АЛГОРИТМИЧЕСКИЙ АНАЛИЗ ГРАФОВ И ЕГО ПРИМЕНЕНИЯ
(Вычислительные системы)
1988 год Выпуск 127

УДК 519.1

КОМПЛЕКС ПРОГРАММ ДЛЯ ИССЛЕДОВАНИЯ ЗАВИСИМОСТИ "СТРУКТУРА-СВОЙСТВО" ХИМИЧЕСКИХ СОЕДИНЕНИЙ

Л.И.Макаров, В.А.Скоробогатов

Введение

Комплекс программ СИСТРАН-МГ (СИстема СТРуктурного АНализа Молекулярных Графов) предназначен для проведения научных исследований зависимости "структура-свойство" химических соединений в различных областях применения: фармакологии, создании химических средств защиты растений, изучении химического загрязнения окружающей среды и т.д.

Исследования на ЭВМ зависимости "структура-свойство" химических соединений основываются на гипотезе [1,2] о том, что близкие по структуре соединения имеют похожие свойства и, следовательно, общие свойства веществ определяются общими фрагментами их структур.

По мере роста количества и сложности новых соединений выбор таких фрагментов и анализ их взаимосвязи становятся затруднительными. Поэтому для поиска признаков структурного сходства необходима машинная методика, позволяющая находить множество общих структурных фрагментов соединений, обладающих некоторым свойством, и выделять те фрагменты, сочетание которых определяет наличие этого свойства у исследуемых веществ.

В основу разработки комплекса программ СИСТРАН-МГ положена методика машинного анализа зависимости "структура-свойство" хи-

мических соединений, описанная в [3]. В качестве формальной модели молекулярной структуры химического соединения выбран молекулярный граф, а общим фрагментом молекул двух соединений считается фрагмент, соответствующий общему порожденному подграфу их молекулярных графов.

Методика нахождения общих фрагментов молекул химических соединений и прогнозирования их свойств состоит в следующем.

Пусть имеется совокупность соединений, разбитая на два класса - проявляющих и не проявляющих некоторое свойство. Из части этой совокупности формируется обучающая выборка, состоящая из представителей разных классов соединений, а из оставшейся части соединений может быть образована контрольная выборка. Для соединений каждого класса обучающей выборки находятся попарно наибольшие общие фрагменты их молекул и определяется вхождение этих фрагментов в молекулы соединений обоих классов.

На основе полученных данных формируются логические решающие правила [4,5], описывающие закономерности вхождения фрагментов в соединения разных классов обучающей выборки и позволяющие относить каждое контрольное соединение к одному из классов. Фрагменты, потенциально ответственные за проявление данного свойства, могут быть выделены из списка фрагментов, соответствующих лучшим по качеству правилам.

Кроме фрагментов соединений, для прогнозирования их свойств в комплексе программ СИСТРАН-МГ предусмотрена возможность использования количественных характеристик: брутто-формул связи соединений [6] и молекулярных индексов - метрических [7] и цепных характеристик [8] молекулярных графов.

При создании программ комплекса СИСТРАН-МГ использованы методы и алгоритмы, разработанные в Институте математики СО АН СССР для решения задач структурно-метрического анализа графов [3, 7-11] и обнаружения эмпирических закономерностей [4,5].

В разработке комплекса программ СИСТРАН-МГ принимали участие Е.И.Беляев, Ю.Е.Бессонов, А.А.Добрынин, А.А.Кочетова, Ю.П.Леоненко, Л.И.Макаров, В.А.Скоробогатов, И.С.Тимохина, П.В.Хворостов, С.А.Федоров.

1. Применяемые методы и алгоритмы

При анализе зависимости "структура-свойство" требуется находить не только общие фрагменты соединений, но и устанавливать их идентичность, а также вхождение фрагментов в соединения и т.д. Все эти задачи могут быть сведены к известным задачам нахождения наибольшего общего подграфа - пересечения графов, установления изоморфизма и изоморфного вложения, поиска клик в графе и т.д.

Пересечение графов. Пересечением $F = G \cap H$ графов $G = (V,X)$ и $H = (U,Y)$ порядков $p_G = |V|$ и $p_H = |U|$, $p_G \le p_H$, назовем максимальный по порядку граф F такой, что существуют порожденные подграфы $G' \subseteq G$ и $H' \subseteq H$, изоморфные F: $G' \simeq F \simeq H'$. Задачи изоморфизма графов и их изоморфного вложения являются частными случаями задачи нахождения пересечения графов, поскольку если $p_F = p_G = p_H$, то $G \simeq H$, а если $p_G = p_F < p_H$, то $G \simeq H' \subset H$, т.е. G изоморфно вложим в H: $G \overset{\sim}{\subset} H$.

Для точного решения задачи нахождения пересечения двух графов G и H применяется следующая общая схема [3]:

$$G,H \to L = G \nabla H \to K_{max} \to F = G \cap H.$$

1. Для графов G и H строится модульное произведение $G \nabla H = L(W,E)$ - граф соответствий, вершины w которого являются всевозможными парами вершин графов G и H, $W = (v,u) \in W = V \times U$ и w смежна с w', т.е. $(w,w') \in E$, если и только если $(v,v') \in X$ и $(u,u') \in Y$ или $(v,v') \notin X$ и $(u,u') \notin Y$, $v \neq v'$, $u \neq u'$. Порядок графа $p_L = p_G p_H$.

2. Кликой графа называют максимальный по включению полный его подграф. Поскольку каждая клика K графа L соответствует общему порожденному подграфу графов G и H, и наоборот, то в графе L находятся максимальные по порядку клики K_{max} и тем самым соответствующие им пересечения $F = G \cap H$.

Перечисленные задачи являются труднорешаемыми [12], т.е. в общем случае точное решение задачи находится с помощью перебора, и поэтому для их практического применения необходимо разрабатывать алгоритмы, позволяющие сократить перебор и находить в приемлемое время точные или приближенные решения этих задач.

Уменьшение времени нахождения общих фрагментов соединений может быть достигнуто разными способами. В комплексе программ ТОПЛОГ [2], разработанном в Институте органического синтеза АН ЛатвССР и предназначенном для отбора структурных признаков биологической активности химических соединений, используются следующие приемы сокращения времени вычислений:

- уменьшение порядков молекулярных графов за счет укрупнения их описания и устранения из них фрагментов, редких для данного класса соединений;

- сокращение числа соединений путем устранения редких и сложных соединений, не влияющих на результат поиска общих фрагментов;

- отсечение неперспективных ветвей дерева поиска по некоторым критериям, например, использование инвариантов молекулярных графов при установлении их изоморфизма.

Подобные способы на практике оказываются достаточно эффективными, но могут привести к потере нужной информации.

Разработанные в Институте математики СО АН СССР [3,9-11] алгоритмы нахождения точного решения вышеуказанных задач позволяют сократить перебор благодаря учету структурных особенностей исследуемых графов - метрических свойств и свойств сим-

метрии. Эти алгоритмы имеют достаточно высокое быстродействие, что дает возможность в приемлемое время находить наибольшие общие фрагменты в реальных семействах молекулярных графов.

Анализ метрических свойств графа проводится методом относительных его разбиений [13]. Относительным разбиением $\hat{V}(v)$ графа G по отношению к вершине $v \in V$ называют упорядоченное множество классов $\hat{V}(v)=\{V_k(v) \mid k=\overline{1,e(v)}, v' \in V_k(v) \Leftrightarrow \Leftrightarrow d(v,v')=k\}$ такое, что расстояние от v до любой v' из класса $V_k(v)$ равно k ; класс $V_k(v)$ называют k-слоем для v.

Симметрии графа описываются с помощью разбиения его вершин на орбиты - множества эквивалентных вершин. Вершины v, $u \in V$ называются эквивалентными (симметричными) относительно группы Γ_G автоморфизмов графа, если существует подстановка $\varphi \in \Gamma_G$ такая, что $\varphi v = u$. Орбитами графа G называются классы эквивалентности $\theta_i(G)$, $i = \overline{1,s}$, $s \le p_G$, разбиения $\hat{\theta}(G)$ множества его вершин по отношению к Γ_G.

В общей схеме поиска пересечения двух графов G и H относительные разбиения графов и их орбиты применяются в следующих случаях:

1) Орбиты графов G и H находятся с помощью алгоритма [9], использующего относительные разбиения и процедуру итеративной классификации вершин.

2) Для вершин v и u, представителей орбит из $\hat{\theta}(G)$ и $\hat{\theta}(H)$, находится множество $\mathcal{L}=\{L(v,u)\}$, где $L(v,u)$ - модульное произведение относительных разбиений $\hat{V}(v)$ и $\hat{U}(u)$, определяемое как порожденный подграф $L(W',E') \subseteq \subseteq L(W,E)$, множество вершин которого $W'=\bigcup_{k=0}^{e} V_k(v) \times U_k(u)$, $e=\min(e(v),e(u))$ (см. [3]).

Применение модульных произведений разбиений позволяет заменить анализ сложного графа $L = G \nabla H$ анализом совокупно -

сти $\mathcal{L}$ более простых графов $L(v,u)$, т.е. сократить порядок анализируемых графов, а учет симметрий графа - уменьшить их число за счет рассмотрения только вершин - представителей орбит графа.

3) Максимальные клики K_{max} графа $L(v,u)$ находятся также с учетом его орбит. Поскольку автоморфные клики графа $L(v,u)$, т.е. переводящиеся друг в друга его автоморфизмами, соответствуют изоморфным подграфам графов G и H , то производится поиск только неавтоморфных клик.

Каждой максимальной клике $K_{max}=\{w_i\}$, $w_i=(v_i,u_i)$ соответствует пересечение $F = G \cap H$, $F \simeq G' = \langle V' \rangle$, $V' = \{v_i\}$, $F \simeq H' = \langle U' \rangle$, $U' = \{u_i\}$.

Задача нахождения пересечений двух графов решается с помощью алгоритмов поиска орбит [9], нахождения модульных произведений, изоморфизма графов [3] и поиска клик (алгоритм рекурсивного разбора) [11].

Алгоритм поиска орбит графа $G(V,X)$ основан [9] на построении разбиений множества вершин V на классы эквивалентности по отношению к набору характеристик графа, инвариантных относительно автоморфизмов, и последовательном "дроблении" полученного разбиения до совпадения с разбиением $\hat{\theta} = \{\theta_i\}$, $i = \overline{1,s}$, множества вершин на орбиты θ_i .

Алгоритм реализован в виде следующей последовательности процедур:

1) построения разбиения $\hat{V}_M$ множества V по меткам вершин.

Вершины графа попадают в один и тот же класс, если они имеют одинаковые метки;

2) "дробления" начального разбиения $\hat{V}_M$ с учетом метрических характеристик вершин.

Вычисляются попарные расстояния между вершинами графа. Вершины из одного и того же класса в разбиении $\hat{V}_M$ будут отне-

сены к различным классам нового разбиения $\hat{V}^*$,если они имеют различное число соседей хотя бы в одном из классов разбиения $\hat{V}_M$. Данная процедура (итеративная классификация) повторяется до наступления стабилизации разбиений;

3) проверки совпадения классов разбиения $\hat{V}^*$ с орбитами графа.

Для каждой пары вершин $v,u \in V$ из одного и того же класса разбиения $\hat{V}^*$ проверяется существование автоморфизма (подстановки) переводящего v в u . Эта проверка производится путем выделения вершин v и u в качестве отдельных классов в разбиении $\hat{V}^*$ (полученные разбиения обозначим $\hat{V}^*_v$ и $\hat{V}^*_u$) и сравнения между собой разбиений $\hat{V}^*_v$ и $\hat{V}^*_u$.

Результатом работы алгоритма является список орбит графа G .

Алгоритм распознавания изоморфизма графов $G(V,X)$ и $H = (U,Y)$ [9] использует упорядоченные разбиения $\hat{V}_\Gamma = (V_1,\dots,V_n)$ и $\hat{U}_\Gamma = (U_1,\dots,U_m)$ на непересекающиеся классы V_i, $i = \overline{1,n}$ и U_j, $j = \overline{1,m}$, полученные с помощью алгоритма поиска орбит графов.

Для изоморфных графов $G \cong H$: $|V| = |U| = p$, $|X| = |Y|$, $n = m$, $|V_i| = |U_i|$, $i = \overline{1,n}$.

Возможны две ситуации:

1) $n = p$, т.е. $V_\Gamma = (v_1,\dots,v_p)$, $U_\Gamma = (u_1,\dots,u_p)$ - каждый класс в разбиениях состоит из одной вершины.

Графы G и H изоморфны тогда и только тогда, когда совпадают их матрицы смежности, определяемые нумерациями $v_1,\dots,v_p$ и $u_1,\dots,u_p$ соответственно;

2) $n < p$, т.е. существует $i \in \{1,\dots,n\}$ такое, что $|V_i| = |U_i| > 1$.

В этом случае выбираем $v \in V_i$, $u \in U_i$ и строим разбиения $(V_1,\dots,V_i \setminus v,\dots,V_n,v)$, $(U_1,\dots,U_i \setminus u,\dots,U_n,u)$, к ко-

торым применяем процедуру итеративной классификации. Получен - ные разбиения обозначим через $\hat{V}^*=(V_1^*,\ldots,V_k^*)$ и $\hat{U}^* = (U_1^*,\ldots,U_l^*)$. Если $k \neq l$ или $k = l$, но $|V_i^*| \neq |U_i^*|$ для некоторого i, то графы G и H неизоморфны. Если $k = p$, то имеем ситуацию п.1. Если $k < p$, то, заменяя n на k, повторяем процесс.

Результатом работы алгоритма является установление изо - морфности или неизоморфности графов G и H.

Алгоритм поиска общих подграфов и изоморфного вложения графов $G(V,X)$ и $H=(U,Y)$ основан на свойствах клик графа соответствий $L(W,E) = G \nabla H$. Каждая клика графа L отвечает общему порожденному подграфу в графах G и H. Поиск клик производится методом рекурсивного разбора графа [11]: в графе L выбирается некоторая вершина $w_0 \in W$ и рассматри - ваются графы $\langle\{w_0\} \cup O(w_0)\rangle$ и $L-w_0$, где $O(w_0)$ - окрестность вершины w_0. Процесс разбора каждого из этих графов продолжается до тех пор, пока рассматриваемый граф либо будет кликой, либо будет входить в состав ранее разобранного графа. В этих случаях разбор прекращается. Для всех найденных максимальных клик графа L рассматриваются соответствующие им подграфы в графах G и H. Различным кликам в L могут соответствовать изоморфные подграфы, поэтому дополнительно проверяется изоморфизм среди подграфов G и H. Результатом работы алгоритма является список пересечений (максимальных общих подграфов) графов G и H.

Задача изоморфного вложения графов сводится к задаче поиска максимальных общих подграфов заданного порядка.

Пример пересечения молекулярных графов, соответствующих пестицидам ЦИАНОКС (рис.1) и БОЛСТАР (рис. 2), приведен на рис. 3. В данном примере пересечение несвязно и состоит из двух компонент.

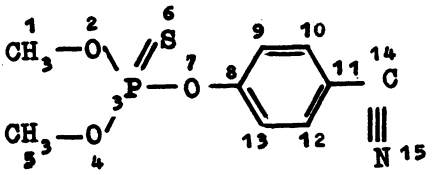

Рис.1

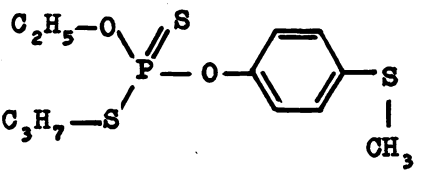

Рис. 2

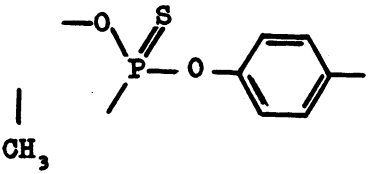

Рис. 3

Количественные характеристики (дескрипторы) графов. Введем определения некоторых количественных характеристик связного графа [7,8]. Пусть граф $G(V,X)$ имеет порядок $p = |V|$ и количество ребер $q = |X|$. Простую цепь в G, соединяющую вершины $i,j \in V$, обозначим через $C(i,j)$. Расстоянием $d(i,j)$ между i и j называют длину кратчайшей $C(i,j)$, а протяженностью $el(i,j)$ – длину дальнейшей $C(i,j)$. Цепным расстоянием $p(i,j)$ между i и j называют сумму длин всевозможных $C(i,j)$. В молекулярном графе длина любого ребра полагается равной единице.

Эксцентриситетом $e(i)$ вершины i называют величину $e(i)=\max_j d(i,j)$, цепным эксцентриситетом величину $e_\tau(i)=\max_j e1(i,j)$. Количественные характеристики графа G, вычисляемые на основе понятия дистанционного расстояния, назовем метрическими дескрипторами, а вычисляемые на основе понятий цепного расстояния и протяженности - цепными дескрипторами. Ниже приводятся наименования десяти метрических дескрипторов, их обозначения и формулы для вычисления:

- эксцентриситет графа $e(G)=\sum_{i\in V} e(i)$,
- средний эксцентриситет графа $e_{cp}(G)=\frac{1}{p}e(G)$,
- эксцентричность графа $\Delta G=\frac{1}{p}\sum_{i\in V}|e(i)-e_{cp}(G)|$,
- дистанция графа

$$D(G)=\frac{1}{2}\sum_{i\in V} D(i),\quad D(i)=\sum_{j\in V} d(i,j),$$

- средняя дистанция графа $D_{cp}(G)=\frac{2}{p}D(G)$,
- компактность графа $\mu(G)=\frac{4}{p(p-1)}D(G)$,
- среднее дистанционное отклонение графа

$$\Delta D(G)=\frac{1}{p}\sum_{i\in V}|D(i)-D_{cp}(G)|,$$

- централизация графа

$$\Delta G^*=\sum_{i\in V}(D(i)-D^*(G)),\ D^*(G)=\min_{j\in V} D(j),$$

- средняя централизация графа $\Delta G_{cp}=\frac{1}{p}\Delta G^*$,
- обратная централизация $L(G)=\sum_{i\in V}\left(\frac{1}{D^*(G)}-\frac{1}{D(i)}\right)$.

Набор цепных дескрипторов содержит одиннадцать дескрипторов. Ниже приводятся наименования, обозначения и формулы для вычисления соответствующих дескрипторов:

- цепной эксцентриситет графа $e_\tau(G)= \sum_{i\in V} e_\tau(i)$,

- средний цепной эксцентриситет графа

$$e_{\tau cp}(G) = \frac{1}{p} e_\tau(G),$$

- цепная эксцентричность графа

$$\Delta_\tau G= \frac{1}{p} \sum_{i\in V} |e_\tau(i)-e_{\tau cp}(G)|,$$

- цепная дистанция графа

$$D_\tau(G)= \frac{1}{2} \sum_{i\in V} D_\tau(i),\ D_\tau(i)= \sum_{j\in V} p(i,j),$$

- средняя цепная дистанция графа $D_{\tau cp}(G)= \frac{2}{p} D_\tau(G)$,

- цепная компактность графа $\mu_\tau(G)= \frac{4}{p(p-1)} \cdot D_\tau(G)$,

- среднее цепное дистанционное отклонение графа

$$\Delta D_\tau(G)= \frac{1}{p} \sum_{i\in V} |D_\tau(i)-D_{\tau cp}(G)|,$$

- цепная централизация графа

$$\Delta_\tau G^*= \sum_{i\in V} (D_\tau(i)-D_\tau^*(G)),\ D_\tau^*(G)=\min_{j\in V} D_\tau(j),$$

- средняя цепная централизация графа

$$\Delta_{\tau cp} G^*= \frac{1}{p} \Delta_\tau G^*,$$

- обратная цепная централизация графа

$$L(G)= \sum_{i\in V} \left(\frac{1}{D_\tau^*(G)} - \frac{1}{D_\tau(i)} \right),$$

- сложность графа

$$\xi(G) = \frac{pq}{2(p+q)} \cdot \sum_{i,j\in V} \gamma(i,j),$$

где $\gamma(i,j)$ - число различных простых цепей, соединяющих вершины i и j.

Дескрипторами связей молекулярного графа (брутто-формулы связей соединений [6]) описываются атомы соединения и типы химической связи между ними в виде ASB, где A и B - обозначения атомов, S - обозначение типа химической связи между A и B. Значением дескриптора связи является количество связей типа S между атомами A и B в соединении.

Классификация соединений. Исследование зависимости "структура-свойство" химических соединений имеет целью установление закономерностей, описывающих корреляцию свойств соединений и их структурных характеристик. Свойства соединений обучающей выборки описываются с помощью двух таблиц "объект-признак"*). Количественная таблица содержит для каждого объекта (соединения) значения его дескрипторов-признаков, а в бинарной таблице для каждого соединения указывается, какие фрагменты-признаки входят в данное соединение. В качестве фрагментов-признаков в бинарную таблицу включаются все неизоморфные компоненты попарных пересечений молекулярных графов соединений каждого класса.

Выявление закономерностей, характеризующих свойства соединений обучающей выборки, и классификация (прогноз свойств) контрольных соединений основаны на применении алгоритма построения логического решающего правила в виде двоичного дерева [4] для заданной таблицы обучающей выборки.

Пусть обучающая выборка $A = \{a_i\}$ состоит из объектов двух классов A_1 и A_2; A_1 - множество из n_1 соединений, обладающих некоторым свойством, A_2 - множество из n_2 соединений, не обладающих этим свойством, $n_1 + n_2 = n$. Обозначим таблицу "объект-признак" для обучающей выборки A через $T = \|t_{ij}\|, i = \overline{1,n}, j = \overline{1,k}$, n - число строк (объ-

*) Далее для краткости будем просто говорить таблица.

ектов), k - число столбцов (признаков), а множество признаков обозначим через $X = \{x_j\}$, $j = \overline{1,k}$. Значением t_{ij} признака x_j на объекте a_i является в количественных таблицах значение j-го дескриптора для i-го объекта, а в бинарных таблицах $t_{ij} = 1$, если j-й признак (фрагмент) содержится в i-м объекте (соединении) и $t_{ij} = 0$, если не содержится.

Назовем элементарным высказыванием $C_j(t_j^0)$ для признака x_j высказывание вида

- "значение признака $t_j > t_j^0$", t_j^0 - порог (для количественных признаков),

- "$t_j = t_j^0 = 1$" (для бинарных признаков).

Высказыванию C_j поставим в соответствие двоичную переменную $y_i \in \{0,1\}$ так, что $y_j = 1$, если C_j истинно, и $y_j = 0$, если C_j ложно. Высказыванию C вида "$C_1 \& \bar{C}_2 \& \ldots \ldots \& C_r$" поставим в соответствие конъюнкцию $q = y_1 \bar{y}_2 \ldots \ldots y_r$.

Для выбранной переменной y_j, соответствующей признаку x_j, и величины t_j^0 в таблице T можно выделить две подтаблицы: T_1, состоящую из объектов обоих классов, для которых $y_j = 1$, и T_2, состоящую из остальных объектов, для которых $y_j = 0$. Обозначим число объектов первого класса таблицы T_1 (или T_2) через m_1 (или l_1), а второго класса через m_2 (или l_2), $m_s + l_s = n_s$, $s = 1,2$. Каждую из полученных таблиц T_1 и T_2 можно таким же способом делить далее по другим переменным y_j' или по той же y_j, но для другого значения t_j^0. Процесс деления может быть завершен по достижении заданного значения некоторого критерия, например, числа объектов в одном из классов.

Результату деления исходной таблицы T по множеству переменных $Y = \{y_j\}$ соответствует двоичное дерево, а каждой

подтаблице, полученной делением T по последовательности переменных $y_{j1}, y_{j2}, \dots, y_{jr}$, соответствует вершина дерева и некоторая конъюнкция, например, $q = y_{j1}\bar{y}_{j2}\cdots y_{jr}$. Если подтаблица, соответствующая висячей вершине дерева, содержит n_s' объектов класса s, $s = 1,2$, и $n_2' = 0$ (или $n_1' = 0$), то соответствующая конъюнкция определяет высказывание, которое истинно только для n_1' объектов первого класса (или для n_2' объектов второго класса). При $n_1' \neq 0$ и $n_2' \neq 0$ полагаем, что соответствующее высказывание истинно для объектов s-го класса с вероятностью $n_s'/(n_1' + n_2')$. Совокупность конъюнкций, соответствующих всем висячим вершинам дерева, называют логическим решающим правилом (в дальнейшем решающее правило).

Выбор переменной y_j (признака x_j) для деления каждой подтаблицы производят, исходя из значения критерия качества признака, учитывающего априорные гипотезы о закономерностях в совокупности исходных объектов. В алгоритме формирования решающего правила принят критерий качества признака, учитывающий веса w значений признаков [5]. Критерий F_j признака x_j для любой подтаблицы имеет вид:

$$F_j = \max[w_1(\mu_1 - c\mu_2), w_0(\lambda_1 - c\lambda_2)],$$

где $\mu_s = \frac{m_s}{n_s}$, $\lambda_s = \frac{l_s}{n_s} = 1 - \mu_s$, $s = 1\ 2, \dots, n_s$ – число объектов класса s подтаблицы, m_s (или l_s) – число объектов класса s, для которых $y_j = 1$ (или 0); w_1 (или w_0) – вес тех значений признака x_j, для которых $y_j = 1$ (или 0), c – параметр, выделяющий признаки, имеющие заданное превышение $\mu_1(\lambda_1)$ над $\mu_2(\lambda_2)$; $w_1 = 1+w, w_0 = 1-w$, $-1 \le w \le 1$, $c \ge 1$. Для деления заданной подтаблицы выбирается признак, имеющий максимальное значение $F_j > 0$, если же $F_j \le 0$ для всех x_j, то деление не производится.

Для бинарной таблицы веса значений признаков могут быть заданы из соотношения $W \geq 0$, поскольку предполагается, что данное свойство соединения определяется наличием в нем некоторого фрагмента, а не его отсутствием, хотя возможно, что для проявления этого свойства необходимо отсутствие определенного другого фрагмента. Параметр C позволяет выбрать для деления признаки, имеющие некоторые особенности, например, при $C = n_2$, $W = 1$ выбор признаков производится только среди собственных признаков первого класса подтаблицы, т.е. среди тех, у которых $m_2 = 0$. Для признаков-дескрипторов количественной таблицы задается значение $W = 0$.

В алгоритме предусмотрено формирование набора решающих правил с помощью последовательного устранения из таблицы тех признаков, которые вошли в ранее сформированные правила.

Классификация контрольных соединений производится следующим образом: для каждого соединения устанавливаются значения признаков и находится конъюнкция решающего правила, равная 1 на этих значениях, т.е. определяется принадлежность соединения одному из классов. Качество прогноза правил определяется относительной ошибкой классификации соединений контрольной выборки. Фрагменты, предположительно ответственные за проявление данного свойства, могут быть выделены экспертом как из общего множества фрагментов, так и из набора фрагментов, вошедших в решающее правило, давшее лучший прогноз.

2. Функционирование комплекса программ СИСТРАН-МГ

Комплекс программ СИСТРАН-МГ построен по модульному принципу и является системой, открытой для наращивания функциональных возможностей.

Структурными компонентами СИСТРАН-МГ являются (рис.4):

- комплекс программ подготовки, предбазовой обработки входной информации и управления базой данных (ПУБД);

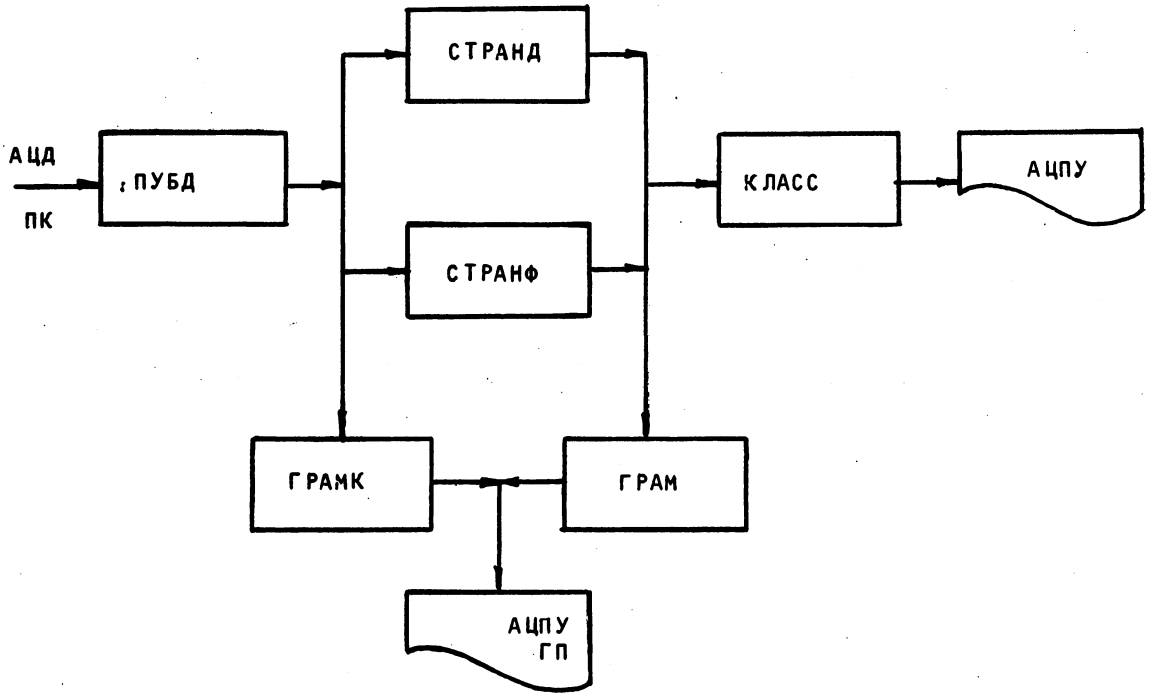

Рис. 4. Структура комплекса программ СИСТРАН-МГ

- комплекс программ структурного анализа молекулярных графов (СТРАН), предназначенный для нахождения общих фрагментов (СТРАНФ) и дескрипторов - количественных значений структурных характеристик соединений (СТРАНД);

- комплекс програми формирования логических решающих правил и классификации химических соединений (КЛАСС);

- программы графического отображения молекулярных графов (ГРАМ, ГРАМК).

Алгоритм работы комплекса программ СИСТРАН-МГ соответствует принятой методике нахождения наибольших общих фрагментов соединений [3] и прогнозирования их свойств.

Комплекс программ СИСТРАН-МГ функционирует следующим образом. В базу данных с перфокарт или АЦД вводятся данные о структурах и свойствах одного или нескольких семейств химических соединений. При проведении эксперимента по выявлению связи "структура-свойство" из базы данных средствами ПУБД извлекается обучающая и контрольная выборки соединений данного семейства. Обучающая выборка содержит представителей двух классов соединений: первый класс - соединения, проявляющие данное свойство, второй класс - непроявляющие.

По данным обучающей выборки комплекс СТРАНФ находит попарные пересечения (наибольшие общие фрагменты) молекулярных графов соединений каждого класса, определяет вхождение найденных фрагментов в соединения обоих классов и формирует бинарную таблицу для признаков-фрагментов. Комплекс СТРАНД находит значения дескрипторов связи, метрических и цепных дескрипторов и формирует таблицу для количественных признаков-дескрипторов.

Комплекс КЛАСС по обучающей таблице "объект-признак" и заданным входным параметрам формирует логические решающие правила, описывающие закономерности вхождения фрагментов или распределения значений дескрипторов в соединениях разных классов обучаю -

щей выборки, и с помощью этих правил производит классификацию (прогноз свойств) соединений контрольной выборки.

Программы ГРАМК, ГРАМ позволяют выводить на АЦПУ или графопостроитель изображения структурных формул химических соединений и их фрагментов с целью обеспечения контроля входных данных и возможности визуального анализа фрагментов.

Блок-схема комплекса программ СИСТРАН-МГ приведена на рис.5.

2.1. Входные данные. Исходная информация о химических соединениях разделена на основную и дополнительную. Основная информация состоит из номера соединений и описания его молекулярной структуры. Дополнительная информация может содержать наименования соединения, источник информации, количественные значения его свойств и т.д.

Под молекулярной структурой химического соединения понимается его структурная формула. Обозначения отдельных атомов или функциональных групп атомов назовем метками. Локализованные химические связи в структурной формуле могут иметь различную кратность - 1,2,3,..., делокализованная (ароматическая) - кратность 1,5.

Для данного химического соединения молекулярным графом назовем такой, у которого вершинам поставлены в соответствие метки атомов или групп атомов, а ребрам - веса, равные кратностям химических связей в структурной формуле соединения.

В качестве представления молекулярного графа во входном языке ОГРА принято его описание с помощью множества фрагментов (цепей и звезд графа), содержащих все его вершины и все ребра [15]. Далее описание графов на языке ОГРА будем называть топологическим кодом. Топологический код определяется при помощи следующих правил:

1) ⟨семейство графов⟩ ::= ⟨имя⟩ * ⟨граф⟩ ... ⟨граф⟩ *

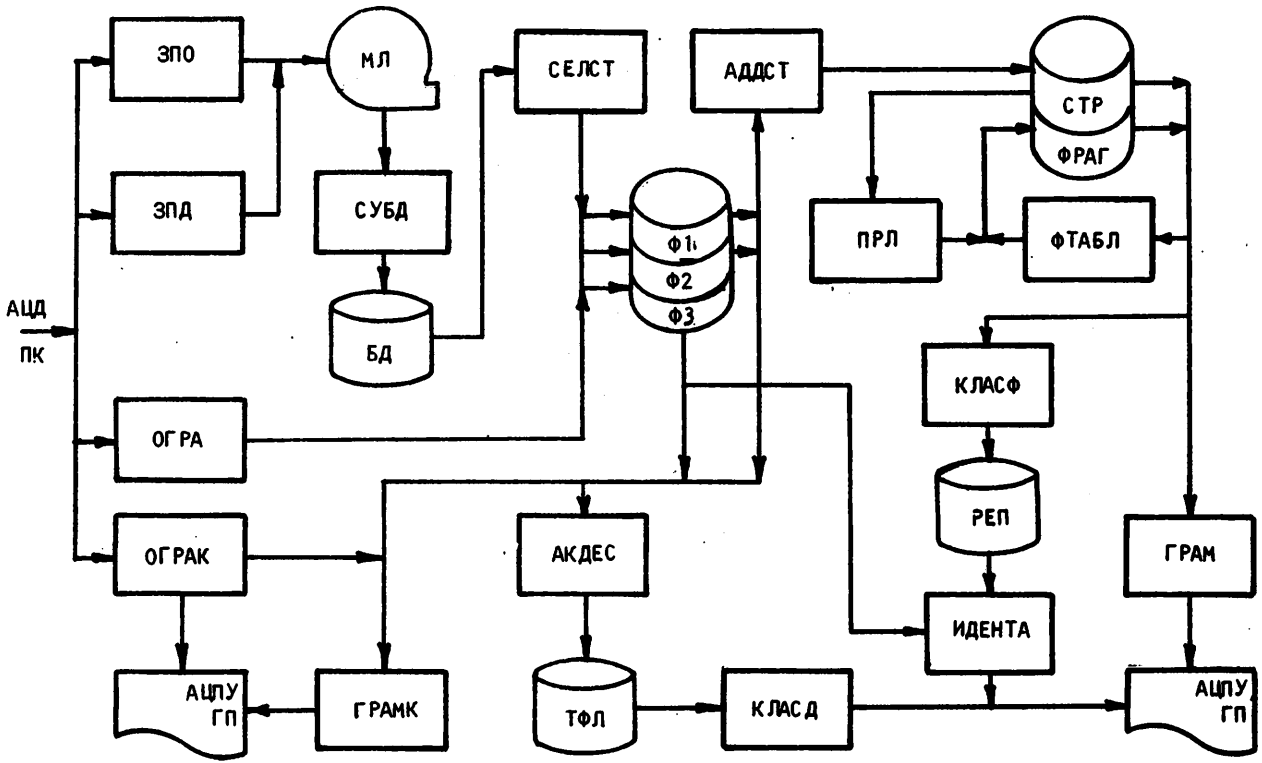

Рис. 5. Блок-схема комплекса программ СИСТРАН-МГ

Семейство графов является наибольшим структурным элемен - том языка и состоит из имени и множества входящих в это семейство графов.

2) ⟨граф⟩ ::= ⟨имя⟩ * ⟨фрагмент⟩ ... ⟨фрагмент⟩ *

Описание графа состоит из имени графа и описаний его подграфов (фрагментов).

3) ⟨имя⟩ ::= набор произвольных символов

В имени не должен встречаться символ "*"

4) ⟨фрагмент⟩ ::= ⟨цепи и звезды⟩ * | ⟨цепь⟩ * | ⟨звезда⟩ *

Под фрагментом графа понимается либо цепь (необязательно простая), либо звезда для фиксированной "звездной" вершины, либо совокупность цепей и звезд, соответствующая некоторому связному подграфу.

5) ⟨цепь⟩ ::= ⟨вершина⟩ - ... - ⟨вершина⟩

При описании цепи через символ "-" перечисляются вершины в том порядке, в котором они встречаются в цепи.

6) ⟨звезда⟩ ::= ⟨вершина⟩ - ⟨вершина⟩,..., ⟨вершина⟩

Описание звезды состоит из описания звездной вершины, разделительного символа "-" и разделенных между собой запятыми описаний вершин, смежных со звездной.

7) ⟨вершина⟩ ::= ⟨номер вершины⟩ | ⟨номер вершины⟩ (⟨вес ребра⟩) | ⟨номер вершины⟩ ⟨метка вершины⟩ | ⟨номер вершины⟩ ⟨метка вершины⟩ (⟨вес ребра⟩)

8) ⟨номер вершины⟩ ::= натуральное число

9) ⟨метка вершины⟩ ::= набор произвольных символов

Метка вершины не должна начинаться с цифры и содержать символы "*", ",", "-", "(". Метку вершины в описании графа достаточно указать только один раз.

10) ⟨вес ребра⟩ ::= вещественное число.

Вес ребра указывается в том случае, когда описываемая вершина графа соединена с предыдущей вершиной в цепи или со звездной вершиной ребром с весом, не равным единице.

При описании химических структур можно не указывать метки типа C, CH, CH_2, CH_3.

Пример описания молекулы цианокса (рис.1):

ЦИАНОКС *1-20-3P-6S(2),70,40-5*7-8-9(2)-10-11(2)-12-13(2)-
8*11-14-15N(3)**.

В языке ОГРА предусмотрены средства задания эквивалентности меток вершин.

В общем случае в топологическом коде молекулярного графа могут быть учтены физико-химические характеристики соединения, его атомов и связей. Для этого значения характеристик должны быть включены соответственно в имя, метки вершин и веса ребер графа. Тогда общие фрагменты соединений могут быть найдены с учетом не только вида атомов и связей, но и с учетом, например, зарядов атомов и расстояний между ними.

В качестве машинного представления молекулярного графа в комплексе СИСТРАН-МГ принята матрица смежности, в которой учтены веса ребер, и список меток вершин.

2.2. Ввод и поиск информации в базе данных. Основная и дополнительная информация о соединениях одного семейства вводится и хранится в двух файлах базы данных, связанных по номерам соединений. В качестве системы управления базой данных принята система СПЕКТР [13] - адаптируемая система, использующая для хранения данных метод частично инвертируемых файлов и обеспечивающая прямой доступ к записям логического файла, значения полей которых соответствуют поисковому предписанию.

В состав ПУБД из штатного математического обеспечения системы управления базой данных СПЕКТР включены базовый язык (ядро системы), макрокоманды языка СПМАКРО, программный модуль СПИНТЕР, реализующий режим пакетной обработки и утилиты сопровождения базы данных.

Работа с комплексом ПУБД включает предбазовую обработку, формирование и сопровождение базы данных, а также поиск в ней по заданным значениям полей в записях данных.

Предбазовая обработка данных состоит в преобразовании исходных данных в формат, необходимый для ввода в базу данных. Результатом предбазовой обработки являются промежуточные файлы на магнитной ленте или диске.

Для формирования и сопровождения базы данных используются утилиты системы СПЕКТР. Утилита загрузки загружает промежуточные файлы с магнитной ленты или диска в базу данных. Другие утилиты выполняют функции добавления записей в файл, связывание и модификацию файлов, отчета о состоянии базы данных и т.д.

Интерактивный язык запросов СПИНТЕР позволяет выбирать записи в соответствии с различными значениями поисковых полей. В ПУБД модуль СПИНТЕР используется для нахождения номеров соединений по различным поисковым полям в файлах дополнительной информации, в том числе по полям дескрипторов связи, и для вывода их либо на экран терминала, либо на АЦПУ.

Номера соединений являются входной информацией для прикладной программы СЕЛСТ комплекса ПУБД, которая с помощью модуля интерфейса, созданного из макрокоманд СПМАКРО, осуществляет поиск соответствующих топологических кодов в файле основной информации базы данных.

Комплекс ПУБД производит ввод входных данных двумя способами: 1) в формате, необходимом для работы комплекса СТРАН, 2) в файлы основной и дополнительной информации.

1. При вводе топологических кодов соединений с перфокарт или АЦД формирование на магнитном диске файлов Ф1, Ф2, Ф3 (для соединений первого и второго классов обучающей выборки и контрольных соединений) производится программами ОГРК, ГРАМК и ОГРА. Программа ОГРК производит синтаксический контроль и выдачу сообщений о синтаксических ошибках данных. Структурные ошибки

данных могут быть установлены при анализе графических изобра - жений молекулярных структур, полученных программой ГРАМК. Программа ОГРА преобразует топологические коды соединений в матрицы смежности молекулярных графов и записывает данные в указанные файлы.

2. Для совокупности соединений основная информация, вводимая с перфокарт или АЦД, проверяется с помощью программ ОГРК и ГРАМК и записывается программой ЗПО на магнитную ленту; дополнительная информация записывается на ленту с помощью программы ЗПД, а затем с помощью утилит системы СПЕКТР все данные переписываются с магнитной ленты в базу данных.

Программные средства комплекса ПУБД позволяют по запросу выбрать из базы данных данные о соединениях обучающей и контрольной выборок, с помощью программы СЕЛСТ записать их на диск в файлы Ф1, Ф2 и Ф3 и распечатать номера и имена соединений.

2.3. Формирование таблиц "объект-признак". Комплекс СТРАН обеспечивает построение двух таблиц "объект-признак": бинарной и количественной. Формирование бинарной таблицы для признаков-фрагментов производится в следующем порядке.

1. По данным файла Ф1 программа АДДСТ формирует и записывает в файл СТР для каждого соединения первого класса обучающей выборки матрицу смежности, список симметрий (орбит), список меток вершин.

2. По данным файла СТР программа ПРЛ для каждой пары соединений находит неизоморфные наибольшие общие фрагменты и заносит их описание в файл ФРАГ в виде матриц смежности и списков соединений, в которые входит каждый фрагмент.

3. По данным файла Ф2 программа АДДСТ формирует для каждого соединения второго класса обучающей выборки требуемую информацию (п.1) и пополняет ею файл СТР.

4. По данным файла СТР программа ПРЛ находит фрагменты соединений второго класса обучающей выборки и пополняет их описаниями (п.2) файл ФРАГ.

5. По данным файлов СТР и ФРАГ программа ФТАБЛ устанавливает вхождения фрагментов в соединения обоих классов обучающей выборки и в файл ФРАГ заносит данные в характеристический вектор каждого фрагмента: если фрагмент входит в i-е соединение файла СТР, то i-й компоненте вектора присваивается значение 1, иначе - 0.

Поскольку число фрагментов одного класса пропорционально квадрату числа его соединений, то выполнение программ ПРЛ и ФТАБЛ требует значительных затрат машинного времени, поэтому в них предусмотрены средства, позволяющие задавать длительность сеанса работы и возобновлять вычисления с точки прерывания при повторном обращении к программам.

Программы ГРАМ, ГРАМК, используя библиотеку шаблонов изображений циклических частей молекулярных графов, производят подготовку графической информации и вывод на АЦПУ или графопостроитель изображения молекулярных графов соединений и фрагментов. Программа ГРАМ использует данные файлов СТР и ФРАГ, а программа ГРАМК - данные файла Ф, в качестве которого могут быть использованы, например, Ф1, Ф2 и Ф3.

Формирование количественной таблицы "объект-признак" и занесение ее в файл ТФЛ производится программой АКДЕС в следующем порядке.

1. По данным файлов Ф1, Ф2 и Ф3 вырабатываются дескрипторы связей соединений обучающей и контрольной выборок.

2. По данным файлов Ф1 и Ф2 формируется таблица метрических и цепных дескрипторов для обучающей выборки.

3. По данным файла Ф3 формируется таблица метрических и цепных дескрипторов для контрольной выборки.

Выходными данными программы АКДЕС являются данные файла ТФЛ, состоящего из шести массивов, в которых содержатся таблицы обучающей и контрольной выборок для дескрипторов связи, метрических и цепных дескрипторов.

2.4. Построение логических и решающих правил и прогноз свойств соединений. Комплекс КЛАСС обеспечивает формирование логических решающих правил для бинарной или количественной таблиц обучающей выборки соединений и классификацию соединений контрольной выборки. Комплекс КЛАСС состоит из трех компонент (КЛАСФ, КЛАСД, ИДЕНТА) и обрабатывает таблицу "объект-признак" следующим образом.

1. При обработке бинарной таблицы обучающей выборки программа КЛАСФ по данным файлов СТР и ФРАГ формирует логические решающие правила и записывает их в файл РЕП.

По данным файлов ФЗ и РЕП программа ИДЕНТА устанавливает вхождение фрагментов в структуры контрольных соединений, производит классификацию (проверяет выполнение логических решающих правил) контрольных соединений и результаты прогноза выводит на печать в виде протокола, содержащего для каждого контрольного соединения и каждого правила результат классификации и суммарные ошибки прогноза для контрольной выборки.

2. При обработке количественной таблицы программа КЛАСД по данным файла ТФЛ формирует логические решающие правила, производит классификацию контрольных соединений и печатает результат и вероятность правильности прогноза.

3. Технические характеристики комплекса программ

Комплекс программ СИСТРАН-МГ функционирует в операционной среде ОС ЕС ЭВМ версии 6.1 в пакетном режиме.

В дополнение к техническим средствам ОС ЕС ЭВМ используются:

- алфавитно-цифровой дисплей с дисплейной станцией 7920,

- графопостроитель 9004.

В дополнение к программным средствам ОС ЕС ЭВМ используются:

- оптимизирующий транслятор ПЛ1 ОП,
- система математического обеспечения графопостроителей СМОГ ЕС ЭВМ,
- система управления базой данных СУБД СПЕКТР,
- диалоговая система ПРИМУС.

Программы комплекса реализованы на языках ПЛ1, ФОРТРАН, АССЕМБЛЕР. Минимальный необходимый объем оперативной памяти - 512 Кбайт.

Ограничения на параметры исходных данных:

- количество вершин молекулярного графа не более 50,
- количество фрагментов до 3000,
- количество соединений обучающей выборки до 400,
- количество контрольных соединений до 250,
- количество дескрипторов связи до 100.

4. Методика проведения машинных экспериментов и их результаты

Структура и функционирование комплекса программ СИСТРАН-МГ обусловлены особенностями принятой методики [3] машинного анализа зависимости "структура-свойство" химических соединений. В комплексе предусмотрены два независимых режима работы.

Первый режим работы предназначен для нахождения попарно общих фрагментов молекулярных графов соединений обучающей выборки, построения логического решающего правила, определения вхождения этих фрагментов в контрольные соединения и прогноза их свойств по найденным фрагментам. Такая последовательность процедур определяется вычислительной сложностью нахождения общих фрагментов соединений и тем, что перечень признаков-фрагментов для данной обучающей выборки заранее неизвестен.

Второй режим работы комплекса предназначен для вычисления значений дескрипторов всех соединений обучающей и контрольной выборок, построения решающего правила для соединения обучающей выборки и прогноза свойств контрольных соединений по значениям дескрипторов. Такой режим обусловлен быстродействием процедур вычисления значений дескрипторов и известным перечнем этих дескрипторов (кроме дескрипторов связи).

Таким образом, в комплексе СИСТРАН-МГ могут быть независимо получены логические решающие правила для таблиц "объект-признак" с признаками-фрагментами и признаками-дескрипторами.

Поскольку количество попарно общих фрагментов соединений обучающей выборки растет с ростом числа соединений, т.е. в общем случае обучающую выборку следует считать малой в статистическом смысле, то в комплексе программ СИСТРАН-МГ применен алгоритм построения решающего правила в виде логического решающего правила [4]. Для формирования последнего использован критерий, определяющий включение в состав его конъюнкций переменных, соответствующих общим фрагментам, при этом в правило могут войти и переменные, соответствующие редким фрагментам, т.е. таким, которые характерны для небольшого количества соединений. Управляющие параметры критерия позволяют получать наборы решающих правил и тем самым выбирать из них лучшие по качеству.

Качество решающих правил определяется относительной ошибкой классификации соединений контрольной выборки, т.е. качеством прогноза их свойств. При проведении экспериментов с комплексом СИСТРАН-МГ обычно исходное множество соединений разбивается на две выборки - обучающую и контрольную. Это диктуется, с одной стороны, вычислительной сложностью построения бинарной таблицы для всех соединений, связанной с трудоемкостью нахождения попарно-общих фрагментов, а с другой - необходимостью получения достоверной оценки качества решающего правила, что требует проверки последнего на достаточно большом количестве конт-

рольных соединений. Однако решающее правило, полученное для части исходных соединений, может иметь невысокое качество, что в основном определяется двумя факторами: присутствием в таблице неинформативных признаков и непредставительностью случайной обучающей выборки малого объема.

В качестве признаков бинарной таблицы используются фрагменты, являющиеся компонентами связности общих фрагментов соединений, поэтому в составе признаков присутствует достаточное количество фрагментов, графы которых имеют небольшой порядок. Такие фрагменты обычно не являются информативными, а значит, решающие правила, построенные на основе этих фрагментов, будут иметь низкое качество прогноза. Однако при малом объеме случайной обучающей выборки переменные, соответствующие этим фрагментам, могут войти в конъюнкции логических решающих правил, тем более что с удлинением конъюнкции уменьшается объем подвыборок. Для устранения неинформативных признаков из таблиц "объект-признак" могут быть использованы управляющие переменные программ ПРЛ (ограничение снизу порядка графов, включаемых в файл ФРАГ) и КЛАСС (устранение из таблиц произвольных неинформативных признаков, в том числе и дескрипторов).

Данное свойство исследуемого класса химических соединений может определяться разными фрагментами (или разными значениями дескрипторов), поэтому каждый класс может состоять из нескольких подклассов, в каждом из которых имеются собственные фрагменты, определяющие свойства соединений. В малой случайной обучающей выборке некоторые подклассы могут иметь недостаточное число представителей, что может привести к формированию решающих правил низкого качества.

Поэтому возможны два подхода к проведению машинных экспериментов с комплексом СИСТРАН-МГ. В первом случае из классов соединений предварительно выделяются подклассы и эксперименты проводятся независимо с каждым из подклассов. Во втором - при

проведении эксперимента организуется итеративный процесс, направленный на повышение качества логических решающих правил путем направленного изменения обучающей выборки на каждом шаге в соответствии с качеством прогноза на предыдущем шаге.

При экспериментах с комплексом программ СИСТРАН-МГ использовался следующий итеративный способ улучшения решающего правила. Пусть на очередном шаге исходная совокупность соединений разбита на две выборки - обучающую и контрольную. По обучающей выборке формируется логическое решающее правило (или их набор) и производится классификация контрольных соединений, в итоге для каждого соединения совокупности становится известен результат его классификации. Исходя из предположения, что ошибки классификации относятся к соединениям подклассов, имеющих недостаточное число представителей в обучающей выборке, для следующего шага итерации формируется новая обучающая выборка. В эту выборку наряду со всеми соединениями (или их частью) предыдущей выборки включается часть ошибочно классифицированных соединений, отобранных в соответствии с некоторым критерием (например, соединения с ошибочным результатом прогноза на большинстве из набора решающих правил). Итеративный процесс заканчивается при исчерпании ресурса машинного времени или при достижении допустимой величины ошибки прогноза полученных правил.

Возможен и другой, более трудный для реализации в комплексе СИСТРАН-МГ способ итеративного улучшения правил, который основан на последовательном формировании решающего правила, состоящего из "лучших" конъюнкций, т.е. тех, ошибка которых не превышает заданную на подмножествах исходной совокупности соединений. На очередном шаге итерации исходная совокупность соединений разбивается на обучающую и контрольную выборки. Из построенного решающего правила выбираются "лучшие" конъюнкции и включаются в результирующее правило. Соединения, значения при -

знаков которых обращают "лучшие" конъюнкции в единицу, устраняются, а оставшаяся часть соединений является исходной совокупностью для следующего шага итерации. Классификация новых соединений производится последовательным применением конъюнкций в порядке их включения в результирующее правило.

Испытания комплекса программ СИСТРАН-МГ проводились на ЭВМ ЕС 1060 с целью получения данных о качестве прогнозирования свойств химических соединений и эффективности поиска фрагментов, потенциально ответственных за их свойства.

4.1. Качество прогнозирования свойств соединений проверялось в нескольких машинных экспериментах. В экспериментах с пестицидами использованы набор фосфорсодержащих соединений ФОС (первый класс - 100 инсектицидов и второй - 53 нейтральных [17]) и набор соединений без фосфора АКР (первый класс - 37 акарицидов, второй - 23 нейтральных). Следует отметить, что в этих наборах для соединений одного класса существуют "близкие" по структуре соединения другого класса.

В экспериментах часть соединений включалась в обучающую выборку, а остальные - в контрольную выборку. Для каждой обучающей выборки сформированы по 5 логических решающих правил при следующих значениях параметров критерия качества: $C = 1,3, 5$; $W = 0; 0,5; 0,95$. Эксперименты проведены для случайной обучающей выборки (ОВ); направленно расширенной обучающей выборки за счет включения в нее ошибочно классифицируемых соединений (ОВН); и выборки, случайно расширенной из обучающей выборки до объема направленно расширенной (ОВС). Результаты приведены в табл. 1, 2. Использованы следующие обозначения: $ОВ(n_1, n_2)$, n_1 - число соединений первого класса, n_2 - число соединений второго класса $d = \frac{n_1+n_2}{N}$, N - число соединений в наборе, K - число фрагментов-признаков. Результатами

экспериментов являются относительные ошибки по контрольной выборке - средние и наименьшие по всем логическим решающим правилам.

Таблица 1

Результаты эксперимента для соединений ФОС

Обучающая выборка	Относительная ошибка, %		d	Время счета, мин
	средняя	наименьшая		
ОВ (24, 24) K = 270	32	18	0,3	65
ОВН(46,28) K = 434	13,5	2,5	0,48	83
ОВС(46,28) K = 441	19	10	0,48	89

Таблица 2

Результаты эксперимента для соединений АКР

Обучающая выборка	Относительная ошибка, %		d	Время счета, мин
	средняя	наименьшая		
ОВ(20,13) K = 146	36,5	22	0,55	28
ОВН(26,16) K = 196	23	6	0,7	37

Наименьшую ошибку дают логические решающие правила при $w > 0$, $c > 1$.

Эксперименты проведены также с наборами токсикантов у которых, данные о структурных формулах и значения их токсичности (ПДК)*) получены из справочников [18-20].

*) ПДК - предельно допустимая концентрация химического вещества в воздухе рабочей зоны (мг/м^3).

Один эксперимент проводился для набора ароматических соединений-токсикантов ТАР: первый класс - 40 соединений с ПДК = 0,5, второй класс - 34 соединения с ПДК = 5. Для случайной обучающей выборки ОВ(13,12) и направленно расширенной выборки ОВН(28,18) сформированы по 3 решающих правила для значений параметров $C = 1;5$, $W = 0;0,95$. В табл. 3 учтены и результаты прогноза по правилу "большинства" - соединение относится к тому классу, к которому его относит большинство из имеющихся логических решающих правил.

Таблица 3

Результаты эксперимента для соединений ТАР

Обучающая выборка	Относительная ошибка, %		d	Время счета, мин
	средняя	наименьшая		
ОВ(13,12) K = 70	38	30	0,33	14
ОВН(28,18) K = 138	18	13	0,62	32

Следующие эксперименты проводились для токсикантов с учетом типа взаимодействия их на организм. Исследованы три семейства ксенобиотиков: 1) метгемоглобинобразователи с ПДК = 0,1-1,0; 2) наркотики с ПДК > 10; 3) ксенобиотики, действующие на печень с ПДК = 1,1-10,0. Количества первых ксенобиотиков равно 20, вторых - 47, третьих - 16.

Для двух обучающих выборок ОВ1(15,33) и ОВ2(12,11), в которых первый класс метгемоглобинобразователи, а второй - в ОВ1 - наркотики, а в ОВ2 - ксенобиотики, действующие на печень, построены по 3 решающих правила для параметров $C = 1$, $W = 0,95$ и правило "большинства". Полученные результаты прогноза свойств оставшихся соединений приведены в табл. 4.

Таблица 4

Обучающая выборка	Относительная ошибка, %		d	Время счета, мин
	средняя	наименьшая		
OB1(15,33) K = 91	12	5	0,72	24
OB2(12,11) K = 52	25	20	0,64	11

Результаты экспериментов показывают, что комплекс программ СИСТРАН-МГ позволяет получить решающие правила с приемлемым качеством прогноза даже для выборок малого объема из семейства соединений, "близких" по структуре, но имеющих разные свойства.

4.2. Эффективность поиска фрагментов, влияющих на свойства соединений, проверена в двух экспериментах. Первый эксперимент проведен для 4-х семейств соединений из набора АКР: 1) амидины (5 соединений первого класса, 5 соединений второго); 2) гидразины (11,5); 3) А-карбаматы (8,4); 4) В-карбаматы (9,9). Разбиение на семейства соединений проведено специалистами химиками в предположении, что фрагменты, определяющие свойства соединений разных семейств, могут быть разными и свойства соединений одного семейства определяются небольшим (не более двух) количеством фрагментов.

Каждое из семейств было использовано как обучающая выборка для построения пяти логических правил при $C = 1,5$; $W = 0,95$. Количество полученных фрагментов для соответствующих семейств K = 37,88,47,57. Во множестве фрагментов каждого семейства специалистами указаны такие, которые, по их мнению, являются потенциально ответственными за свойства соединений данного семейства.

В каждом решающем правиле для каждого семейства соединений выделены короткие конъюнкции длиной не более двух, образующие усеченное логическое решающее правило.

Результаты испытания следующие: 1) усеченные решающие правила отделяют в семействах соответственно 100%, 91%, 75%, 64% соединений первого класса от соединений второго класса; 2) все указанные специалистами фрагменты содержатся в конъюнкциях усеченных решающих правил, при этом среди фрагментов, входящих в последние, присутствуют и такие, влияние которых на свойства соединений ранее известно не было. Таким образом, для большинства соединений семейств исходное предположение оказывается справедливым и фрагменты-признаки, входящие в конъюнкции решающих правил, могут быть рекомендованы для проверки их влияния на акарицидные свойства соединений.

Второй эксперимент проведен с набором из 30 соединений, принадлежащих рядам производных индола и фенилэтиламина, описанных в [16]. Эти соединения использовались как обучающая выборка ОВ(11,19),состоящая из 11 соединений первого класса (подавляющих влечение к алкоголю) и 19 нейтральных. По выборке ОВ(11,19) построены решающие правила и проведено сравнение фрагментов, входящих в решающие правила, с фрагментами, рассмотренными в [16]. Это сравнение показало, что фрагменты,указанные в [16] в качестве характерных для 1-го и 2-го классов, содержатся в совокупности фрагментов построенных логических решающих правил.

Таким образом, эти эксперименты подтверждают предположение о том, что фрагменты, потенциально ответственные за свойства соединений, входят во множество фрагментов логических решающих правил, построенных комплексом программ СИСТРАН-МГ.

4.3. Сравнительная значимость признаков-дескрипторов определялась в эксперименте с соединениями из набора АКР,заданными в качестве обучающей выборки ОВ(26,15) для построения решающих правил по всему множеству из 56 признаков: 35 дескрипторов связи, 10 метрических и 11 цепных дескрипторов. Наиболее часто в полученных решающих правилах встречаются следующие де-

скрипторы: $e(G)$, $L(G)$ - 3 раза; $\xi(G)$, ΔG^* - 2 раза. Средняя относительная ошибка классификации соединений контрольной выборки составляет 28%.

При использовании в качестве признаков только дескрипторов связи относительная ошибка составила 17%, при этом в решающее правило, как наиболее информативные, вошли дескрипторы NN, CS, NO .

Ошибки классификации контрольных соединений с помощью решающих правил, построенных только для метрических или только для цепных дескрипторов, составляют 33% и 22% соответственно. Решающее правило, построенное только для наиболее часто встречающихся дескрипторов $e(G)$, $L(G)$, $\xi(G)$, ΔG^*, дает на контрольных соединениях ошибку 17%.

Заключение

Комплекс программ СИСТРАН-МГ может быть использован для поиска зависимостей типа "структура-свойство" путем

- нахождения множества общих фрагментов в обучающей выборке, содержащей представителей разных классов соединений;
- анализа закономерностей между наличием фрагментов или распределением значений дескрипторов в выборке и свойствами соединений;
- использования полученных закономерностей (в виде логических решающих правил) для прогнозирования свойств соединений контрольной выборки.

Алгоритмы, использующие метрические свойства графов и их симметрии, позволяют значительно сократить перебор при поиске точного решения задачи нахождения пересечения графов.

Достаточно высокое быстродействие разработанных алгоритмов позволяет применять комплекс программ СИСТРАН-МГ для анализа свойств реальных семейств соединений.

Управляющие параметры критерия отбора фрагментов-признаков в логические решающие правила дают возможность получать различные варианты правил с целью выбора из них лучшего по качеству.

Фрагменты, потенциально ответственные за проявление данного свойства соединений, практически могут быть выделены из набора фрагментов, вошедших в качественные правила, и рекомендованы для использования при синтезе соединений с данными свойствами.

Применение комплекса программ СИСТРАН-МГ для прогноза свойств новых соединений может позволить снизить трудоемкость их синтеза и экспериментальной проверки свойств за счет предварительного отсеивания неперспективных соединений.

Литература

1. СТЬЮПЕР Э. и др. Машинный анализ связи химической структуры и биологической активности /Стьюпер Э., Брюггер У., Джурс П. - М.: Мир, 1982. - 233 с.

2. РОЗЕНБЛИТ А.Б., ГОЛЕНДЕР В.Е. Логико-комбинаторные методы конструирования лекарств. - Рига: Зинатне, 1984. - 395 с.

3. ЗАГОРУЙКО Н.Г., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Вопросы анализа и распознавания молекулярных структур на основе общих фрагментов //Алгоритмы анализа структурной информации. - Новосибирск. - 1984.-Вып.103: Вычислительные системы.-С.26-50.

4. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ЛБОВ Г.С. Алгоритмы обнаружения эмпирических закономерностей. - Новосибирск, Наука, 1985. - С. 46-54.

5. МАКАРОВ Л.И., СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Прогнозирование биологической активности химических соединений //Тез. докл. 5 Всесоюзн. симпозиума. Машинные методы обнаружения закономерностей. Минск, дек. 1985 г. - Минск,1985.-С. 252-253.

6. МИЩЕНКО Г.Л., ГЛАДКОВА Г.И. О поиске групп структурно-родственных соединений с помощью языка брутто-формул связей соединений //НТИ. - 1980. - № 12. - С. 14-17.

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

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

9. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Методы и алгоритмы анализа симметрий графов //Алгоритмы анализа структурной информации. - Новосибирск. - 1984. - Вып. 103: Вычислительные системы. - С. 6-26.

10. СКОРОБОГАТОВ В.А. Нахождение общих частей в семействах графов //Прикладные задачи на графах и сетях: Материалы Всесоюзн. совещания, Новосибирск, сентябрь 1980 г. - Новосибирск, 1981. - С. 117-132.

11. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. Об одном семействе схем рекурсивного разбора графов //Машинные методы обнаружения закономерностей, анализа структур и проектирования. - Новоси - бирск. - 1982. - Вып. 92: Вычислительные системы. - С. 3-49.

12. ГЭРИ Г., ДЖОНСОН Д. Вычислительные машины и трудноре - шаемые задачи. - М.: Мир, 1982. - 416 с.

13. СКОРОБОГАТОВ В.А. Относительные разбиения и слои гра - фов //Вопросы обработки информации при проектировании систем. - Новосибирск. - 1977. - Вып. 69: Вычислительные системы. - С. 3-10.

14. Специализированный комплекс телеобработки разнородных баз данных, СУБД СПЕКТР: Сб. тр. /НПО АСУ "МОСКВА". -М., 1982.

15. 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.

16. БОРИСОВ М.И., АВИДОН В.В., МУФАЗАЛОВА Т.П. Влияние структуры некоторых химических соединений на противоалкоголь - ную активность //Химико-фармацевтический журн. - 1984. - № 4. - С. 457-461.

17. МЕЛЬНИКОВ Н.Н. Пестициды. Химия, технология и применение. - М.: Химия, 1987. - 287 с.

18. БЕСПАМЯТНОВ Г.П., КРОТОВ Ю.А. Справочник. Предельно допустимые концентрации химических веществ в окружающей среде. - Л.: Химия, 1985.

19. ИЗМЕРОВ И.Ф., САНОЦКИЙ И.В., СИДОРОВ К.К. Параметры токсикометрии промышленных ядов при однократном воздействии. - М.: Медицина, 1977. - 240 с.

20. ЛАЗАРЕВ И.В. Вредные вещества в промышленности. Т.1. - Л.: Химия, 1971.

Поступила в ред.-изд.отд.
30 сентября 1988 года