому сочетанию вершин соответствует L матриц смежности, где $L = \rho_1! \cdot \ldots \cdot \rho_m!$. Эти матрицы поэлементно сравниваются с матрицей G. Таким образом, из матрицы A_Z генерируется $K = C_1 \cdot \ldots \cdot C_m \cdot \rho_1! \cdot \ldots \cdot \rho_m!$ матриц смежности порядка P ($P = \rho_1 + \ldots + \rho_m$), каждая из которых сравнивается с G. Проверки на соответствие строк вводятся для того, чтобы уменьшить число матриц K. Даже при незначительном уменьшении набора $d_1, \ldots, d_m$ число $K = K(d_1, \ldots, d_m)$ существенно уменьшается. При этом значительно сокращается объем необходимых вычислений.

Описанный генератор структур обладает большим быстродействием за счет применения метода одновалентных заместителей для построения ациклической части структуры, а также за счет разнообразного и эффективного учета ограничительной информации. Генератор структур, реализованный на ЭВМ СМ-4, позволяет строить структуры молекул, содержащие до 20-25 атомов в скелете, что является достаточным для решения реальных задач идентификации.

ЛИТЕРАТУРА

1. Митрофанов Ю.П., Разников В.В., Шкуров В.А. // ЖАХ, 1982. Т. XXXVII, вып. 8. С. 1477.

2. Серов В.В., Эляшберг М.Е., Грибов Л.А. Комплекс алгоритмов и программ математического синтеза и анализа структурных формул химических соединений. Деп. в ВИНИТИ, 1975.