

УДК 519.17:681.3:278.141

ПРОГРАММНЫЙ КОМПЛЕКС МЕТАХИМ

А.А.Добрынин, Е.В.Константинова, Л.И.Макаров,
Н.М.Мейрманова, А.А.Палеев, В.А.Скоробогатов .

Пакет программ МЕТАХИМ предназначен для изучения и использования методов алгоритмического анализа структурной информации в химических и граф-теоретических исследованиях, для обучения и повышения квалификации студентов, аспирантов, специалистов. Программные средства пакета ориентированы на различные виды анализа структурной информации, позволяющие в конечном итоге устанавливать закономерные связи между топологическими свойствами структур и их химическими и биологическими свойствами, классифицировать и упорядочивать выборки структур по нужным топологическим и химическим параметрам, выделять группы (таксоны) структурно-подобных соединений.

Для этих целей в пакете предусмотрены широкие функциональные возможности, реализованные в пяти его блоках.

1. Блок ввода структурной информации (блок ДАННЫЕ)

В блоке ДАННЫЕ осуществляется компьютерный ввод молекулярных структур на основе простого линейного языка ОГРА [1], выразительные средства которого позволяют описывать структуры с учетом требований номенклатуры. Можно описывать и обрабатывать не только молекулярные структуры, но и вообще любую структурную информацию, представимую в виде графов.

Имеются возможности работы как с молекулярными скелетами, так и со структурными формулами соединений, представленными с любой степенью детальности.

При работе со структурными формулами молекул предусматривается независимый ввод весов атомов и связей, что обеспечивает многовариантность и многоаспектность анализа.

Имеется возможность работы с химическими соединениями как классического строения, так и неклассического, например, с металлоорганическими и комплексными соединениями, имеющими многоцентровые связи. Использование гиперграфов позволяет моделировать неклассические химические связи и описывать структуры на удобном для компьютерной обработки языке. Допускается также обработка структурной информации, представимой в виде графа с обыкновенными ребрами и гиперребрами.

В перспективе предусматриваются методы обработки структур с топологическими связями, таких как катенаны, узлы, зацепления и сети ДНК.

В блоке ДАННЫЕ также формируются таблицы "объект-признак" и таблицы расстояний между структурами, необходимые при анализе зависимости "структура-свойство" химических соединений, а также при поиске групп подобных соединений. Элементами таблицы "объект-признак" являются признаки (свойства) изучаемых объектов (структур). Объектами могут быть молекулярные графы химических соединений, а признаками - физико-химические характеристики, электронно-топологические и другие параметры. Выбран нужные параметры и заданы значения этих параметров на дисплее, можно сформировать таблицу. Кроме этого, таблицы "объект-признак", а также таблицы расстояний между объектами формируются автоматически в блоке МЕТАН на основе метрических свойств этих объектов. Полученные данные используются в алгоритмах таксономии, представленных в блоке ТАКС.

2. Блок МЕТрического АНАлиза (блок МЕТАН)

Методы метрического анализа [2] предназначены для изучения метрических и цепных свойств молекулярных графов на основе большого числа топологических индексов, как широко известных, так и новых. Конструирование топологических индексов, основанных на расстояниях между элементами молекулярного графа, является перспективным направлением исследований взаимосвязи между структурой и свойствами химических соединений. Методы метрического анализа позволяют исследовать свойства графов, определяемые расстояниями, описывать структурные свойства графа в терминах метрических аналогов, что может быть использовано при построении новых индексов. Метрические свойства в зависимости от рассматриваемого расстояния отражают различные структурные особенности графов.

Блок МЕТАН работает в двух режимах: ИНДЕКСЫ и УПОРЯДОЧЕНИЕ. Работая в первом режиме, можно познакомиться не только с известными топологическими индексами молекулярных структур, но и расширить свои знания в области информационных индексов, топологических спектров, вершинных индексов. Второй режим работы блока МЕТАН предоставляет возможность упорядочить структуры по значениям выбранных индексов.

Молекулярные графы в блоке МЕТАН анализируются на основе двух типов расстояний: обычного и цепного. Обычное расстояние, равное длине кратчайшей по числу ребер простой цепи, соединяющей две вершины, определяет метрические индексы. Цепное расстояние, равное сумме длин всех цепей, соединяющих две вершины, определяет цепные индексы. Второй вид индексов является обобщением первого.

Ниже дана краткая характеристика предлагаемых индексов.

2.1. Интегральные индексы. Интегральные индексы, являясь функциями от параметров графа и расстояний между его вершинами, позволяют характеризовать граф в целом. Предлагаемые индек-

сы образуют два класса: эксцентриситетные и дистанционные [2]. Первый класс индексов дает представление о средних, минимальных и максимальных расстояниях в графе. Второй класс индексов основан на учете всех возможных расстояний в графе. Здесь же представлены топологические индексы: Винера (дистанция графа), Платта, Гордона-Скантлбери, индекс расстояний в графе и т.д. [3,4].

2.2. Информационные индексы. Информационные индексы, учитывая различные свойства молекулярных графов, дают количественную оценку информации, содержащейся в рассматриваемом объекте. Это позволяет сравнивать свойства молекулярных графов в единой количественной шкале. Предлагаемые информационные индексы [5] учитывают следующие свойства молекулярных графов:

- информационный индекс автометричности основан на отношении автометричности вершин - особом виде метрической или цепной эквивалентности вершин;

- информационный индекс дистанционности основан на дистанционной эквивалентности вершин в графе;

- информационный усредненный индекс дистанционности есть информационный эквивалент среднего значения дистанций вершин в графе.

2.3. Топологический спектр. Метрический и цепной спектры (топологические спектры [6]) графа определяют диаграмму, по осям которой откладываются значения вершинных метрических или цепных индексов. Предлагаемый спектр является 2-мерным спектром, функции X и Y которого основаны соответственно на дистанционных и эксцентриситетных индексах графа. Топологические 2-мерные спектры по своему графическому изображению сходны, например, с масс-спектрами, по осям которых откладываются интенсивности и массы ионов. В данной версии МЕТАХИМ спектр представлен в виде таблицы его значений.

2.4. Вершинные индексы. Вершинные индексы отражают метрические или цепные свойства вершин графа. Предлагаемые вершинные индексы являются аналогами интегральных индексов в молекулярном графе.

3. Блок СТруктурного Анализа (блок СТРАН)

Методы структурного анализа позволяют выявлять структурные особенности графов. К классическим задачам анализа структуры графа относятся задачи нахождения мостов, точек сочленения, блоков графа, выделения циклов и т.п. Решение задач анализа циклической структуры и связности молекулярных графов имеет важное значение для приложений в химии и химической информатике (номенклатура органических соединений, кодирование, классификация). Кроме того, структурный анализ позволяет исследовать молекулярные графы на основе их общих фрагментов. Типичные задачи этого вида анализа - это нахождение наибольших общих фрагментов в парах графов, вложение одного графа в другой, распознавание изоморфизма графов. Результаты данного вида анализа могут оказаться полезными в системах обработки структурной химической информации, в информационно-поисковых системах и банках данных по химии.

Блок СТРАН работает в четырех режимах, краткая характеристика которых дана ниже.

3.1. Изоморфизм графов. В режиме ИЗОМОРФИЗМ устанавливается тождественность (изоморфизм) двух молекулярных графов [7]. Графы являются изоморфными, если существует взаимно-однозначное отображение вершин графов друг на друга, сохраняющее смежность вершин, а также метки вершин и ребер. Все графы рассматриваемого набора данных проверяются попарно на изоморфизм. Результатом работы является сообщение об изоморфизме двух структур с указанием либо всех возможных соответствий вершин, либо с указанием какого-либо одного соответствия (по желанию пользователя).

3.2. Вложение графов. В режиме ВЛОЖЕНИЕ устанавливается вложение одного молекулярного графа в другой [7]. Граф H вкладывается в граф G , если в G найдется подграф, изоморфный графу H . Требуется два набора данных для установления вложения одной структуры в другую: каждая структура из второго набора данных будет проверяться на вложение в каждую структуру из первого набора данных.

Результатом работы является сообщение о всех вложениях (или невложениях) одной структуры в другую с указанием либо всех возможных соответствий вершин, либо с указанием какого-либо одного соответствия (по желанию пользователя).

3.3. Пересечение графов. В режиме ПЕРЕСЕЧЕНИЕ осуществляется поиск общих фрагментов (пересечений) структур [7]. Под общим фрагментом двух структур G_1 и G_2 понимается граф H , который изоморфен некоторому подграфу из G_1 и некоторому подграфу из G_2 . Если H максимален, т.е. имеет наибольшее число вершин, для которого еще сохраняется указанное свойство, то H называется максимальным общим фрагментом. Такой общий фрагмент называют также пересечением двух структур. Поиск общих фрагментов осуществляется для всех графов рассматриваемого набора данных фрагментов. По желанию пользователя могут быть найдены:

- все пересечения;
- максимальные пересечения;
- пересечения заданного порядка.

При поиске пересечений заданного порядка указывается нужный порядок фрагментов. Получаемые общие фрагменты могут быть как связными, так и несвязными, что учитывается с помощью соответствующего управляющего параметра.

Результатом работы является сообщение о всех общих фрагментах, или всех максимальных фрагментах, или всех общих фрагментах заданного порядка двух структур с указанием соответствия вершин.

3.4. Анализ циклических фрагментов. Методы данного вида анализа позволяют определить цикличность или ацикличность структур, выявить циклы и циклические системы в структуре. Предоставляется возможность решить следующие задачи.

1. Определить мосты, точки сочленения и блоки графа. Точкой сочленения графа называется вершина, удаление которой вместе с инцидентными ей ребрами нарушает связность графа, т.е. увеличивает число его компонент связности. Ребро графа, при удалении которого нарушается связность графа, называется мостом. Блоком графа называется максимальный связный подграф без собственных точек сочленения.

Результатом работы является перечисление всех точек сочленения, мостов, блоков с указанием вершин.

2. Найти циклические блоки графа и их граничные вершины. Циклический блок - это блок, все вершины которого принадлежат циклам графа. Если блок не является ребром, то он всегда будет циклическим блоком. Граничной вершиной циклического блока называется вершина циклического блока, смежная с вершинами, не принадлежащими данному циклическому блоку. Через граничные вершины циклический блок соединяется с остальной частью графа.

Результатом работы является сообщение о циклических блоках графа с указанием вершин, а также сообщение о граничных вершинах для каждого циклического блока.

3. Найти циклические части графа и их граничные вершины. Циклической частью графа называется максимальный связный подграф без мостов. Циклическая часть графа может содержать собственные точки сочленения. Граничной вершиной циклической части называется вершина циклической части, смежная с вершинами, не принадлежащими данной циклической части. Через граничные вершины циклическая часть соединяется с остальной частью графа.

Результатом работы является сообщение о циклических частях графа с указанием вершин, а также сообщение о граничных вершинах для каждой циклической части.

4. Установить принадлежность вершин графа его циклическим фрагментам. Вершины, принадлежащие циклическим фрагментам графа, называют циклическими вершинами. Результатом работы является перечисление всех вершин, принадлежащих циклам графа.

4. Блок Структурно-Метрического Анализа (блок СМАН)

Методы структурно-метрического анализа предназначены для метрической характеристики взаимного и относительного положения фрагментов, имеющих различные свойства в молекулярных структурах. Один и тот же фрагмент в разных структурах может занимать различные положения, и при этом соединение может проявлять различные свойства. Структурно-метрический анализ позволяет определять значения метрических и цепных индексов фрагментов с учетом их положения в графе. Кроме самих индексов, вычисляются также нормированные значения этих индексов.

Блок структурно-метрического анализа позволяет решать следующие задачи.

1. Вычисление метрических индексов любого фрагмента в графе, заданного перечислением вершин.

2. Определение вложимости заданного графа в другой граф и вычисление метрических индексов всех фрагментов-вложений.

Предлагаемые в блоке индексы представлены тремя значениями.

1. Метрические индексы, определяемые как сумма соответствующих индексов вершин графа, содержащихся в этом фрагменте.

2. Метрические индексы, нормированные по значению аналогичного интегрального индекса графа. Данная нормировка определяет относительный метрический индекс фрагмента.

3. Метрические индексы, нормированные по отклонению значения метрического индекса фрагмента от наименьшего и наибольшего возможных значений данного метрического индекса фрагмента с тем же числом вершин в заданном графе. Значение нормирован -

ного таким образом метрического индекса фрагмента заключено между 0 и 1. Значение 0 соответствует самому "центральному" расположению фрагмента, а значение 1 - самому "периферийному". Указанный способ нормировки может использоваться для сравнения значений метрических индексов фрагмента в семействе графов, существенно различающихся своими параметрами, например, числом вершин.

Результатом работы являются значения метрических индексов фрагмента и их нормированные значения. Если имеется несколько вложений заданного графа в граф, то вычисляются метрические индексы для всех фрагментов-вложений.

Нахождение зависимостей между структурой и свойствами молекулярных графов - одна из основных областей применения структурно-метрического анализа.

5. Блок ТАКСонимия (блок ТАКС)

Таксономия предназначена для обработки информации, представленной в виде таблицы. Результатом работы данного блока является разбиение заданного множества объектов на группы (таксоны, кластеры, классы) объектов, "близких" по своим свойствам (признакам). Предлагаемые алгоритмы таксономии разбивают все множество объектов либо на таксоны "сферической" формы, когда все объекты одного таксона покрываются сферой в пространстве признаков, либо на таксоны произвольной формы,

Блок ТАКС работает с данными двух типов: с таблицей "объект-признак", элементами которой являются значения признаков рассматриваемых объектов и которую можно создать на основе данных пользователя, и с таблицей расстояний между объектами, создаваемой программно в блоке МЕТАН на основе метрических свойств этих объектов.

В качестве объектов таблицы "объект-признак" могут быть взяты химические соединения или их молекулярные графы, а в ка-

честве признаков - физико-химические свойства, электронно-топологические параметры соединений, молекулярные индексы, структурные, метрические и другие индексы молекулярных графов и т.д.

Для выявления общих свойств используют разные способы объединения в группы (таксоны, классы, кластеры) объектов, "близких" по значениям признаков. Такая группировка производится с помощью алгоритмов таксономии, отличающихся способами объединения объектов в таксоны и критериями качества. Для нахождения таксонов разной пространственной формы в МЕТАХИМ применяются следующие алгоритмы.

5.1. Алгоритм СКАТ. Предназначен для разбиения множества объектов на таксоны сферической формы с проверкой на устойчивость (типичность) таксонов.

Алгоритм СКАТ состоит из двух этапов [8]. Вначале в пространстве количественных признаков производится таксономия с помощью следующего алгоритма. Центр сферы радиуса R помещается в начальную точку множества объектов, находится центр тяжести точек внутри этой сферы и центр сферы перемещается в найденный центр тяжести и т.д., пока центр тяжести не стабилизируется. Внутренние точки полученной сферы относятся к первому таксону. Затем центр сферы помещается в одну из оставшихся точек и процесс повторяется, пока все точки не будут распределены по таксонам. На втором этапе производится испытание таксонов на устойчивость: сферы радиуса R "запускаются" по очереди из центров всех таксонов, при этом одни сферы остаются на месте, а другие (неустойчивые) "скатываются" к центрам других таксонов.

На этом же методе поиска таксонов основан алгоритм СКАТ2 [8], отличающийся от алгоритма СКАТ процедурой изменения радиуса.

Результаты работы таксономии определяются набором параметров, управляющих режимом работы алгоритма, способом задания и

изменения радиуса сферы, количеством получаемых таксонов, и выводятся в виде распределения объектов по таксонам.

Алгоритм целесообразно использовать при компактном расположении объектов каждого таксона в пространстве признаков.

5.2. Алгоритм КРАБ. Предназначен для разбиения множества объектов на заданное число таксонов произвольной формы.

Алгоритм КРАБ [8] в пространстве количественных признаков строит кратчайший незамкнутый путь, соединяющий все точки множества объектов. Для заданного числа k таксонов с помощью полного или ограниченного перебора находятся такие $(k-1)$ ребер, устранение которых приводит к лучшему по качеству разбиению кратчайшего незамкнутого пути на k таксонов.

При большом числе объектов (>100) применяется алгоритм КРАБ2 [8]: сначала проводится предварительная таксономия с помощью первого этапа алгоритма СКАТ, а затем таксономия центров полученных таксонов с помощью алгоритма КРАБ.

Результаты работы таксономии определяются набором параметров, управляющих режимом работы алгоритма, выбором вида критерия качества, количеством получаемых таксонов, и выводятся в виде распределения объектов по таксонам.

Алгоритм целесообразно использовать при отсутствии априорных данных о расположении объектов в пространстве признаков.

5.3. Алгоритм ТАКЛИ. Предназначен для разбиения множества объектов, описанного в виде графа с заданной матрицей расстояний между вершинами (объектами), на таксоны сферической формы.

Алгоритм ТАКЛИ [9] для заданного значения порога строит граф, вершины которого соединены, если расстояния между ними меньше порога, т.е. если вершины находятся в сфере с диаметром, равным порогу. В полученном графе находятся все клики (подграфы, все вершины которых соединены), и из них выбираются клики-таксоны. Порог изменяется последовательно с шагом, равным D/N , где D - максимальное расстояние между объектами, N - задаваемое число итераций (управляющий параметр).

Алгоритм целесообразно применять для множества с компактным расположением объектов каждого таксона и заданными расстояниями между объектами.

В следующих версиях МЕТАХИМ предполагается в качестве признаков объектов использовать не только данные, полученные в блоке МЕТАН, но и данные, получаемые в блоках СТРАН и СМАН. Комбинирование этих данных поможет качественно расширить признаковое пространство объектов.

Результаты ТАКС можно использовать при анализе зависимости "структура-свойство" химических соединений, для выделения групп объектов по значениям их структурных характеристик и для сопоставления этих объектов по значениям их физико-химических свойств.

З а к л ю ч е н и е

Работа с комплексом МЕТАХИМ осуществляется в интерактивном режиме на персональной ЭВМ типа IBM PC/AT/XT. Основные языки программирования - ПАСКАЛЬ, ФОРТРАН. Требуемый объем памяти ОЗУ - 640 КБ. Размерность обрабатываемой структурной информации, представимой в виде графов, до 100 вершин.

Л и т е р а т у р а

1. КОЧЕТОВА А.А., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Язык описания структурной информации ОГРА-3.0 //Машинные методы обнаружения закономерностей, анализа структур и проектирования. - Новосибирск, 1982. - Вып. 92: Вычислительные системы. - С.70-79.
2. SKOROBOGATOV V.A., DOBRYNIN A.A. Metric Analysis of Graphs //Math. Chem. (MATCH). - 1988. - N 23. - P. 105-151.
3. РУВРЭ Д. Следует ли заниматься разработкой топологических индексов? //Химические приложения топологии и теории графов. - М.: Мир, 1987. - С. 183-205.
4. СТАНКЕВИЧ М.И., СТАНКЕВИЧ И.В., ЗЕФИРОВ Н.С. Топологические индексы в органической химии //Успехи химии. - 1988. - Т. 57. Вып. 3. - С. 337-366.

5. НЕКРАСОВ Ю.С., СУХАРЕВ Ю.Н., МОЛГАЧЁВА Н.С., ТЕПФЕР Э.Э., ЗАГОРЕВСКИЙ Д.В., СКОРОБОГАТОВ В.А., МЖЕЛЬСКАЯ Е.В. Информационные индексы масс-спектров и их корреляция с инвариантами молекулярных графов металлоорганических соединений //Тезисы докладов на УШ Всесоюзной конференции по использованию вычислительных машин в спектроскопии и химических исследованиях. -Новосибирск, 1989. - С. 256.

6. МЖЕЛЬСКАЯ Е.В., СКОРОБОГАТОВ В.А. Метрические спектры графов //Математические вопросы химической информатики. -Новосибирск, 1990. - Вып. 130: Вычислительные системы. -С.68-83.

7. МАКАРОВ Л.И., СКОРОБОГАТОВ В.А. Комплекс программ для исследования зависимости "структура-свойство" химических соединений //Алгоритмический анализ графов и его приложения. -Новосибирск, 1988. -Вып. 127: Вычислительные системы. -С. 92-129.

8. Пакет прикладных программ ОТЭКС //Загоруйко Н.Г., Елкина В.Н., Емельянов С.В., Лбов Г.С. - М.: Финансы и статистика, 1986.-158 с.

9. СКОРОБОГАТОВ В.А., МЖЕЛЬСКАЯ Е.В., МЕЙРМАНОВА Н.М. Изучение метрических характеристик ката-конденсированных полибензолов //Алгоритмический анализ графов и его приложения. - Новосибирск, 1988. - Вып. 127: Вычислительные системы. -С.40-91.

Поступила в ред.-изд.отд.

18 июня 1990 года