ВОПРОСЫ АЛГОРИТМИЧЕСКОГО АНАЛИЗА СТРУКТУРНОЙ ИНФОРМАЦИИ
(Вычислительные системы)
1987 год Выпуск 119

УДК 519.17:5/6

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ АВТОМАТИЧЕСКОГО ИЗОБРАЖЕНИЯ МОЛЕКУЛЯРНЫХ СТРУКТУР ГРАМ-ЕС

Ю.П. Леоненко

В работе описывается первая версия программного комплекса ГРАМ-ЕС (графика молекулярная, ЕС ЭВМ), предназначенного для получения двумерных изображений структурных формул молекул органических соединений или их фрагментов по бескоординатному топологическому коду (матрицам смежности).

По топологическому коду с заданными координатами расположения символов атомных группировок можно построить наглядное изображение молекулы. Если же информация о координатах отсутствует, то возникает проблема воспроизведения привычного химику изображения только на основании сведений о наличии связей между отдельными атомами или их группами.

Известны попытки создания изображений двумерных проекций объемных моделей молекул, которые могут быть получены на основании трудоемких квантово-химических расчетов внутренних координат атомов молекулы (координаты могут быть получены и более простым путем с привлечением методов молекулярной механики). Однако при этом возникают определенные трудности в получении привычных химикам изображений.

Один из подходов построения двумерных структурных формул молекул основан на использовании библиотек заготовок изображений наиболее распространенных циклических фрагментов молекул (шаблонов). При этом подходе возникает проблема компоновки удачных изображений на основе шаблона структурных фрагментов, не входящих в библиотеку.

В [1] описаны идеи построения алгоритма отображения структурных формул в специализированной информационно-поисковой системе

для узкого класса соединений с сопряженными связями. Используется "библиотека" простых колец, из которых осуществляется сборка циклических систем. В [2] предложена программа рисования химических структур, в которой реализованы следующие функции: формирование трехмерного изображения молекулы на основе методов молекулярной механики и выделения ее проекции с наибольшей площадью; нахождение распределения связей вокруг каждого атома при их возможных фиксированных ориентациях, сборка отдельных "рисунков атомов" с учетом ориентаций связей; рисование структуры с изменением длин связей. Программа имеет существенные ограничения, и во многих случаях изображения получаются неудачными либо не могут быть получены вовсе. В [3] приведен алгоритм, который дает результаты, приближающиеся к требованиям CHEMICAL ABSTRACTS SERVICE. Применяются "библиотечный принцип" и библиотека из 17000 циклических шаблонов [4], каждый из шаблонов состоит из одной или более циклических систем. Этот алгоритм дает удовлетворительные результаты для большинства классов соединений. Недостаток его в трудностях создания большой библиотеки циклических шаблонов.

В данной работе также был использован "библиотечный" подход, но в библиотеку включались лишь те шаблоны из тезауруса классов органических соединений [5] и справочника по пестицидам [6] которые соответствовали циклическим частям наиболее часто встречающихся в химических соединениях и были рекомендованы химиками.

Состав библиотеки стандартных изображений циклических частей молекулярных структур комплекса ГРАМ-ЕС (в значительной мере определяющий его возможности) в случае необходимости может быть расширен или изменен, что позволяет адаптировать комплекс к различным условиям эксплуатации.

Состав и функционирование комплекса ГРАМ-ЕС.

Комплекс состоит из программных блоков, структура и взаимодействие которых представлены в виде блок-схемы на рис. I.

Блок I обеспечивает ввод и контроль описаний структурных формул. Для описания химических структур использован язык описания молекулярных графов ОГРА-30-Х [II]. Основой блока I является транслятор с этого языка.

Комплекс не требует специальной аппаратуры для ввода структур, но она может быть использована при соответствующей адаптации. Комплекс ГРАМ-ЕС может работать при вводе информации на языке ОГРА-30-Х с перфокарт или с магнитного диска. Запись на последний может быть осуществлена любым доступным способом.

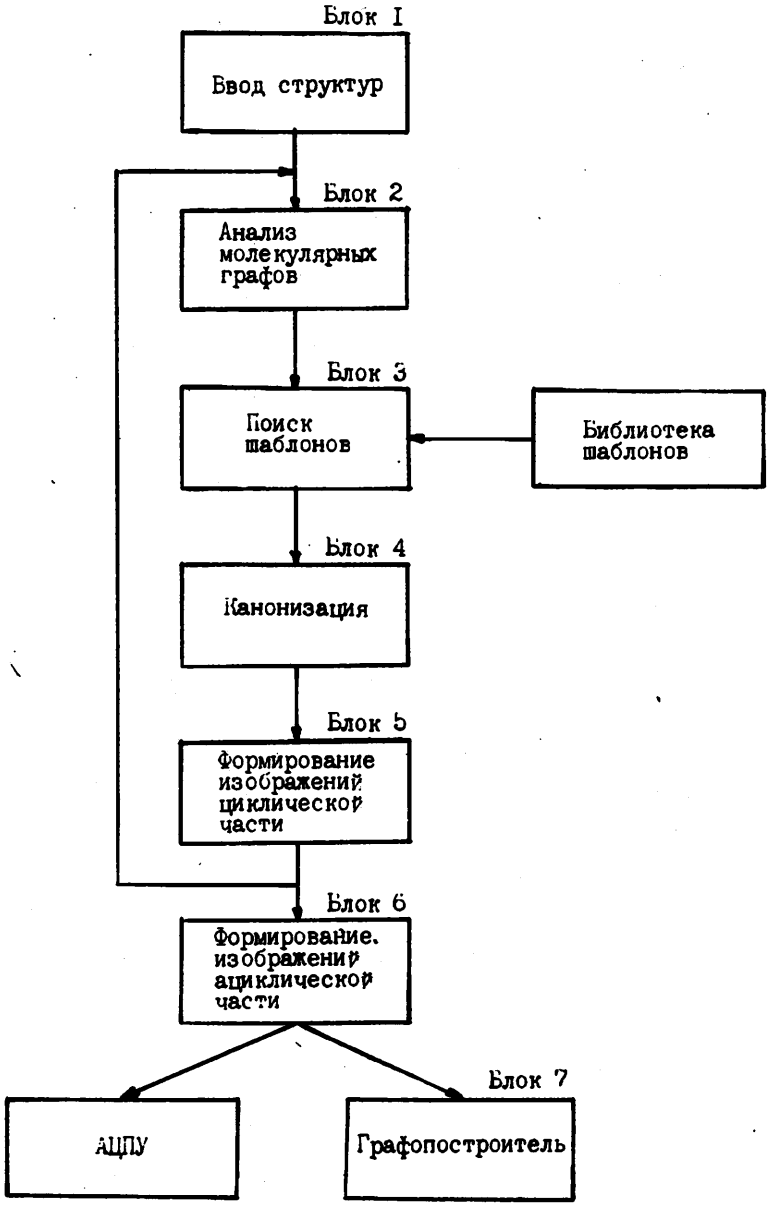


Рис. I. Блок-схема комплекса ГРАФ-ЕС

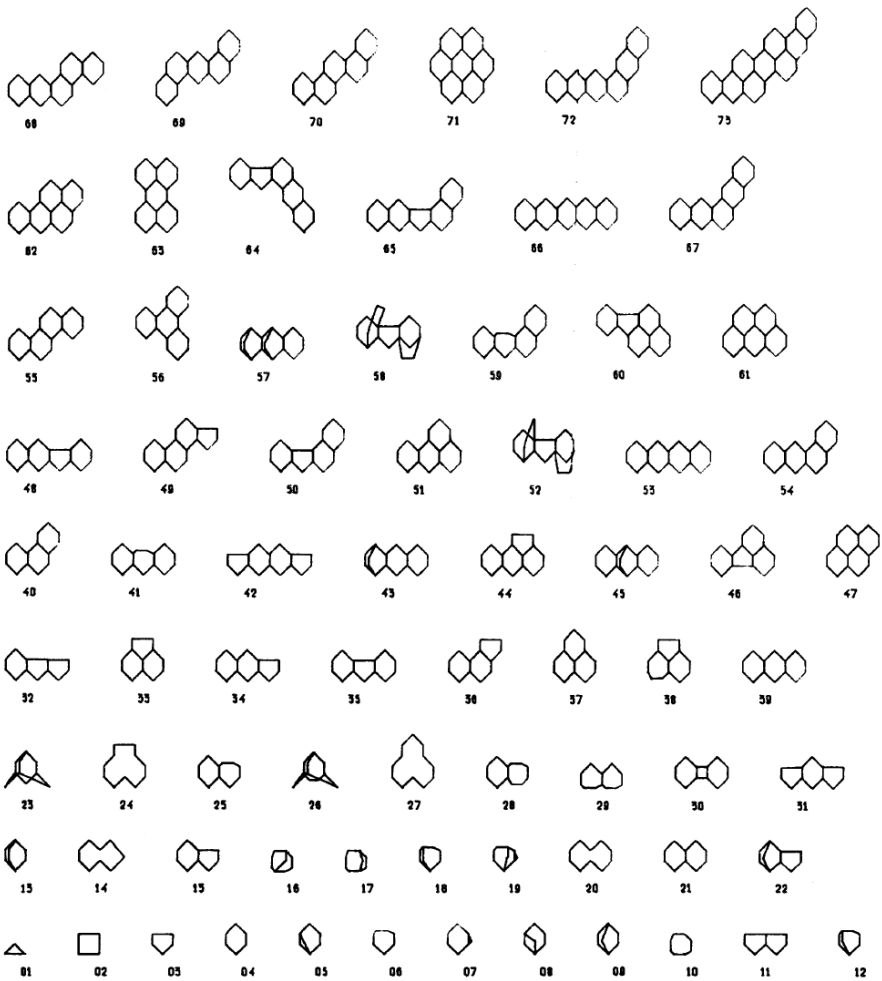


Рис.2. Библиотека стандартных циклических частей (шаблонов)

Блок 2 осуществляет циклический анализ молекулярных графов. Для этого вычисляется значение цикломатического числа молекулярных графов, равное $\nu = q-p+1$, где q - число ребер (без учета кратности), p - число вершин графа. Если $\nu = 0$, то структура ациклическая и обрабатывается в блоке 6; если $\nu > 0$, то в структуре имеются циклы. Вершина, принадлежащая циклу, называется ц и к л и ч е с к о й. Циклическая вершина, смежная хотя бы с одной ациклической, называется г р а н и ч н о й. Подграф молекулярного графа, порожденный множеством циклических вершин, называется его циклической частью. В общем случае циклическая часть может быть несвязна. Компонента циклической части называется циклической компонентой молекулярного графа.

В результате анализа находятся циклические компоненты, множества ациклических и циклических вершин. Для анализа используется алгоритм, основанный на применении относительных разбиений графа [9].

Блок 3 образуется библиотекой шаблонов программ, осуществляющих поиск соответствующих шаблонов для каждой из циклических компонент. Находятся также поисковые признаки, позволяющие отыскать в библиотеке шаблонов соответствующее данной компоненте изображение.

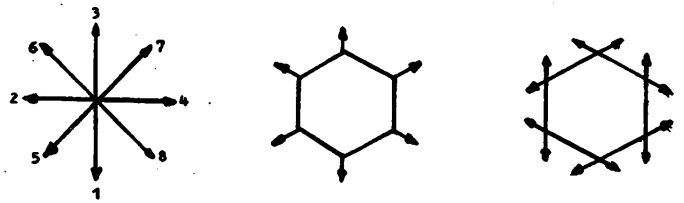
Библиотека состоит из 73 шаблонов (рис.2). Организована она таким образом, чтобы при поиске нужного шаблона число просмотренных библиотечных шаблонов было по возможности наименьшим. С этой целью все элементы библиотеки разбиты на классы по числу вершин в шаблоне. Внутри классов шаблоны характеризуются кодом, однозначно определяющим шаблон. Код шаблонов представляет собой вектор значений метрических характеристик [7] графов, соответствующих шаблонам. Для данной библиотеки в качестве таких характеристик выбраны дистанция и среднее дистанционное отклонение графа. Для связных графов дистанция графа $D(G) = \frac{1}{2}\sum_{v\in V} D(v)$, где дистанция вершины $D(v) = \sum_{u\in V} d(v,u)$, а $\Delta D(G) = \frac{1}{p}\sum_{v\in V} |D(v) - \frac{2}{p}D(G)|$ - среднее дистанционное отклонение графа.

Используемые здесь характеристики графов оказались достаточными для идентификации шаблонов рассматриваемой библиотеки.

При изменении состава библиотеки и ее размеров может возникнуть необходимость в расширении набора характеристик. Для этих це-

лей могут быть использованы другие характеристики, как метрические, так и цепные [12].

С каждым библиотечным шаблоном связаны пять типов массивов: 1) массив кодов шаблонов, 2) массив канонических номеров вершин, 3) массив координат вершин шаблона для АЦПУ, 4) массив координат для изображения на графопостроителе (наличие двух массивов координат шаблонов вызвано тем, что изображение на графопостроителе отличается от изображения на АЦПУ), 5) массив направлений изображения ветвей (ациклических компонент). Направление изображения и их номера показаны на рис.3,а. С каждой вершиной шаблона связано



а) б) в)

Рис. 3

направление, разрешенное для рисования (рис. 3,б), обусловленное алгоритмом работы блока переориентации присоединяемых компонент. От любой вершины можно строить не более двух ветвей (рис.3,в).

В блоке 4 устанавливается соответствие между циклическими компонентами и шаблонами путем установления соответствия между каноническими нумерациями вершин циклической компоненты и соответствующего библиотечного шаблона.

Каноническим видом графа называют граф, матрица смежности которого упорядочена по свойствам вершин, не зависящих от исходной нумерации. Нумерация вершин, соответствующая каноническому виду графа, называется канонической.

Для построения канонической нумерации вершин графа используется упорядоченное орбитальное разбиение множества вершин графа. Порядок на множестве орбит определяется алгоритмом нахождения орбит из [8].

Каноническая нумерация позволяет установить одно из множества однозначных соответствий между номерами вершин из циклических компонент и номерами вершин графов, соответствующих шаблонам. Вершины последних имеют относительные координаты, что позволяет выб-

рать некоторый вариант изображения циклической компоненты. Соответствие, определяемое сравнением канонических нумераций шаблонов и компоненты, позволяет установить координаты вершин шаблона. Естественно, таких эквивалентных соответствий может быть много, и число их зависит от свойств симметрий компонент. Из множества эквивалентных соответствий выбирается любое.

В блоке 5 формируются изображения структур, содержащих более одной циклической компоненты. Формирование начинается с выбора некоторой циклической компоненты, по отношению к которой осуществляется "сборка" изображения всей структуры. Для этого могут быть использованы различные критерии, от их выбора существенно зависит окончательный вид изображения. Для выбора "центральной" циклической компоненты применяется метрический критерий. "Центральная" циклическая часть выбирается с наибольшим числом вершин и граничных точек, причем вершины должны иметь наименьшие эксцентриситеты [7]. Для однозначности выбора при условии равенства значений указанных характеристик выбирается циклическая часть с наименьшим порядковым номером. Найденная циклическая компонента размещается в координатном поле.

От каждой граничной точки "центральной" компоненты с учетом возможных направлений рисования строятся ветви (не более двух).

Так как вершины ветвей, в свою очередь, могут также являться граничными точками, принадлежащими другим циклическим компонентам, то такие вершины рассматриваются как "текущие" граничные точки.

Для очередной "текущей" граничной точки определяется циклическая компонента, которой она принадлежит. От всех вершин этой циклической компоненты строятся ветви по такому же правилу, как и от центральной циклической части. Вершины построенных ветвей и вершины циклической компоненты объединяются в один блок. Поскольку выбранная "текущая" граничная точка имеет номер центральной циклической части, то относительно нее перерассчитываются координаты вершин блока. Построение заканчивается после обработки последней "текущей" точки.

После того, как построены все ветви для очередной циклической компоненты, происходит переход к следующей компоненте.

Блок 6 формирует изображение ациклической структуры или ациклической части структуры.

Находится относительное разбиение графа для произвольной вершины, и выделяются вершины последнего слоя. Затем для каждой вершины из последнего слоя также строится относительное разбиение, в результате находятся цепи, имеющие наибольшую длину [9]. Если таких цепей несколько, то из них выбирается главная цепь, содержащая наибольшее число вершин степени более двух. Если таких цепей окажется несколько, то выбирается та, у которой суммарный атомный вес меток вершин - наибольший. Для вершин, принадлежащих главной цепи, степени больше двух, выполняется построение ответвлений от этой вершины.

Блок 7. Вывод изображения на АЦПУ или графопостроитель. Процедура изображения использует в виде параметров число вершин графа, матрицу смежности, массив меток вершин и их координат.

Для изображения структур химических соединений наряду с АЦПУ (рис.4) используются графопостроитель ЕС-7052 (рис.5) и система математического обеспечения графопостроителя СМОГ ОС ЕС [10]. Из системы СМОГ используются подпрограммы изображения отрезков прямых, символов, чисел, текстов, а также подпрограмма масштабирования.

Комплекс написан на языке ПЛ/I для ЕС ЭВМ, требует 200К байт памяти, состоит из 2500 операторов, содержит 26 процедур,6 из которых вызываются из системы СМОГ. Применяется оверлейная структура описания программ. Изображение на АЦПУ молекулярной структуры с не более чем 50 вершинами на ЭВМ ЕС 1050 получается не более чем за 5 сек, а на графопостроителе ЕС-7052 за 60 сек.

В случае изображения фрагментов химических структур "незанятые" химические связи изображаются в виде "незаконченных" линий.

З а к л ю ч е н и е

В данной работе представлена первая версия комплекса программ построения изображений молекулярных структур по бескоординатному топологическому коду.

Испытания комплекса ГРАМ-ЕС проводились для нескольких сотен химических соединений, взятых из справочников по пестицидам. Для более чем 90% структурных формул соединений были получены легко "читаемые" химиками изображения. В ряде случаев изображения имели самоналожения и пересечения (рис.6), которые могут быть устранены введением в комплекс блоков анализа и коррекции изображений. Комп-

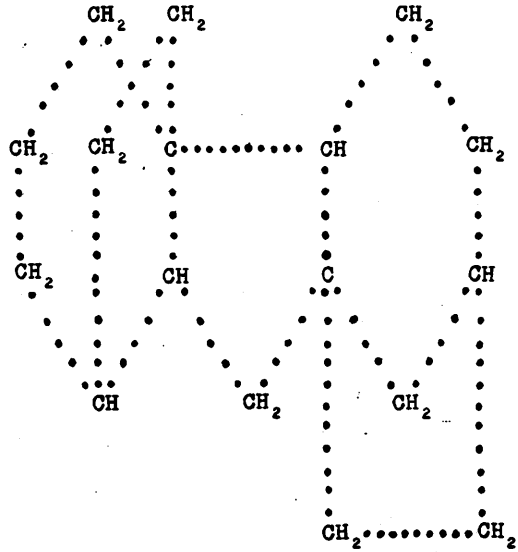


Рис.4. Изображение, выполненное на АЦПУ

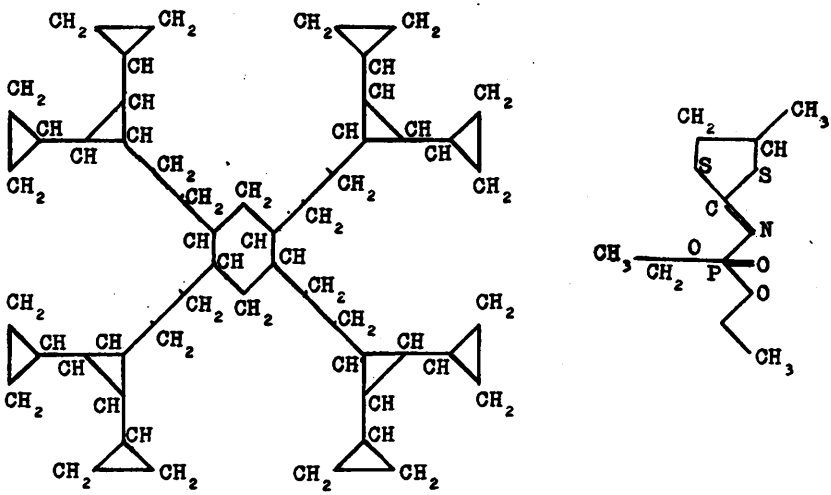


Рис.5. Изображение, выполненное на графопостроителе

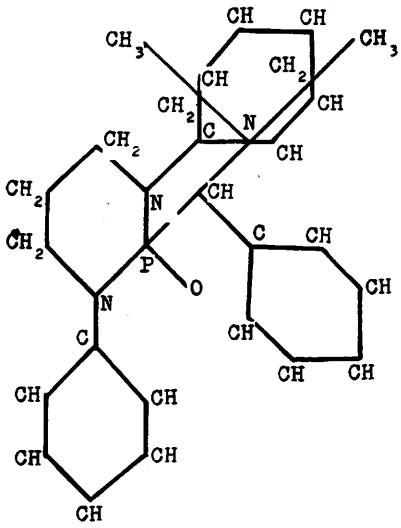


Рис. 6

лекс также прошел испытания и показал свою работоспособность как составная часть в програм - мной системе, предназначенной для прогнозирования биологической активности химических соединений, где он используется для получения изображений структурных формул соединений и их фрагментов.

Следующие важные вопросы построения изображений: автоматическое пополнение библиотеки; анализ и коррекция полученных изображений; вопросы предва - рительного анализа структуры с целью улучшения качества изоб - ражений; улучшение качества изображений за счет использования специальных приемов рисования, специальных символов и стандартов, использующихся при изображении химических структур, и другие предполагается рассмотреть в дальнейшем.

Автор благодарит научного руководителя В.А.Скоробогатова за помощь в данной работе.

Литература

I. ГЕЙВАНДОВ Э.А., ШОЛОХОВ В.Г. Система ввода химической информации в автоматизированной информационно-поисковой системе в виде двумерного изображения структурных формул //Научно-техни - ческая информация. ВИНИТИ. Сер.2. - 1971. - №8.-Вып.2.-С.22-23.

2. CARHART R.E. A model-based approach to the teletype printing of chemical structures//J.Chem.Inform.and Comput.Sci.- 1976. - Vol.16,N 2.- P.82-88.

3. DITTMAR P.G.,MOSKUS J., COUVREUR K.M. An algorithmic computer diagrams// J.Chem.Inform.and Comput.Sci.- 1977. - Vol. 17, N 3. - P.186-192.

4. ZIMMERMAN B.L. Computer - generated chemical structural formulas with standart ring orientations: Thes ... dokt.phyloso - hy.- Univ.of Pensylvania,1971.

5. Тезаурус классов органических соединений /составитель Н.Е.Голубева. - М.: ВИНИТИ, 1981.

6. Справочник по пестицидам /Под ред.Н.Н.Мельникова.-М.: Химия, 1985.

7. СКОРОБОГАТОВ В.А.,ХВОРОСТОВ П.В.Анализ метрических свойств графов //Методы обнаружения закономерностей с помощью ЭВМ. - Новосибирск, 1981. - Вып.91: Вычислительные системы. - С. 1-20.

8. Их же. Методы и алгоритмы анализа симметрий графов //Алгоритмы анализа структурной информации.-Новосибирск,1977.-Вып.103: Вычислительные системы. - С.6-25.

9. СКОРОБОГАТОВ В.А. Относительные разбиения и слои графов //Вопросы обработки информации при проектировании систем. - Новосибирск, 1977. - Вып. 69: Вычислительные системы. - С. 6-9.

10. Математическое обеспечение графопостроителей //Под ред. Ю.А.Кузнецова. - Новосибирск, 1976 (ВЦ СО АН СССР).

11. 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.

12. ДОБРЫНИН А.А., СКОРОБОГАТОВ В.А. Свойства цепей графов и изотопичность //Алгоритмический анализ структурной информации.-Новосибирск, 1985. - Вып. 112: Вычислительные системы. - С.33-45.

Поступила в ред.-изд.отд.
5 февраля 1987 года