АЛГОРИТМЫ АНАЛИЗА СТРУКТУРНОЙ ИНФОРМАЦИИ
(Вычислительные системы)
1984 год Выпуск 103

УДК 519.I

ВОПРОСЫ АНАЛИЗА И РАСПОЗНАВАНИЯ МОЛЕКУЛЯРНЫХ СТРУКТУР НА ОСНОВЕ ОБЩИХ ФРАГМЕНТОВ

Н.Г.Загоруйко, В.А.Скоробогатов, П.В.Хворостов

Известное направление химических исследований, посвященных изучению зависимости между химическими или биологическими свойствами веществ и их строением, интенсивно развивается [I,2]. Эти исследования затрагивают различные аспекты, например, в [3] изучаются зависимости физико-химических параметров многоатомных молекул углеводородов от структуры взаимного расположения атомов и их связей. Одним из важнейших является вопрос о связи биологической активности соединений и их химического строения, в частности, для лекарственных препаратов. Впервые этот вопрос возник в середине I9 века. Позже было установлено, что всякое изменение в структуре молекулы органического вещества изменяет в ту или иную сторону характер его действия на живой организм. Более общее предположение, которое подтверждается экспериментально, состоит в том, что сходные по структуре соединения должны оказывать сходные воздействия. Подходы к определению структурного сходства или подобия различны. Для этих целей можно использовать любые конструктивно выражаемые свойства структур, такие, например, как метрические, конформационные, циклические или другие топологические свойства. Традиционно подобие строения соединений определяется через наличие в молекулах определенных структурных фрагментов. Фрагменты могут состоять из функциональных групп или каких-то других частей молекул, которые возникли на основе привычных классификационных признаков химических классов.

По мере роста сложности новых соединений и их проявлений возникают нетрадиционные ситуации, когда объективный выбор классификационных фрагментов становится затруднительным, поскольку

обнаружить достаточно сложные (крупные) фрагменты в большом количестве структур невозможно. Тогда либо фрагменты неоправданно упрощаются, т.е. теряется информация о возможном наличии общих структурных свойств в данном множестве соединений, либо они выбираются субъективно с целью проверки при последующей классификации структур на обучающих выборках. В этом случае успех при выборе подходящей системы фрагментов существенно зависит от объема вычислительной работы, в основе которой лежит перебор большого числа всевозможных подходящих фрагментов.

Наличие эффективной машинной методики нахождения достаточно полных систем структурных фрагментов для реальных классов соединений позволит получать объективно существующие признаки подобия и, следовательно, получать более объективные решения.

Результат выбора систем фрагментов зависит еще и от степени детальности описания структур: можно описывать молекулы с точностью до отдельных атомов, т.е. во входном описании структуры каждый атом кодировать отдельно. В этом случае анализ структур может быть осуществлен наиболее детально. Но и сложность такого анализа существенно возрастет, так как увеличивается размерность комбинаторных задач, которые лежат в основе такого анализа. По-видимому, для определенных приложений должна подбираться необходимая степень детализации описания молекулярных структур. Так, например, при исследовании активности лекарственных препаратов [2] используются так называемые фармакофоры - небольшие структурные фрагменты, состоящие из определенным образом соединенных атомов. Установление структурного подобия в указанных задачах может проводиться с точностью до фармакофоров.

При исследовании других видов активности, по-видимому, также могут быть определены группы атомов, аналогичные фармакофорам.

В [4] приводятся требования к системам признаков. Признаки должны порождаться автоматически, в результате анализа должен быть ясен их физический смысл и число их должно быть небольшим.

Из последнего требования можно сделать вывод, что признаки должны быть достаточно крупными (в пределе максимально крупными) и, следовательно, информационно емкими.

I. Некоторые функции системы SISTRAN. В данной работе приводится описание алгоритмической системы обработки структурной информации, предназначенной для поиска признаков структурного подобия молекулярных графов органических соединений. В системе преду-

сматриваются средства для экспериментов по распознаванию и прогнозированию свойств; она может также применяться при топологическом анализе других видов структурной информации.

I.I. О п и с а н и е и в в о д с т р у к т у р. Может проводиться одновременная обработка ряда семейств молекул, описанных на языке ОГРА-30 [5]. Точность описания фиксируется в виде базисного набора меток вершин и связей молекулярных графов. После необходимого контроля и проверки исходной информации путем применения контролирующего модуля информация записывается в память ЭВМ в виде матриц смежности молекулярных графов.

Правила описания на языке ОГРА-30 напоминают правила кодирования Висвессера, однако, в связи с тем, что группы атомов при кодировании на языке ОГРА-30 имеют номера, фактически отдельно описываются топология структуры и дополнительно ее нетопологические особенности, в то время как в коде Висвессера [6] топология описывается одновременно и наглядность записи теряется (рис.I).

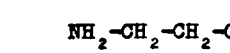


Рис.I

Код на языке ОГРА-30: INH$_2$--3-4CO-5OH или I2--3-4V-5Q.

Код Висвессера: Z2VQ , где Z-группа NH$_2$, V-группа >C=O, Q-группа - OH.

По этой причине ОГРА-30 обладает значительной универсальностью и может быть без переделки приспособлена для описания различных форм структурной информации только изданием соответствующих специальных инструкций по кодированию в данной области приложений.

Следует отметить, что как вся представляемая здесь система, так и версия языка ОГРА-30 ориентированы на пакетный режим обработки структурной информации, при котором информация готовится заранее, а не непосредственно при решении задач на ЭВМ в режиме диалога. Такой подход обладает известными недостатками, но не требует уникального оборудования. Ввод структур может быть осуществлен с перфокарт или магнитной ленты, либо с алфавитно-цифрового дисплея.

2. Хранение и поиск структур. Для увеличения гибкости при обработке множества структур, состоящего из отдельных семейств, может возникнуть необходимость изменить отдельные семейства путем их пополнения или замены в них некоторых структур. Для этого необходимо иметь средства идентификации, позволяющие устанавливать

наличие или отсутствие данной структуры в семействе. Эта манипуляция обеспечивается двумя функциями системы: хранением структур в виде линейных канонических кодов на языке ОГРА-30 и канонизацией запрашиваемой структуры с последующим сравнением канонических кодов.

Приведем более подробное описание метода канонизации графов на языке ОГРА-30.

2.1. К а н о н и з а ц и я г р а ф о в. При массовом распознавании изоморфизма графов, их хранении и поиске удобно использовать следующий прием. В каждом классе попарно изоморфных графов выбирается один граф, называемый каноническим видом любого графа данного класса. После этого распознавание изоморфизма графов сводится к построению и сравнению канонических видов. Каноническим видом графа называется граф, матрица смежности которого определяется переупорядочиванием вершин в соответствии с их свойствами, не зависящими от исходной нумерации. Нумерацию вершин графа, соответствующую каноническому виду, назовем канонической.

Для построения канонической нумерации вершин помеченного графа $G(V,X)$, $|V| = p$, сначала рассматривается упорядоченное разбиение множества V на классы эквивалентности, совпадающие с орбитами графа: $V = \{V_1,\dots,V_S\}$, $V_i \cap V_j = \emptyset$, $i \neq j$. Считаем, что $V_i > V_j$, если $m(v) > m(u)$, $\forall v \in V_i, \forall u \in V_j$, где $m: V \to R' \subset R$ – однозначная функция, ставящая в соответствие меткам вершин действительные числа. Если $m(v) = m(u)$, то $V_i > V_j$, в том случае, когда вектор $D(v) = (d_1(v),\dots,d_s(v))$ лексикографически старше вектора $D(u) = (d_1(u),\dots,d_s(u))$, где d_i – число ребер, соединяющих вершину v (или u) с вершинами из класса V_i, $i = \overline{1,s}$. Для мультиграфа вместо одного числа d_i рассматривается набор весов ребер.

Если $s = p$ (в каждом классе V_i содержится ровно одна вершина), то полученный порядок вершин является канонической нумерацией.

Если $s < p$, то выбираются первый по порядку класс V_i, $|V_i| > 1$, и некоторая вершина $v \in V_i$. К разбиению $\{V_1,\dots,V_i\setminus v,\dots, V_S, v\}$ применяется итеративная процедура доразбиения классов в соответствии с векторами D[7]. Если число полученных классов совпадает с p, то имеем каноническую нумерацию, иначе снова выбираем $|V_i| > 1$ и повторяем процесс до тех пор, пока не получим $s = p$.

Очевидно, что канонические виды изоморфных графов совпадают.

Для компактного хранения графов в каноническом виде используется запись на языке ОГРА-30, в которой описание графа представ-

лено в виде его покрытия множеством цепей. Каждая цепь формируется следующим образом. Выбирается вершина v_1 с минимальным канонический номером и записывается последовательность ребер (v_1,v_2), $(v_2,v_3),\ldots,(v_{i-1},v_i),\ldots$, где v_i – вершина с минимальным номером, смежная с v_{i-1}. После записи ребро помечается. Если вершина v_i не имеет непомеченных инцидентных ребер, то цепь заканчивается. Далее выбирается вершина с минимальным номером, имеющая непомеченные инцидентные ребра, и начинается формирование новой цепи. Граф полностью покрыт, когда помечены все ребра.

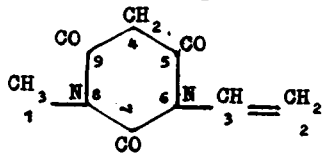


Рис. 2

Например, для графа G на рис.2 будет сформирована запись
G * 1-8 N-7CO-6N-3-2(2) * 4-5CO-6 * 4-9CO-8 **

При формировании записи опускаются метки C, CH, CH_2, CH_3, так как они могут быть однозначно восстановлены при учете валентности.

3. Нахождение общих максимальных фрагментов в парах структур. Под общим фрагментом двух структур G_1 и G_2 понимается граф, который изоморфен некоторому подграфу из G_1 и некоторому подграфу из G_2, причем хотя бы один из этих подграфов максимален, т.е. имеет наибольшее число вершин, для которого еще сохраняются указанные свойства. Такой общий фрагмент называют также пересечением двух структур. Так, например, бензол и нитробензол (рис.3) имеют общий фрагмент – бензольное кольцо.

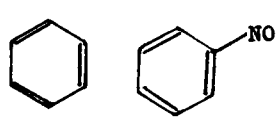


Рис. 3

Многие практические задачи обработки структурных данных сводятся к нахождению общих фрагментов. При помощи этой операции можно обеспечить поиск структур, содержащих заданный фрагмент или их набор, что позволяет сформировать множество или ряды похожих структур для дальнейшего анализа. Найдя все попарные пересечения для данного семейства структур и их взаимные вхождения после некоторого дополнительного анализа, можно сформировать набор фрагментов, объективно характеризующих данное семейство. Если решить такую задачу для достаточно представительной выборки, то можно сформировать правило для отнесения неизвестных молекулярных структур к данному химическому классу.

Предусматривается возможность осуществить не попарный, а одновременный поиск общих частей в семействе структур на основе k-местной операции пересечения [8] или путем многократного применения двухместной операции.

4. Задачи распознавания образов. Пусть дана обучающая выборка, состоящая из нескольких семейств структур. Найдем общие признаки-фрагменты (п.3) для каждого из семейств, а также будем использовать информацию о наличии или отсутствии этих признаков в структурах из других семейств. На основе такой информации можно построить решающее правило, позволяющее установить к какому семейству может быть отнесена некоторая новая структура.

Если дано множество структур, о котором предварительно ничего не известно, то можно найти разбиение этого множества на классы, пользуясь понятием подобия, введенного на основе метрики пространства признаков.

4.1. П о д о б и е г р а ф о в. Для определения степени подобия классов изоморфных графов введем понятие расстояния между графами в метрическом пространстве графов M. Под расстоянием в M будем понимать функцию $\rho(G,H) = p_G + p_H - 2p_{GH}$, где $G,H \in M$ – произвольные графы; p_G, p_H – порядки графов G и H соответственно; p_{GH} – порядок пересечения $G \cap H$.

Функция $\rho(G,H)$ является в некотором смысле обобщением метрики, введенной в [9] для p-вершинных графов: $d(G, H) = t$, где $p+t$ – наименьшее число вершин в графе, содержащем порожденные подграфы, изоморфные G и H.

Покажем, что функция $\rho(G,H)$ удовлетворяет аксиомам метрики

1) $\rho(G,H) \geq 0$, $\rho(G,H) = 0 \Rightarrow G \simeq H$,

2) $\rho(G,H) = \rho(H,G)$,

3) $\rho(G,H) + \rho(H,K) \geq \rho(G,K)$ для любых $G,H,K \in M$.

Аксиомы 1 и 2 – очевидны; аксиому 3 можно переписать $p_H + p_{GK} \geq p_{GH} + p_{HK}$. Если $p_H < p_{GH} + p_{HK}$, то $p_H \geq p_{GH} + p_{HK} - p_{GHK}$, а $p_{GHK} \leq p_{GK}$, что и доказывает 3. Здесь p_{GHK} – порядок пересечения $G \cap H \cap K$.

В практических задачах рассматриваются конечные семейства графов $S = \{G_1, \ldots, G_n\}$. Любой граф $G_j \in S$, $j=\overline{1,n}$, можно представить точкой в $(n-1)$-мерном пространстве M и, зная расстояния между каждой парой графов из S, вычислить их координаты как координаты точек в M.

Пусть $G_i, G_j \in S$, тогда

$$\rho(G_i, G_j) \equiv \rho_{ij} = \sqrt{\sum_{k=1}^{n-1} (x_k^i - x_k^j)^2},$$

где x_k^i или x_k^j – k-я координата графа G_i или G_j. Если начало координат поместить в точку, соответствующую некоторому графу, помечаемому индексом "0", тогда координаты следующего графа, помеченного индексом "1", будут $(x_1^1=\rho_{01}, x_2^1=0, \ldots, x_{n-1}^1=0)$ и т.д., для i-го графа имеем систему из i уравнений относительно его координат

$$\begin{cases} (x_1^i)^2 + \ldots + (x_i^i)^2 = \rho_{0i}^2, \\ (x_1^i - x_1^1)^2 + (x_2^i)^2 + \ldots + (x_i^i)^2 = \rho_{1i}^2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ (x_1^i - x_1^{i-1})^2 + \ldots + (x_{i-1}^i - x_{i-1}^{i-1})^2 + (x_i^i)^2 = \rho_{i-1,i}^2; \; 1 \le i \le n-1, \end{cases}$$

и ее решение

$$x_k^i = (\rho_{0i}^2 - \rho_{ki}^2 + \sum_{j=1}^{k} (x_j^k)^2 - 2 \sum_{j=1}^{k-1} x_j^i x_j^k)/2x_k^k; \quad 1 \le k \le i.$$

Представление графов точками в метрическом пространстве позволяет вводить различные критерии близости и решать задачи таксономии [10] на графах. Например, все p-вершинные графы, содержащие заданный подграф порядка p_1, можно заключить в шар диаметра $2(p-p_1)$. Конкретные алгоритмы таксономии, приспособленные для работы в метрических пространствах и с различными критериями близости, изложены в [22].

Указанные задачи могут возникать при анализе связей типа "структура-свойство", и данные методы могут применяться для исследования, например, биологической активности химических соединений.

5. Генерация структур. В системе предусматриваются средства генерации структур по заданной совокупности фрагментов и критерию подобия. Такая операция может понадобиться, например, если для некоторой выборки найдены фрагменты и есть необходимость дополнить эту выборку другими возможными структурами. Для этих целей использован программный модуль из [11].

6. Кратность и относительное положение признаков. Предусматриваются средства для описания кратности вхождения и относительного положения фрагментов в структурах и решения перечисленных задач с учетом этой информации. Такая возможность появляется благо-

даря достаточно хорошо развитому программному аппарату метрического анализа структур [I2].

Более детально остановимся на описании методики исследования зависимости между структурными особенностями и свойствами молекулярных графов.

7. Корреляция между структурой и свойством. Для установления корреляции между структурой вещества и его свойствами используется следующий подход. Пусть класс соединений $S_1 = \{G_1^1, \ldots, G_{N_1}^1\}$, где G_i^1, $i = \overline{1, N_1}$, - графы с помеченными вершинами и ребрами, представляющие структурные формулы соединений, проявляющие некоторый вид активности. Это может быть вызвано присутствием некоторого фрагмента F, который входит в структуру каждого из данных соединений, что может быть выражено в виде $F = \bigcap_{i=1}^{N_1} G_i^1$. На практике может оказаться, что общий фрагмент, принадлежащий пересечению одновременно всех структур из S_1, представляет собой "недостоверный" признак, так как достаточно часто встречается у соединений, не проявляющих данного вида активности. Такая ситуация может быть вызвана существованием нескольких структурных фрагментов, ответственных за наличие данного вида активности. Вследствие этого представляется более разумным рассматривать все максимальные попарные пересечения структур $G_i^1 \cap G_j^1$, $i,j \in \{1, \ldots, N_1\}$, $i < j$, и формировать набор потенциальных признаков, из которого затем выбирать подмножество таких структурных характеристик, по сочетанию которых можно достаточно надежно определять наличие интересующего свойства. Как уже говорилось в п.3, под пересечением графов G и H понимается максимальный по включению порожденный общий подграф K, компоненты связности которого рассматриваются в качестве множества признаков.

Для повышения надежности распознавания можно одновременно с классом S_1 рассматривать и другой класс соединений $S_2 = \{G_1^2, \ldots, G_{N_2}^2\}$, обладающий другим видом активности и формировать набор потенциальных признаков одновременно для двух классов (можно рассматривать и большее число классов).

Такой подход к отбору совокупности признаков позволяет исключать из рассмотрения признаки, достаточно часто встречающиеся у веществ, не обладающих заданной активностью.

Признак, который часто встречается у представителей одного класса и редко (или совсем не встречается) у представителей другого класса, имеет большую вероятность войти в информативное подмножество. Отсюда следует, что нет смысла тратить усилия на поиск общих фрагментов для молекул из разных классов - если такие и найдутся, они не войдут в информативный набор. Это замечание позволяет уменьшить перебор, сосредоточив внимание на поиске общих фрагментов только в молекулах веществ из одного и того же класса.

В качестве объекта, к которому применяется методика выбора информативных признаков, используется секционированная таблица $T = \|t_{ij}\|$ типа "объект-признак". Строки таблицы с номерами $i = \{1,\dots,m\}$ соответствуют графам G_i, представляющим структуры, а столбцы с номерами $j = \{1,\dots,l\}$ соответствуют признакам (или фрагментам) F_j. Секция таблицы S_α, $\alpha = \overline{1,q}$, - это множество ее строк, соответствующих структурам одного класса. Значения элементов таблицы t_{ij} вычисляются следующим образом: $t_{ij} = 1$, если F_j содержится в G_i, или $t_{ij} = 0$, если F_j не содержится в G_i.

В задачах распознавания сначала формируется полная таблица, а затем делается попытка построить решающее правило, определяющее связь между значениями t_{ij} признаков F_j объекта G_i и его принадлежностью к своему классу S_α.

8. Поиск общих подграфов и заполнение таблицы "объект-свойство". Для нахождения общих подграфов графов G и H используется метод, являющийся развитием метода, описанного в [24], и основанный на свойствах графов соответствий L, который определяется как модульное произведение этих графов $L = G \nabla H$.

ОПРЕДЕЛЕНИЕ 1. Пусть $G(V,X)$, $H(U,Y)$ - неориентированные, непомеченные, без петель и кратных ребер графы; V,U - множества вершин, X,Y - множества ребер графов G и H соответственно. Графом соответствий для графов G и H (или их модульным произведением) называется граф $L(W,E)$, множество вершин которого $W = V \times U$; т.е. $W = \{w=(v,u) \mid v \in V,\ u \in U\}$, а множество ребер определяется следующим отношением смежности вершин $w = (v,u)$, $w' = (v',u')$ графа L:

$$(w,w') \in E \Leftrightarrow \{(v,v') \in X,\ (u,u') \in Y\} \vee$$

$$\vee\ \{(v,v') \notin X,\ (u,u') \notin Y,\ v \neq v',\ u \neq u'\}.$$

Табличное представление операции модульного произведения приведено на рис. 4.

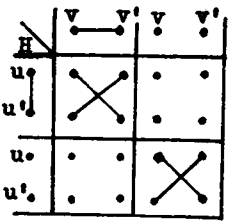


Рис.4

Граф соответствий содержит полную информацию о возможных соответствиях вершин графа G вершинам графа H, и основное его свойство состоит в том, что каждый максимальный полный подграф (клика) графа L соответствует максимальному общему подграфу графов G и H. Иначе говоря, если $\{w_1,\ldots,w_k\}$ – клика графа L, $w_i = \{v_i, u_i\}$, $v_i \in V$, $u_i \in U$, $i=\overline{1,k}$, то подграфы $\langle v_1,\ldots,v_k\rangle \subseteq G$ и $\langle u_1,\ldots,u_k\rangle \subseteq H$ максимальны и изоморфны.

Таким образом, для получения максимальных пересечений графов G и H достаточно построить $L = G \nabla H$ и найти в L клики наибольшего порядка и, применяя операцию модульного произведения к каждой паре химических структур из некоторого семейства, как к графам с помеченными вершинами и ребрами, можно получить полное множество их максимальных попарных пересечений. Но для этого требуется определить операцию модульного произведения для помеченных графов.

ОПРЕДЕЛЕНИЕ 2. Пусть $G(V,X)$ и $H(U,Y)$ – графы с помеченными вершинами и взвешенными ребрами. Введем однозначные отображения $m: V,U \to R' \subset R$ и $r: X,Y \to R'' \subset R$, где R – множество действительных чисел. Графом соответствий для графов G и H называется граф $L(W,E)$, множество вершин которого $W \subseteq V \times U$:

$$W = \{w = \{v,u\} \mid m(v) = m(u),\ v \in V,\ u \in U\},$$

$$(w,w') \in E \Rightarrow \{r(v,v') = r(u,u')\} \& \{(v,v') \notin X,\ (u\,u') \notin Y,\ v \neq v',\ u \neq u'\} \vee \{(v,v') \in X,\ (u,u') \in Y\}.$$

Заметим, что определение I является частным случаем определения 2, когда все метки вершин совпадают и веса ребер одинаковы.

Величина перебора, определяемая порядком графа соответствий, который в худшем случае равен произведению порядков исходных графов, и сложность алгоритма поиска клик [I3], даже в сравнительно простых случаях, велика и затрудняет практическое решение задач. Для сокращения перебора используются данные топологического анализа структур и способы упрощения графа соответствий, основанные на учете метрических свойств графов [I2,I4], использовании свойств симметрий [I5,I6] и рекурсивном разборе графов при поиске клик [I7-I9].

8.I. Анализ метрических свойств графов и свойств симметрий. Пусть G и H – помеченные графы. Вершины, имеющие одинаковые метки, будем называть сравнимыми. Зафиксируем некоторую пару сравнимых вершин $v \in V$, $u \in U$ и будем считать, что все попарно сравнимые вершины в графах G и H должны находиться на одинаковых расстояниях от v и u соответственно.

Множество сравнимых вершин в этих графах может быть построено путем определения относительных разбиений [I4].

ОПРЕДЕЛЕНИЕ 3. Относительным разбиением графа $G(V, X)$ для $v \in V$ называется множество $\hat{G}(v) = \{V_i(v),\ i = \overline{0,e(v)}\}$, $\bigcup_{i=0}^{e(v)} V_i = V$, $V_i \cap V_j = \emptyset$, $i \neq j$; $V_0 = \{v\}$, $v' \in V_i \Leftrightarrow d(v,v') = i$, $i = \overline{1,e(v)}$, где $e(v)$ – эксцентриситет вершины v, $d(v,v')$ – расстояние между v и v' . На рис.5 приведены примеры относительных разбиений.

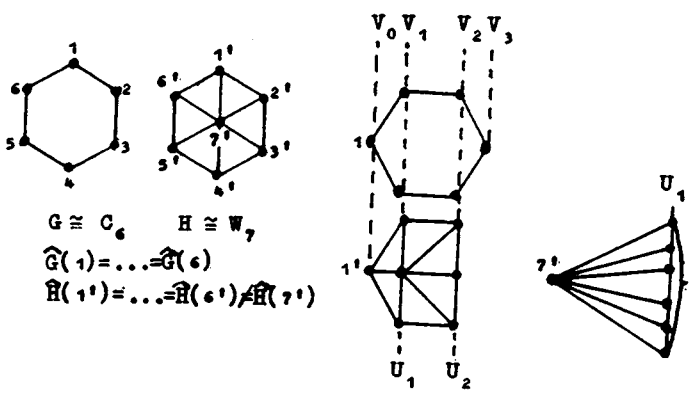


Рис. 5

Чтобы найти максимальные общие подграфы графов G и H, состоящие из сравнимых вершин, достаточно искать клики не во всем графе соответствий L, а только в некоторых его подграфах, определяемых как модульные произведения относительных разбиений: $\hat{L} = \hat{G}(v) \nabla \hat{H}(u)$.

ОПРЕДЕЛЕНИЕ 4. Пусть $\hat{G}(v)$ и $\hat{H}(u)$ – относительные разбиения графов G и H. Модульным произведением относительных разбиений называется порожденный подграф $\hat{L}(W', E')$ графа $L(W, E)$, множество

вершин которого определяется как $w' = \{ v' , u' \} \in W' \Leftrightarrow w' \in \in \bigcup_{i=0}^{k} V_i(v) \times U_i(u)$ и $m(v') = m(u')$, $k = \min(e(v), e(u))$.

Применение операции модульного произведения над разбиениями позволяет заменить анализ сложного графа анализом множества более простых графов. В общем случае такое множество порождается всеми парами относительных разбиений для сравнимых вершин графов G и H.

Эффективность вычислений зависит не только от количества разбиений для сравнимых вершин, но и от их сложности, которая определяется числом и порядком слоев. Чем длиннее разбиения и чем меньше вершин в слоях, тем проще графы соответствий.

ЗАМЕЧАНИЕ. Клики в модульных произведениях относительных разбиений могут не порождать максимальных общих подграфов исходных графов. Это обстоятельство связано с тем, что в построении модульного произведения участвуют только вершины графов G и H из слоев с одинаковыми номерами, т.е. в модульном произведении $\hat{G}(v) \nabla \hat{H}(u)$ не может присутствовать $w' = \{v', u'\}$, где $v' \in V_i(v)$, $u' \in H_j(u)$, $i \neq j$, что не позволяет охватить все возможное множество соответствий вершин графа G вершинам графа H. Так, для графов на рис. 5 вершины из слоя V_3, а в случае, когда рассматриваются произведения с $\hat{H}(7')$, и вершины из V_2 не участвуют в построении модульных произведений разбиений и поэтому не могут попасть в общие подграфы. Тем не менее очевидно, что $G \simeq H \backslash \{7'\}$.

В общем случае вопросы "полноты" модульных произведений разбиений составляют предмет специального исследования. Здесь заметим, что такая ситуация может возникнуть, когда в одном из графов присутствуют вершины с большой степенью, что крайне редко встречается в химических структурах. С другой стороны, потеря максимальности общего подграфа происходит, как правило, за счет отдельных мелких компонент связности, которые в большинстве случаев все же попадают в число признаков (при анализе других пар структур), и такая потеря компенсируется значительным сокращением времени решения задачи.

Переход к модульным произведениям разбиений позволяет сократить порядок анализируемых графов соответствий, а для того, чтобы уменьшить их число, используется информация об орбитах групп автоморфизмов графов $\Gamma(G)$ и $\Gamma(H)$. Поиск орбит осуществляется алгоритмом [17].

Пусть $g \in \Gamma(G)$ и $gv = v'$, $v, v' \in V$, $v = v'$. Поскольку автоморфизмы сохраняют смежность, то под действием g разбиение $\hat{G}(v)$ перейдет в $\hat{G}(v')$. Очевидно, $\forall u \in U$ графы $\hat{G}(v) \nabla \hat{H}(u)$ и $\hat{G}(v') \nabla \hat{H}(u)$ изоморфны, и, следовательно, порождают изоморфные подграфы графа G. Аналогично можно рассуждать и для графа H. Это свойство графов позволяет исключить из рассмотрения множество лишних изоморфных структурных фрагментов, порождаемых симметриями, при этом их количество может быть найдено без перебора. Если известно разбиение множества вершин графа на непересекающиеся части, в которых для каждой пары вершин существуют автоморфизмы, переводящие их друг в друга, то говорят, что заданы орбиты графа.

Как отмечено выше, любая пара вершин из одной и той же орбиты при анализе модульных произведений соответствующих относительных разбиений будет порождать изоморфные общие подграфы, и так как нас интересует только наличие того или иного фрагмента в исследуемой молекуле (число вхождений фрагмента в молекулу может быть вычислено), то при построении модульных произведений достаточно рассматривать по одной вершине из каждой орбиты графов G и H.

Например, граф G (рис.5) имеет только одну орбиту: (1,2,3,4, 5,6), а граф H - две орбиты: (1,2,3,4,5,6) (7). Чтобы получить все неизоморфные общие подграфы для G и H, достаточно рассмотреть произведения $\hat{G}(1) \nabla \hat{H}(1')$ и $\hat{G}(1) \nabla \hat{H}(7')$.

8.2. Рекурсивный разбор и клико - вая база графа. Для поиска клик в графе модульного произведения используется алгоритм рекурсивного разбора графов [18,19] и его модификация - алгоритм поиска клик максимального порядка, которые и определяют наибольшие общие подграфы исходных графов. Экспоненциальный рост числа клик с увеличением порядка графа ограничивает практическую применимость алгоритмов перечисления клик, однако, в ряде задач оказывается достаточным находить не все клики графа, а только кликовую базу [15], т.е. множество клик, непереводящихся друг в друга автоморфизмами графа.

Используемый алгоритм [16] нахождения максимальных клик из базы графа основан на рекурсивном разборе и использует в качестве входной информации орбиты группы автоморфизмов.

Рассмотрим граф модульного произведения. Вершины $w = \{v,u\}$ и $w' = \{v',u'\}$ будем считать принадлежащими одной орбите, если выполнено условие $(\exists g \in \Gamma(G) \mid gv = v') \wedge (\exists h \in \Gamma(H) \mid hu = u')$. Заметим, что такое определение орбит не охватывает всех симметрий мо-

дульного произведения, но оказывается достаточным для практических целей.

Пусть $K = \langle w_1,\dots,w_n\rangle$ и $K' = \langle w'_1,\dots,w'_n\rangle$ – автоморфные клики, т.е. w_i и w'_i принадлежат одной и той же орбите модульного произведения, $i = \overline{1,n}$. Тогда очевидно, что K и K' порождают изоморфные подграфы в исходных графах G и H. Таким образом, при поиске максимальных пересечений графов можно ограничиваться нахождением максимальных клик из кликовой базы модульного произведения.

При поиске максимальных пересечений можно пользоваться следующей оценкой [24] порядка клик в $\hat{G}(v) \nabla \hat{H}(u)$. Пусть $m_1(V_i),\dots,m_{k_1}(V_i)$ – количества вершин с метками $M_1,\dots,M_{k_1}$ в слое V_i разбиения $\hat{G}(v)$, $i = \overline{1,e(v)}$, а $m_1(U_i),\dots,m_{k_2}(U_i)$ – количества вершин с метками $M_1,\dots,M_{k_2}$ в слое U_i разбиения $\hat{H}(u)$, $i = \overline{1,e(u)}$. Если в слое нет вершин с некоторой меткой M_j, то $m_j(V_i) = 0$ или $m_j(U_i) = 0$. Порядок максимальной клики в $\hat{G}(v) \nabla \hat{H}(u)$ не может превышать величины

$$Q(v,u) = \sum_{i=1}^{k} \sum_{j=1}^{l} \min(m_j(V_i), m_j(U_i)),$$

где $k = \min(e(v), e(u))$, $l = \max(k_1, k_2)$.

Если порядок уже найденной клики в $\hat{G}(v) \nabla \hat{H}(u)$ превышает величину $Q(v',u')$, то нет необходимости рассматривать $\hat{G}(v') \nabla \hat{H}(u')$, так как максимальной клики здесь заведомо не содержится.

Итак, при нахождении наибольших общих подграфов в графах G и H выполняются следующие процедуры.

1. Находятся симметрии (орбиты групп автоморфизмов) графов G и H и производится выбор "вершин-представителей".

2. Строятся относительные разбиения для "вершин-представителей" и вычисляются оценки.

3. Строится модульное произведение разбиений и находятся его симметрии.

4. Происходит поиск максимальных базовых клик в модульном произведении и сравнение их порядка с оценками нерассмотренных пар разбиений.

5. Исключаются тождественные пересечения графов G и H. (Если некоторая клика K порождает подграф $F_1 \simeq \langle v_1,\dots,v_n\rangle \subseteq G$ и $F \langle u_1,\dots,u_n\rangle \subseteq H$, а клика K' порождает подграф $F_2 \simeq \langle v'_1,\dots$

..., $v'_n\rangle \subseteq G$ и $F_2 \simeq \langle u_1,...,u_n\rangle \subseteq H$, то $F_1 \simeq F_2$ и F_2 исключается. Аналогично исключается $F_2 \simeq \langle v_1,...,v_n\rangle \simeq \langle u'_1,...,u'_n\rangle$.)

8.3. Исключение изоморфных признаков-подгрфов и формирование таблицы "объект-признак". Выше были изложены принципы и методы поиска наибольших общих подграфов двух графов. Теперь переходим к нахождению общих подграфов для семейств графов, формированию множества признаков и заполнению таблицы "объект-свойство".

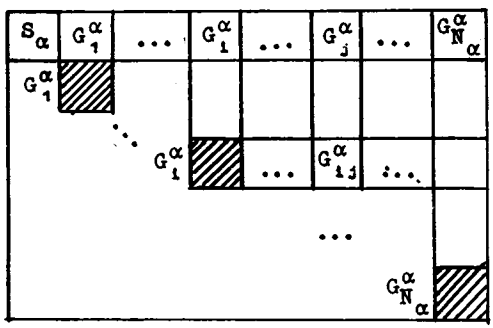


Рис.6

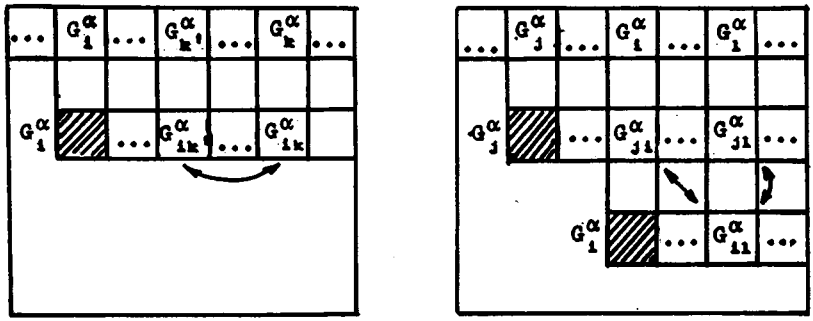
Пусть получены все попарные пересечения графов из класса химических структур $S_\alpha = \{G^\alpha_1,..., G^\alpha_{N_\alpha}\}$. Если некоторый подграф $G^\alpha_i \cap G^\alpha_j$ состоит из нескольких компонент, то каждая компонента рассматривается в качестве отдельного подграфа-признака. Множество пересечений $G^\alpha_i \cap G^\alpha_j$, $1 \le i \le j \le N_\alpha$, удобно изображать в виде таблицы (рис.6). Каждое пересечение $G^\alpha_i \cap G^\alpha_j$ определяется подстановкой вида

$$\tau^\alpha_{ij} = \begin{pmatrix} i_1 & \dots & i_n \\ j_1 & \dots & j_n \end{pmatrix},$$

где $\langle i_1,...,i_n\rangle$, $\langle j_1,...,j_n\rangle$ - порожденные подграфы G^α_i и G^α_j соответственно.

Для каждого фиксированного i, $1 \le i \le N_\alpha - 1$, подстановка τ^α_{ik}, $i+1 < k \le N_\alpha$, сравнивается со всеми подстановками $\tau^\alpha_{ik'}$, $i < k' < k$, по наборам номеров вершин в графе G^α_i (рис.7,а), при совпадении очевидно, что подграф, определяемый подстановкой τ^α_{ik}, изоморфен подграфу, порожденному $\tau^\alpha_{ik'}$, и подграф $G^\alpha_i \cap G^{\alpha_{ik}}_k$ исключается из набора признаков.

Для $1 \le i \le N_\alpha - 1$ каждая τ^α_{il}, $i < l \le N\alpha$, дополнительно сравнивается с τ^α_{ji} по наборам номеров вершин графа G^α_i и с τ^α_{jl}, $j < i$, по наборам номеров вершин графа G^α_l (рис.7,б).



а) б)

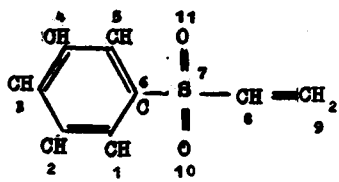
Рис. 7

ПРИМЕР. Исключение изоморфных пересечений для семейства хи - мических структур $S = \{G_1, G_2, G_3\}$:

G_1: $CH_3-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-S(=O)(=O)-CH=CH_2$ (1 2 3 4 5 6 7 8 9 10 11 12)

G_2: $CH_3-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-S(=O)(=O)-CH=CH_2$ (1 2 3 4 5 6 7 8 9 10 11 12 13)

G_3:



$$\tau_{12} = \begin{pmatrix} 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12 \\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13 \end{pmatrix},$$

$$\tau_{13} = \begin{pmatrix} 8\ 9\ 10\ 11\ 12 \\ 7\ 10\ 11\ 8\ 9 \end{pmatrix}, \qquad \tau_{23} = \begin{pmatrix} 9\ 10\ 11\ 12\ 13 \\ 7\ 10\ 11\ 8\ 9 \end{pmatrix}.$$

Набор пересечений состоит из трех фрагментов, определяемых подстановками: $F_1 \equiv \tau_{12}$, $F_2 \equiv \tau_{13}$, $F_3 \equiv \tau_{23}$. Фрагменты F_2 и F_3 изоморфны, так как совпадают наборы элементов в подстановках для графа G_3, поэтому F_3 исключается.

После того, как исключены изоморфные признаки-подграфы внутри каждого рассматриваемого класса S_α, $\alpha = \overline{1,q}$, производится непосредственная проверка изоморфизма между признаками в различных классах. Если некоторый признак F_1 изоморфен признаку F_2, то при исключении F_2 запоминаются структуры, в которые признак F_2 входит как подграф, и считается, что в эти структуры входит признак F_1.

Когда сформирован набор неизоморфных признаков $\{F_1,\dots,F_l\}$ решается задача вложения признака F_j, $j=\overline{1,l}$, в каждую из исходных структур, в которых не было зафиксировано присутствие признака F_j (задача вложения сводится к поиску максимального общего подграфа, порядок которого должен совпадать с порядком графа F_j) и заполняется таблица "объект-признак" $T = \|t_{ij}\|$, $i=\overline{1,n}$, $j=\overline{1,l}$, где

$$n = \sum_{\alpha=1}^{q} N_\alpha;\quad t_{ij} = \begin{cases} 1, & \text{если } F_j \text{ входит } i\text{-ю структуру,} \\ 0, & \text{если } F_j \text{ не входит в } i\text{-ю структуру.} \end{cases}$$

9. Нахождение систем признаков структурного различия [20]. Если таблица T содержит l столбцов, то в общем случае для выбора разделяющей совокупности признаков, по значениям которых можно было бы с определенной надежностью распознать, к какому классу принадлежит та или иная молекула, необходимо осуществить перебор, равный C_n^l. Для реальных ситуаций эта величина достаточно велика.

Существуют эвристические алгоритмы [21], которые достаточно быстро позволяют найти разделяющие системы признаков. При помощи этих алгоритмов из двоичных*) признаков x_j формируются "вторичные" признаки $y_i = x_{i_1}^{\sigma_1} x_{i_2}^{\sigma_2} \dots x_{i_k}^{\sigma_k}$, $\sigma_j \in \{0,1\}$, $x^0=\bar{x}$, $x^1=x$. Каждый признак y_i должен быть "истинным" на возможно большем числе объектов одного класса и ложным на в с е х объектах других классов. Набор признаков $Y = \bigvee_i y_i$ должен полностью "покрыть" соответствующий класс таблицы T. Это значит, что условие $Y=1$ будет

*) Для случая бинарных таблиц .

выполняться для всех структур данного класса, а для всех структур других классов $Y = 0$.

Набор Y есть булева функция от переменных (признаков) x, представленная в дизъюнктивной нормальной форме над множеством конъюнкций y_i. Эта функция на множестве из первого класса равна единице, а на множестве другого класса - равна нулю. Поскольку задачи анализа структур обладают спецификой, по сравнению с другими аналогичными задачами, то была предпринята попытка проанализировать существующие программы, предназначенные для построения решающих правил [22], с целью оценки возможности их приложения для решения указанных задач.

10. Экспериментальные данные о работе комплекса SISTRAN и исследовании программ построения решающих правил.

10.1. Программная реализация комплекса SISTRAN осуществлена в рамках ОС ЕС на языке ПЛ/I и использует мультизадачный режим для совмещения операций ввода-вывода с процессорными вычислениями. В комплекс включены следующие программы, имеющие свое функциональное назначение:

1. Транслирующая программа ОГРА-30.

2. Программы, обеспечивающие загрузку и накопление структур и найденных фрагментов.

3. Анализ симметрий структур.

4. Нахождение и формирование связных неизоморфных подграфов-признаков: поиск всех попарных пересечений исходных структур в семействе, нахождение компонент связности в пересечениях, определение симметрий пересечений, исключение изоморфных фрагментов и их накопление в виде таблицы, проверка изоморфного вхождения фрагментов в исходные структуры.

Комплекс позволяет обрабатывать семейства графов, состоящие из неограниченного числа классов. Каждый класс может состоять из не более 400 структур порядка ≤ 50. Эти ограничения не принципиальны, они являются следствием выбранного варианта форматов данных и могут быть легко изменены. Для опробывания возможностей комплекса программ SISTRAN было выбрано семейство из 160 барбитуратов [1]. Из данного семейства в качестве обучающей выборки были рассмотрены два класса соединений с определенными ограничениями времени угнетающего воздействия.

К первому классу были отнесены 18 соединений (табл.1) $t > 300$ мин, а ко второму классу 14 соединений (табл.2) $t < 15$ мин.

Таблица I.

Класс барбитуратов, t > 300 мин.

№	R	R'	T
54	CH_3CH_2-	$CH_2=C(CH_3)CH_2-$	326
55	-"-	$CH_3CH=CHCH_2-$	372
56	-"-	$CH_3CH_2OCH(CH_3)-$	460
53	-"-	$CH_2=CHCH(CH_3)-$	720
14	-"-	CH_3CH_2-	1400
15	-"-	$CH_3CH_2CH_2-$	1140
16	-"-	$CH_3CH(CH_3)-$	1520
17	-"-	$CH_3CH_2CH_2CH_2-$	450
18	-"-	$CH_3CH(CH_3)CH_2-$	540
19	-"-	$CH_3CH_2CH(CH_3)-$	600
1	CH_3-	$(CH_3)_3CCH_2-$	580
12	-"-	$CH_3CH_2SCH_2-$	330
88	$CH_3CH_2CH_2-$	$CH_2=CHCH(CH_3)-$	420
127	$CH_2=CHCH_2-$	$CH_2=CHCH(CH_3)-$	456
131	-"-	$(CH_3)_3CCHOCH_2$	900
132	-"-	$CH_2=C(CH_3)CH_2-$	380
139	-"-	$CH_2=CHCH_2-$	880
140	-"-	$(CH_3)CH_2-$	720

Из этих классов была сформирована 2-х секционная таблица T(I,2) (табл.3), в которой число признаков-фрагментов равно 6I.

Приведем данные, характеризующие параметры комплекса SISTRAN.

Попарные пересечения для первого класса строились I53 раза, для второго класса - 9I раз. С учетом связности набор признаков состоял из 660 (325 - первый и 335 - второй классы) фрагментов. После исключения изоморфных под-графов набор содержал 6I признак. Время работы - 25 мин (ЕС-I050).

Для обработки полученной таблицы T (табл.3) с целью получения решающих правил применялись программы TREE [23], DW [22].

I0.2. Построение решающих правил. Программа TREE для T(I,2) построила следующие решающие правила: если $Y_i(G) = 1 \Rightarrow G$ принадлежит i-му классу, $i = \{1,2\}$, то

$$Y_1 = x_{32}x_{23} \vee \bar{x}_{32}\bar{x}_{49}\bar{x}_{46}\bar{x}_{47} \vee \bar{x}_{32}\bar{x}_{49}\bar{x}_{46}x_{47}x_{11} , \qquad (1)$$

$$Y_2 = \bar{x}_{32}x_{49} \vee x_{32}\bar{x}_{23} \vee \bar{x}_{32}\bar{x}_{49}\,x_{46} \vee \bar{x}_{32}\bar{x}_{49}\bar{x}_{46}x_{47}\bar{x}_{11} . \qquad (2)$$

В этих правилах каждая конъюнкция истинна только на объектах своего класса, т.е. по этим правилам все структуры данной обучаю-

Таблица 2

Класс барбитуратов, t < I5 мин

№№	R	R'	T
48	CH_3CH_2-	$(CH_3)_2CHCH=CH$-	12
51	-"-	$CH_3(CH_2)_3CH=C(CH_3)$-	6
52	-"-	$CH_3CH_2CH=C(CH_3CH_2CH_2)$-	6
69	-"-	$CH_3(CH_2)_5SCH_2$-	15
71	-"-	$CH_3CH_2SCH(CH_3CHCH_3)$-	12
77	$CH_3CH_2CH_2$-	$CH_3CH_2CH_2CH_2CH_2$-	4
79	-"-	$CH_3(CH_2)_5$-	1
80	-"-	$CH_3(CH_2)_6$-	15
100	$(CH_3)_2CH$-	$(CH_3)_2CHCH=CH$-	12
107	$CH_3(CH_2)_3$-	$CH_3CH_2CH(CH_3)$-	16
108	-"-	$(CH_3)_3C$-	1
109	-"-	$CH_3CH=CH$-	12
155	$CH_3CH_2SCH_2$-	$(CH_3)_3CCH_2$-	8
157	$CH_3CH_2SCH(CH_3)$-	$CH_3(CH_2)_5$-	12

щей выборки распознаются без ошибок. Программа не делает оценки качества решающих правил, что делает затруднительным выбор необходимых решающих правил. Время работы - 37 сек (БЭСМ-6, FORTRAN).

Программа DW для T(I,2) построила правила:

$$Y_1 = x_6\vee\bar{x}_6\bar{x}_{32}x_5\vee\bar{x}_6x_{32}x_{32}\vee\bar{x}_6\bar{x}_{32}\bar{x}_5\bar{x}_{14}\bar{x}_{47}\vee\bar{x}_6\bar{x}_{32}\bar{x}_5\bar{x}_{14}x_{47}x_{11}\,, \qquad (3)$$

$$Y_2 = \bar{x}_6x_{32}\bar{x}_{23}\vee\bar{x}_6\bar{x}_{32}\bar{x}_5x_{14}\vee\bar{x}_6\bar{x}_{32}\bar{x}_5\bar{x}_{14}x_{47}\bar{x}_{11}\,. \qquad (4)$$

Все структуры распознаются без ошибок. Оценка качества решающего правила путем скользящего экзамена составила I8,75%. Время - I7 сек (ЕС-I050, FORTRAN). В табл.4 приведены признаки-фрагменты, используемые в решающих правилах.

С целью оценки качества решающих правил, полученных по T(I,2), из семейства барбитуратов была сформирована контрольная выборка, в

Таблица 3

Таблица "структура-признак" для двух классов

№ структуры		
54	1111111111010000000000000000100	0000000110000000000000000000000
12	1110001001110000000000000000100	0000000110000001000000010001000
55	1110101111011110000000000000100	0000000110000000000000000000000
132	1011011011011111110000000000100	0000000110000000000000000000000
88	1110011001010101001111000000100	0010000111000000000110010100101
17	1110101000010000001000100000101	1010000111100000000000000000000
127	1010011001010111100110010000100	0000000110000000000110010100000
56	1110001101010000000010001100100	0000000110000000000110010100000
18	1110101111010000000000000000100	0000000110000000000100000000000
1	1010001001110000000000000011100	0000000110000000000000000000000
19	1110001101010000000010001001100	0000000110000000000110010100000
53	1110011101010101000110001001100	0000000110000000000110010100000
140	1010011001010111100010010000110	0000000110000000000110010100000
139	0010011011011110000000000000100	0000000110000000000000000000000
131	1010011011011111110000000111100	0000000110000000000000000000000
15	1110101111010000001001100000100	0010000111000000000000000000000
14	1110101111010000000000000000100	0000000110000000000000000000000
16	1110001101010000000010001001100	0000000110000000000110010100000
79	1110001011010000001001000001011	1111111111011100000000000000100
108	1110001001010000001000000011011	0010000111110000000000000000000
77	1110001011010000001001000001011	1110100111011100000000000000000
51	1110001101010100000000000000100	1010000111010010000000000000000
52	1110001101010100000000000001100	0010000111000010000000000000000
155	1110001011010000000000000010100	0000000110000001100000010001000
48	1110001101010100000000000000100	0000000110000000011100001000000
71	1110001101010000000000000000100	0000000110000001000111110010000
100	1000000010101000000000000000100	0000000110000000010110011000000
109	1110001001010100001000000001011	0010000111110000001000001000000
157	1110001001010000001010000001011	0011110111011101000111110111110
69	1110101111010000000000000000100	1011100111010100100000100010000
80	1110001011010000001001000001011	1111101111011100000000000001000
107	1110001001010000001010000001011	0010000111110000000011001010011

которую вошли соединения со временем угнетающего воздействия $t = 300$ мин (№ 23,28,89,I28,I29,I50) и со временем $I5 < t < 30$ мин (№ II,43,47,70,82,83,86,87,99,I0I,I02,II0,I24,74,96,I58).

Результаты распознавания приведены в табл. 5 (I – объект классифицирован правильно, 0 – неправильно). Из таблицы видно, что соединения со временем воздействия $t = 300$ мин по решающим правилам (3)-(4), (I)-(2) относятся к первому классу. Вещества со временем воздействия $I5 < t < 30$ распознаются хуже, так, по правилам (3)-(4) неправильно распознаны 4 объекта, а по правилам (I)-(2) – I2 объектов. Таким образом, для данных выборок решающие правила, построенные программой DW, обеспечивают более надежное

Таблица 4

№№	Признак-фрагмент
6	CH_2-
14	-CH=
32	$CH_3(CH_2)_3$-
47	CH_3-CH_2-S-
5	-CH_2-Б-CH_2-CH_3
11	CH_3-Б-CH_2-
23	-$(CH_2)_2$- Б-CH_2-CH_3
46	CH_3-CH_2-Б-C=CH-CH_2
49	- Б-CH=CH-CH(CH_2)-

распознавание контрольной выборки, чем правила, построенные программой TREE.

ЗАМЕЧАНИЕ. Результаты применения правил (3)-(4) становятся лучше, если рассматривать I5 < t < 20 мин (неправильно распознается только один объект). Это позволяет предположить, что "граница" первого класса лежит в области значения t = 20.

Для рассмотренных программ безразлично какое значение принимают признаки, нуль или единица, - главное, чтобы их значения максимально различали объекты из разных классов. Отсюда в решающих правилах может возникать большое количество отрицаний наличия признаков у объектов.

В некоторых задачах предсказания свойств химических соединений большее значение имеет факт наличия некоторого фрагмента, а не его отсутствия. В связи с этим возникает необходимость в построении и анализе таких алгоритмов, которые бы позволили учитывать приоритет признаков и пытались бы сначала строить решающее правило по ненулевым значениям признаков и только в случае неудачи - по нулевым.

Легко проверить, что часть таблицы T, построенная для первого класса и признаков 5,6,II,I4,23,32,46,47,49, не содержит единиц в строчках для структур 8,II,I8. Откуда следует, что эти структуры по наличию признаков не классифицируются. Следует также отметить, что желательна реализация методики предварительного отбора признаков, т.е. дополнительные алгоритмические условия построения таблицы T, при которых она строилась бы постепенно и одновременно с попытками получения решающих правил. Такой подход может сократить время вычислений, так как для реальных семейств структур решающие правила могут быть найдены раньше, чем построены все попарные пересечения, что требует наиболее трудоемких вычислений.

Приведем один способ построения таких систем признаков. Пусть часть признаков построена. После формирования каждого нового признака будем проверять его на необходимость, а систему признаков с его участием - на достаточность. Необходимым считается признак, у которого не все элементы одинаковы, и если он не пол-

Таблица 5

N	T	DW	TREE
23	300	1	1
28	300	1	1
89	300	1	1
128	300	1	1
129	300	1	1
150	300	1	1
11	24	1	1
43	18	1	0
47	24	1	0
70	22	0	0
82	18	1	0
83	18	1	1
86	18	1	0
87	24	1	0
99	18	1	0
101	18	1	0
102	18	1	0
110	18	1	1
124	18	0	0
74	28	1	1
96	25	0	0
158	28	0	0

ностью совпадает с любым другим ранее найденным признаком. Если признак оказался необходимым, его добавляют в таблицу и полученную таким путем систему признаков проверяют на достаточность.

Достаточной для распознавания класса является такая система признаков, по которой удается построить решающую функцию, отличающую молекулу этого класса от молекулы всех других классов в достаточно большом числе случаев.

Если на очередном шаге система признаков оказалась недостаточной, то порождается очередной признак, проверяется на необходимость, а новая система проверяется на достаточность. Процедура продолжается до тех пор, пока все или заданный процент молекул данного класса не будет распознаваться правильно.

Затем аналогичный процесс построения решающего правила делается для молекул следующего класса. При этом вначале проверяется, не достаточна ли имеющаяся система признаков для построения надежного правила распознавания молекул класса, и если нет - порождается очередной признак.

Процесс заканчивается тогда, когда будет построен набор правил, по которым с заданной точностью будут распознаваться молекулы всех классов, представленных в таблице "объект-признак".

Заключение

Рассмотрена методика нахождения наибольших общих подграфов для всех пар графов из данного семейства, основанная на применении относительных разбиений. Рассмотрены вопросы применения данной методики для исследования связи "структура-активность" для структур молекулярных графов, а также указаны необходимые функции пакета программ, предназначенного для анализа и распознавания химических структур.

Исследована применимость методов распознавания образов для задачи прогнозирования свойств, основанной на общих фрагментах.

Обращается внимание на необходимость разработки методов построения решающих правил для бинарных таблиц, которые бы, по возможности, не использовали отрицаний признаков.

Литература

I. СТЬЮПЕР Э., БРЮГГЕР У., ДЖУРС П. Машинный анализ связи химической структуры и биологической активности. -М.: Мир, 1982. - 235 с.

2. Методы представления и обработки структурной информации для анализа связи структура-активность /Гитлина Л.С., Голендер В.Е., Дрбоглав В.Б. и др. -Рига, 1981. - 74 с. (Препринт/ Ин-т органического синтеза АН ЛатвССР).

3. ЯРОВОЙ С.С. Методы расчета физико-химических свойств углеводородов. -М.: Химия, 1978. - 256 с.

4. Исследование эффективности некоторых статистических алгоритмов предсказания биологической активности многоатомных молекул /Нигматуллин Р.С., Осипов А.Л. и др. - В кн.: Использование вычислительных машин в спектроскопии молекул и химических исследованиях. УI Всесоюз. конф. Тез. докл. Новосибирск, 1983, с. 22-23.

5. КОЧЕТОВА А.А., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Язык описания структурной информации ОГРА-30. -В кн.: Машинные методы обнаружения закономерностей, анализа структур и проектирования (Вычислительные системы, вып. 92), Новосибирск, 1982, с. 70-79.

6. ВЛЭДУЦ Г.Э., ГЕЙВАНДОВ Э.А. Автоматизированные информационные системы для химии. -М.: Наука, 1974. - 48 с.

7. CORNEIL D.G., GOTLIEB C.G. An efficient algorithm for graph isomorphism. - Journal ACM, 1970, v. 17, p. 51-64.

8. ДЕНИЩИК Е.Ю., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Нахождение общих подструктур в семействах химических соединений. -В кн.: Использование вычислительных машин в спектроскопии молекул и химических исследованиях: УI Всесоюз. конф. Тезисы докл. Новосибирск, 1983, с. 197-198.

9. ZELINKA B. On a certain distanse between isomorphism classes of graphs. - Cas. pěst. mat., 1975, v. 100, N 4, p. 371-373.

10. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. -М.: Сов. радио, 1972. - 206 с.

11. МОЛОДЦОВ С.Г., ПИОТТУХ-ПЕЛЕЦКИЙ В.Н. Построение всех неизоморфных химических графов из заданного набора структурных фрагментов. -Настоящий сборник, с. 51-58.

12. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Анализ метрических свойств графов. -В кн.: Методы обнаружения закономерностей с помощью ЭВМ (Вычислительные системы, вып. 91). Новосибирск, 1981, с. 3-20.

13. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. Применение относительных разбиений при поиске клик. -В кн.: Автоматизация проектирования в микроэлектронике. Теория. Методы. Алгоритмы (Вычислительные системы, вып. 77). Новосибирск, 1978, с. 25-33.

14. СКОРОБОГАТОВ В.А. Относительные разбиения и слои графов. -В кн.: Вопросы обработки информации при проектировании систем (Вычислительные системы, вып. 69). Новосибирск, 1977, с. 3-10.

I5. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Орбиты, клики, канонизация. -В кн.: Методы и программы решения оптимизационных задач на графах и сетях: Тезисы докл. Всесоюз. совещ. Новосибирск, I980, с. 85-87.

I6. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Анализ симметрий графов. -В кн.: Методы и программы решения оптимизационных задач на графах и сетях: Тезисы докл. П Всесоюз.совещ. Ч.2. Новосибирск, I982, с. 43-45.

I7. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Алгоритм нахождения кликовой базы графа. -Там же, с. I6-I8.

I8. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. Об одном семействе схем рекурсивного разбора графов. -В кн.: Машинные методы обнаружения закономерностей, анализа структур и проектирования (Вычислитель - ные системы, вып. 92). Новосибирск, I982, с. 3-49.

I9. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. О рекурсивном разборе графов. -В кн.: Алгоритмические основы обработки структурной информации (Вычислительные системы, вып. 85). Новосибирск, I98I,с.3-20.

20. ЗАГОРУЙКО Н.Г., СКОРОБОГАТОВ В.А. Выбор признаков структурного различия классов химических веществ. -Там же, с. 92-93.

2I. ЛБОВ Г.С., КОТОКОВ В.И., МАШАРОВ Ю.П. Метод поиска логических закономерностей на эмпирических таблицах. -В кн.: Эмпири - ческое предсказание и распознавание образов (Вычислительные системы, вып. 67). Новосибирск, I976, с. 29-4I.

22. ЗАГОРУЙКО Н.Г., ЛБОВ Г.С., МАШАРОВ Ю.П. Пакет прикладных программ для обработки таблиц экспериментальных данных ОТЭКС-I.-В кн.: Вопросы обработки информации при проектировании систем (Вы - числительные системы, вып. 69). Новосибирск, I977, с. 93-I0I.

23. ЛБОВ Г.С., ФРОЛОВА Т.И. Комплект программ для анализа данных с использованием логических решающих функций. -В кн.: Те - зисы докл. П Всесоюз. школы-семинара "Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа".М., I983, с. I03-I04.

24. СКОРОБОГАТОВ В.А. Нахождение общих частей в семействах графов. -В кн.: Материалы Всесоюзного совещания "Прикладные задачи на графах и сетях". Новосибирск, I98I, с. II7-I32.

Поступила в ред.-изд.отд.
I0 мая I984 года