ТЕОРИЯ ГРАФОВ И ЕЕ ПРИЛОЖЕНИЯ
(Вычислительные системы)
1994 год Выпуск 151

УДК 519.17:519.72:547.1:543.51

ИНФОРМАЦИОННЫЕ ИНДЕКСЫ СПЕКТРОВ МЕТРИЧЕСКИХ ХАРАКТЕРИСТИК МОЛЕКУЛЯРНЫХ ГРАФОВ*)

Ю.С.Некрасов, Э.Э.Тепфер, Ю.Н.Сухарев,
В.А.Скоробогатов, Е.В.Константинова

В в е д е н и е

Ключевая проблема при установлении спектро-структурных корреляций заключается в выборе соответствующих друг другу спектральных и структурных параметров. Обычно для описания структуры молекулы используют некоторый набор микрофрагментов, которым ставится в соответствие некоторый набор спектральных линий. Например, молекулярные фрагменты CO, OH, C=C и др. имеют в ИК-спектрах характеристические частоты поглощения. Другой подход основан на формировании интегральных структурных и спектральных характеристик в рамках теории графов и теории информации. В частности, масс-спектр представляет собой распределение N ионов по k непересекающимся подмножествам, которые характеризуются определенными значениями массовых чисел m/z. Этому распределению соответствует конечная вероятностная схема, каждое из k подмножеств которого содержит N_i ионов с вероятностью образования каждого сорта ионов $p_i = N_i/N$. Энтропия

*) Работа поддержана Российским фондом фундаментальных исследований (направления 94-08-08126, 93-03-18657) и приоритетным направлением 08 "Методы и средства химического анализа природных и промышленных объектов".

информации этого распределения рассчитывается по формуле Шеннона [1]

$$H = -\sum_{i=1}^{k} p_i \cdot \log_2 p_i \qquad (1)$$

и может использоваться в качестве информационного индекса масс-спектра [2].

По аналогичной схеме, путем разбиения множества вершин (или ребер) молекулярных графов на непересекающиеся классы по некоторому фиксированному критерию эквивалентности, могут быть получены информационно-топологические индексы, характеризующие структуру соединения [3-9]. Ранее [2] на примере серии производных ферроцена нами были установлены линейные зависимости между информационно-топологическими индексами молекулярных графов, основанными на концепции расстояния, и информационными индексами масс-спектров. Предложенные в [10] спектры метрических характеристик молекулярных графов позволяют получать новые информационно-топологические индексы молекулярных структур и провести их классификацию в терминах, применяемых при формировании информационных индексов масс-спектров [11].

Приведем некоторые понятия метрического анализа молекулярных графов [12-16], используемые в данной работе.

Пусть $G(V,E)$ - конечный, неориентированный, связный граф без петель и кратных ребер, $V(G)$ - множество вершин графа G, $|V(G)| = n$, $E(G)$ - множество ребер графа G, $|E(G)| = q$. Граф G называется молекулярным, если его вершины соответствуют атомам, а ребра - связям молекулярной структуры [17]. Под расстоянием $\rho(u,v)$ между вершинами $u,v \in V(G)$ понимается длина кратчайшей по числу ребер цепи, соединяющей вершины u и v [18]. В связном графе естественное расстояние ρ удовлетворяет аксиомам метрики [18], порождая метрическое пространство (G,ρ). Введенное расстояние индуцирует локальные (вершинные и реберные) и

интегральные метрические характеристики [7,18-20] графа $G(V,E)$, определяемые как функции от расстояний между его вершинами.

Среди локальных вершинных характеристик выделяют два различных по свойствам класса - эксцентриситетные и дистанционные характеристики. Формирование этих характеристик основано на понятиях эксцентриситета $e(v)$ и дистанции $d(v)$ вершины графа. Эксцентриситет $e(v)$ вершины v в графе G есть величина $e(v) = \max_{u \in V(G)} \rho(v,u)$. Дистанция $d(v)$ вершины v графа G (центральность вершины) определяется выражением $d(v) = \sum_{u \in V(G)} \rho(v,u)$.

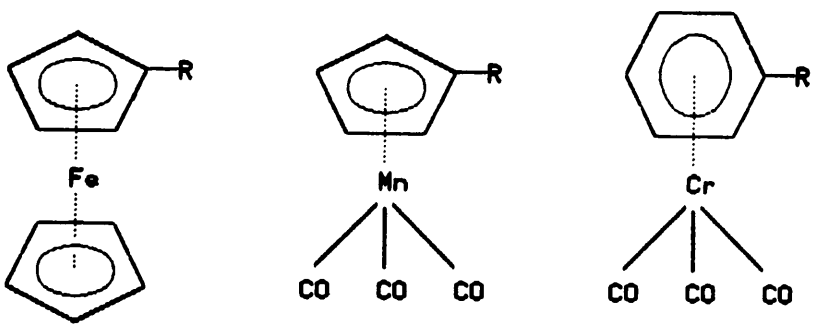
Примерами интегральных характеристик, описывающих свойства молекулярного графа в целом, являются эксцентриситет и дистанция графа. Эксцентриситетом графа G называется величина $e(G) = \sum_{v \in V(G)} e(v)$. Дистанция графа G есть сумма дистанций вершин $D(G) = \frac{1}{2} \sum_{v \in V(G)} d(v)$.

Совокупность значений эксцентриситетов вершин и дистанций вершин можно рассматривать как функции, определенные на множестве вершин молекулярного графа.

Пусть $X = (x_1, x_2, \ldots, x_n)$ и $Y = (y_1, y_2, \ldots, y_n)$ есть значения некоторых функций вершинных метрических характеристик графа $G(V,E)$, $|V(G)| = n$, и пара значений (x_i, y_i) соответствует i-й вершине графа G. Пусть (x_i, y_i) - евклидовы координаты точки, образующей вместе с $(x_i, 0)$ в системе координат (X,Y) отрезок прямой линии. Такой отрезок называют "спектральной линией" (или "спектральной полосой") i-й вершины. Пару (x_i, y_i) называют координатами или значениями спектральной полосы i-й вершины. Если $x_{i_1} = \ldots = x_{i_k}$, $k \in \overline{\{2,n\}}$, то координата y_k, $k = \{i_1, \ldots, i_k\}$, вычисляется по формуле $y_k = y_{i_1} + \ldots + y_{i_k}$ и соответствующая спектральная полоса для вершин $(i_1, \ldots, i_k)$ име-

ет кратность k. Если $y_i = 1$, $i = \overline{1,n}$, то описанное множество спектральных линий определяет спектр метрической характеристики X графа G.

В настоящей работе осуществлен теоретико-множественный анализ спектров метрических характеристик молекулярных графов трех родственных классов металлокомплексов ряда ферроцена (I), цимантрена (II) и бензолхромтрикарбонила (III) (рис.1) и предложена схема формирования по таким спектрам информационных индексов, а на их основе - классифицирующих супериндексов.



(I) (II) (III)

Рис. 1

На рис.2б в качестве примера представлен спектр дистанционной вершинной метрической характеристики молекулярного графа молекулы ферроцена (I; R = H), показанного на рис.2а. В этом случае, в силу симметрии молекулы, имеется только три класса, эквивалентных по значению дистанции $d(v_i)$, где $i = 1,\ldots,21$. Кратности n_i, $i = \overline{1,3}$, дистанций вершин отложены по оси Y.

Множеству элементов каждой метрической вершинной характеристики молекулярного графа можно также поставить в соответствие ряд распределения частот. В нем упорядоченные по возрастанию характеристики рассматриваются совместно с их кратностями,

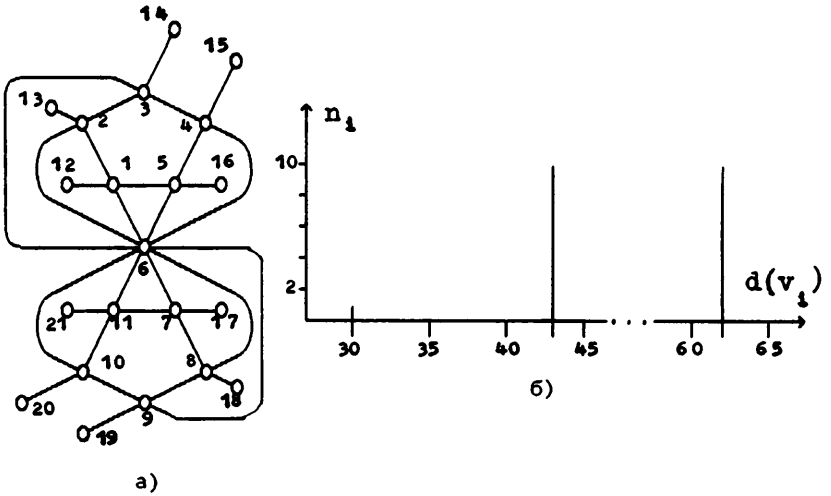


Рис. 2

а спектр метрической характеристики молекулярного графа - это график мультиномиального, в общем случае, распределения, соответствующего рассматриваемой характеристике. По оси X графика этого распределения откладываются значения метрических характеристик, а по оси Y - их кратности.

Индексы, основанные на мере множества

Общим при построении информационных индексов является следующий известный принцип.

Пусть $Y = (y_1, y_2, \ldots, y_n)$ - множество, состоящее из n элементов. Предположим, что по некоторому критерию эквивалентности элементы множества разбиваются на m непересекающихся классов Y_i так, что $n = \sum_{i=1}^{m} n_i$, где n_i - число элементов подмножества Y_i. Тогда величина $p_i = n_i/n$ характеризует вероятность некоторого элемента $y_i \in Y$ попасть в подмножество Y_i и

для количественной оценки информации, приходящейся на один элемент множества Y, можно находить (учитывая, что верхний индекс суммирования в формуле Шеннона в этом случае равен m) энтропию распределения вероятностей p_i непосредственно по формуле (1). Величины p_i здесь представляют собой отношения мощностей конечных множеств, природа элементов которых не существенна.

Знание природы элементов и/или частей некоторого множества, характеризующего спектр метрической характеристики молекулярного графа, становится необходимым в том случае, если это множество является основным. Для основного множества спектра можно ввести алгебру подмножеств и меру множеств. Использование аддитивной меры множеств (см., например, [21]) позволяет представлять Шенноновские вероятности в виде отношения мер элементов и/или частей основных множеств, что обеспечивает правомерность построения для таких множеств информационных индексов по спектру метрической характеристики молекулярного графа.

Множество $X = \{x_1, x_2, \ldots, x_n\}$ элементов x_k, $k = \overline{1,n}$, каждой вершинной характеристики молекулярного графа представляет собой мультимножество, которое характеризуется основным множеством (основанием) $X' = \{x'_1, x'_2, \ldots, x'_m\}$, т.е. множеством из m, $m \leq n$, различных элементов x'_k, $k = \overline{1,m}$, и списком кратностей $Y = [X] = [y_1, y_2, \ldots, y_m]$ элементов основания мультимножества. Поставим в соответствие множествам X' и Y, элементы которых откладываются на координатных осях спектра метрической характеристики графа и являются, по определению, вещественными числами, множество $X'Y = \{x'_1y_1, x'_2y_2, \ldots, x'_my_m\}$, элементами x'_ky_k, $k = \overline{1,m}$, которого являются произведения соответствующих друг другу координат спектра. Множество X'Y может быть либо основным, либо мультимножеством.

Пусть $Z' = \{z_1, z_2, \ldots, z_m\}$ одно из основных множеств (X' или X'Y) спектра метрической характеристики графа. В качестве

системы множеств, являющейся алгеброй, рассматриваем множество всех подмножеств основного множества $B(Z')$, $|B(Z')| = 2^{|Z'|}$, на котором вводим неотрицательную, конечную и аддитивную функцию $\mu(A)$, где $A \in B(Z')$ - некоторое подмножество системы множеств $B(Z')$.

В соответствии с приведенным описанием построения конечно-аддитивной меры μ каждой части $\{z_k\}$, $k = \overline{1,m}$, основного множества Z' спектра поставим в соответствие значения соответствующих ей элементов z_k, $k = \overline{1,m}$, этого множества. Имеем $\mu(\{z_k\}) = z_k$ и для любого подмножества $A \subset B(Z')$: $\mu(A) = \sum\limits_{z_k \in A} z_k$. В частности, $\mu(Z') = \sum\limits_{z_k \in Z'} z_k$ - конечное число, в силу конечности основного множества. Для расчета информационных индексов вводим функцию $p(A)$ (общепринятое обозначение вероятности сообщения в формуле Шеннона), определив ее равенством $p(A) = \mu(A)/\mu(Z')$. Для функции $p(A)$, при $A = Z'$, будет выполняться условие нормировки: $p(Z') = 1$, а вероятность появления элемента z_k в списке основного множества Z' спектра метрической характеристики графа определится выражением:

$$p(z_k) = \mu(\{z_k\})/\mu\{Z'\} = z_k / \sum_{k=1}^{m} z_k. \qquad (2)$$

Используя выражения (1),(2) и обозначая через H информационно-вершинный индекс, характеризующий среднее значение энтропии информации элементов списка основного множества Z', имеем:

$$H = - \sum_{k=1}^{m} p(z_k) \cdot \log_2 p(z_k) = - \sum_{k=1}^{m} (z_k / \sum_{k=1}^{m} z_k) \cdot \log_2 (z_k / \sum_{k=1}^{m} z_k). \qquad (3)$$

Вывод из H других информационных индексов по спектру метрической характеристики графа связан с характеризующими спектр следующими параметрами: числом n вершин молекулярного графа, числом m классов основного множества, числом l полос, статистиче-

скими характеристиками. Обозначая,например, через nH количество информации, приходящееся на все элементы множества Z,по которому формируется основное множество Z', имеем:

$$nH = n \cdot H. \qquad (4)$$

Нормируя информационно-вершинный индекс H на логарифм двоичный числа n вершин молекулярного графа, получаем информаци - онный индекс H^n спектра метрической характеристики графа:

$$H^n = H/\log_2 n. \qquad (5)$$

Построение информационного вершинного индекса кратности H_Y по множеству Y основано на общем принципе разбиения множества. Энтропия H_Y распределения вероятностей p_k, приходящаяся на один элемент множества Y, определится, в этом случае, выражением:

$$H_Y = -\sum_{k=1}^{m} p_k \cdot \log_2 p_k. \qquad (6)$$

Выражения (4) и (5) принимают вид:

$$nH_Y = n \cdot H_Y, \qquad (7)$$

$$H_Y^n = H_Y/\log_2 n. \qquad (8)$$

Абсциссами спектральных полос спектра являются все различные значения рассматриваемой вершинной метрической характеристики, которые можно интерпретировать как сообщение длины m над m различными буквами некоторого алфавита, значения букв которого задаются и равновероятны. В качестве основного мно - жества Z', в этом случае, рассматривается основание $X' = \{x'_1, x'_2, \ldots, x'_m\}$ ($|X'| = m$, $m \leq n$) спектра метрической характеристики графа. Исходя из специфики этого множества, учитывая выражения (3)-(5), получим семейство индексов H_X, nH_X и H_X^n основания спектра:

$$H_X = -\sum_{k=1}^{m} p(x'_k)\cdot\log_2 p(x'_k), \qquad (9)$$

$$nH_X = n\cdot H_X, \qquad (10)$$

$$H_X^n = H_X/\log_2 n. \qquad (11)$$

В качестве основного множества Z', при формировании семейства мультипликативных индексов H_{XY}, nH_{XY} и H_{XY}^n спектра метрической характеристики графа, можно рассматривать множество $X'Y = \{x'_1y_1, x'_2y_2, \ldots, x'_my_m\}$, для которого в соответствии с (3)-(5), имеем:

$$H_{XY} = -\sum_{k=1}^{m} p(x'_ky_k)\cdot\log_2 p(x'_ky_k), \qquad (12)$$

$$nH_{XY} = n\cdot H_{XY}, \qquad (13)$$

$$H_{XY}^n = H_{XY}/\log_2 n. \qquad (14)$$

Для множества X'Y, не являющегося основным, рассчитываются две группы индексов. Первую группу составляют информационные индексы, получаемые по формулам (12)-(14) для основного множества X'Y, а во вторую группу входят информационные индексы кратности элементов основания такого множества, получаемые по формулам (6)-(8).

Бинарные информационные индексы

Исходя из спектра некоторой метрической характеристики $X = (x_1, x_2, \ldots, x_n)$ молекулярного графа, можно сформировать два множества L_X и L_Y. Будем считать, что первое из них характеризует диапазон изменения значений элементов основания характеристики, а второе - диапазон изменения значений кратностей элементов основания. Обозначим через x_i, $i = \overline{1,m}$, элементы основания этой характеристики, располагаемых на оси X спектра, а через y_i, $i = \overline{1,1'}$, - кратности этих элементов, располагаемых на оси Y спектра. Здесь m и 1' соответственно числа классов,

на которые разбиваются множество элементов рассматриваемой характеристики и множество кратностей этих элементов (в соответствии с выбранными, в обоих случаях, критериями эквивалентности). Оба множества $\{x_i/i = \overline{1,m}\}$ и $\{y_i/i = \overline{1,l'}\}$ рассматриваются, таким образом, как основные.

Результат расчета элемента x_i, $i = \overline{1,n}$, метрической характеристики X представляет собой приближенное число, точность которого определяется числом значащих цифр в нем после запятой и погрешностью δ вычисления последней значащей цифры этого числа. Обозначим через δ_X погрешность вычисления последней значащей цифры элементов x_i, $i = \overline{1,m}$, основного множества рассматриваемой характеристики и будем считать, что все x_i вычисляются с одинаковым числом значащих цифр после запятой. Полагая, что $|\delta_X|$ является одним из чисел ряда 0,1; 0,01; ... , число элементов множества L_X определим выражением: $L_X = [(\max\limits_i x_i - \min\limits_i x_i) + |\delta_X|]/|\delta_X|$. Значения кратностей y_i, $i = \overline{1,l'}$, элементов основных множеств (а также значения элементов основных множеств, формируемых эксцентриситетной и дистанционной характеристиками) являются целыми положительными числами. Считая, что $|\delta_X| = |\delta_Y| = 1$ в этом случае, для числа элементов множеств L_X и L_Y имеем соответственно выражения:

$$|L_X| = (\max_i x_i - \min_i x_i) + 1;$$

$$|L_Y| = (\max_i y_i - \min_i y_i) + 1.$$

Элементы множества L_X, совпадающие по значению с элементами основания характеристики, и элементы множества L_Y, совпадающие по значению с кратностями, назовем заполненными позициями, а несовпадающие - незаполненными. Рассмотрим также множество L_{XY}, характеризующее совместный диапазон изменения элементов множеств L_X и L_Y и состоящее из заполненных и незаполненных

позиций обоих множеств L_X и L_Y. Совпадающие заполненные и/или незаполненные позиции множеств L_X и L_Y при образовании позиций множества L_{XY} будем учитывать дважды. Множества L_X, L_Y и L_{XY} могут служить основой для формирования бинарных информационных индексов.

Пусть L^1 - множество заполненных, а L^o - множество незаполненных позиций из диапазона изменений одного из множеств (L_Y, L_X или L_{XY}) рассматриваемой характеристики. Пусть L - множество всех позиций в рассматриваемом диапазоне, мощность $|L|$ которого обозначим через n^{ol}. Мощности $|L^1|$, $|L^o|$ множеств L^1 и L^o обозначим через n^1 и n^o соответственно.

Множество L можно рассматривать как сообщение, использующее алфавит s,состоящий из двух букв {1,0}, $|s| = 2$, и имеющее длину n^{ol}. Так как при $|s| = 2$ вероятности букв алфавита $p_1+p_2 = 1$, то положив $p_1 = p$ и $p_2 = 1-p$, где $p = n^1/n^{ol}$, а $(1-p) = n^o/n^{ol}$, для энтропии информации, приходящейся на элемент данного сообщения L, а следовательно, и для бинарного информационно-вершинного индекса спектра метрической характеристики молекулярного графа, получаем выражение:

$$H^b = - p\cdot\log_2 p - (1-p)\cdot\log_2(1-p) =$$

$$= - \frac{n^1}{n^{ol}}\cdot\log_2\frac{n^1}{n^{ol}} - \frac{n^o}{n^{ol}}\cdot\log_2\frac{n^o}{n^{ol}} .$$

Бинарный индекс достигает максимального значения при $p = 1/2$, равного одному биту, поэтому нормированный бинарный индекс

$$H^{nb} = \frac{H^b}{\log_2|s|} = \frac{H^b}{\log_2 2} = H^b$$

совпадает по значению с бинарным индексом.

В соответствии с предложенной схемой на основе множеств L^{l}_{Y}, L^{l}_{X}, L^{l}_{XY} и L^{o}_{Y}, L^{o}_{X}, L^{o}_{XY} получим серию бинарных информационных индексов спектра (кратности H^{b}_{Y}, основания H^{b}_{X} и мультипликативного H^{b}_{XY}):

$$H^{b}_{Y} = -\frac{n^{l}_{Y}}{n^{ol}_{i}}\cdot\log_2\frac{n^{l}_{Y}}{n^{ol}_{Y}} - \frac{n^{o}_{Y}}{n^{ol}_{Y}}\cdot\log_2\frac{n^{o}_{Y}}{n^{ol}_{Y}}, \qquad (15)$$

$$H^{b}_{X} = -\frac{n^{l}_{X}}{n^{ol}_{X}}\cdot\log_2\frac{n^{l}_{X}}{n^{ol}_{X}} - \frac{n^{o}_{X}}{n^{ol}_{X}}\cdot\log_2\frac{n^{o}_{X}}{n^{ol}_{X}}, \qquad (16)$$

$$H^{b}_{XY} = -\frac{n^{l}_{XY}}{n^{ol}_{XY}}\cdot\log_2\frac{n^{l}_{XY}}{n^{ol}_{XY}} - \frac{n^{o}_{XY}}{n^{ol}_{XY}}\cdot\log_2\frac{n^{o}_{XY}}{n^{ol}_{XY}}. \qquad (17)$$

Таким образом, предложенная схема позволяет на основе спектра метрической характеристики молекулярного графа с использованием одной (X или XY) или обеих (XY) спектральных характеристик сформировать серию информационно-топологических индексов: H_X, H_Y, H_{XY}, характеризующих структуру молекулы. При этом, индексы H_X, H_Y и H_{XY} отражают среднее информационное содержание спектра, приходящееся на один элемент спектральной характеристики X, Y и XY соответственно. Индексы nH_X, nH_Y и nH_{XY} отражают полное информационное содержание спектра по данной характеристике. Нормированные индексы H^{n}_{X}, H^{n}_{Y} и H^{n}_{XY} представляют собой отношение H_X, H_Y, H_{XY} к максимально возможному информационному содержанию спектра и изменяются от 0 до 1.

Рассмотрим свойства этих индексов и возможности их использования для характеризации молекул на примере трех родственных классов металлоорганических соединений ряда ферроцена (I), цимантрена (II) и бензолхромтрикарбонила (III) (см.рис.1).

Обсуждение результатов

В соответствии с предложенной схемой для каждого соединения из трех классов монозамещенных производных ферроцена (I), цимантрена (II) и бензолхромтрикарбонила (III) построены по два спектра, где в качестве оснований использованы дистанции и эксцентриситеты вершин. Для каждого спектра рассчитано по 12 информационных индексов (H_i, nH_i, H_i^n и H_i^b, где i обозначает Y, X или XY): кратности H_Y(6) , основания H_X (9), мультипликативный H_{XY} (12), полные индексы nH_Y (7), nH_X (10), nH_{XY} (13), нормированные индексы H_Y^n (8), H_X^n (11), H_{XY}^n (14),бинарные индексы: кратности H_Y^b (15), основания H_X^b (16), мультипликативный H_{XY}^b (17). В табл.1, в качестве примера, приведены типы замес - тителей и значения информационных индексов H_{XYd}, H_{Xd}, H_{Yd}, H_{Yd}^n, H_{Xd}^b, nH_{Yd}, рассчитанных по спектрам метрических характеристик молекулярных графов (по оси X откладывались значения дистанций вершин) монозамещенных производных цимантрена. Здесь n - число вершин соответствующего молекулярного графа. Индекс d (или e) означает, что в качестве основания спектра использовалась дистанция (или эксцентриситет) вершины.

Таблица 1

Заместитель	n	H_{XYd}	H_{Xd}	H_{Yd}	H_{Yd}^n	H_{Xd}^b	nH_{Yd}
1	2	3	4	5	6	7	8
H	17	2.096	2.271	2.162	.529	.663	36.757
CMe_3	29	2.456	2.968	2.611	.537	.592	75.707
$C(OH)Me_2$	27	2.806	3.425	2.972	.625	.729	80.244
COMe	22	3.106	3.283	3.197	.717	.730	70.334
CN	18	3.140	3.277	3.197	.767	.821	57.550
C(Me)=CHMe	27	3.087	3.423	3.208	.675	.656	86.609
Me	20	3.106	3.284	3.209	.742	.842	64.174
CHO	19	3.140	3.281	3.221	.758	.831	61.201
NH_2	19	3.140	3.281	3.221	.758	.831	61.201

1	2	3	4	5	6	7	8
$CHMe_2$	26	3.158	3.433	3.277	.697	.768	85.192
COOH	20	3.217	3.410	3.284	.760	.737	65.684
Pr	26	3.280	3.430	3.353	.713	.681	87.191
Ph	27	3.320	3.542	3.366	.708	.629	90.871
CH=CHCN	22	3.400	3.533	3.413	.765	.678	75.088
COOMe	23	3.290	3.547	3.414	.755	.709	78.531
$CH=CH_2$	21	3.396	3.545	3.463	.788	.787	72.729
CH_2OH	21	3.412	3.541	3.463	.788	.772	72.729
Et	23	3.370	3.551	3.469	.767	.803	79.778
CH=CHMe	24	3.439	3.665	3.542	.773	.747	85.020
OCOMe	23	3.461	3.664	3.555	.786	.754	81.776
COOEt	26	3.477	3.672	3.565	.758	.687	92.700
$CMe=CH_2$	24	3.481	3.666	3.574	.779	.819	85.774
COPr	28	3.470	3.777	3.580	.745	.705	100.243
Bu	29	3.530	3.673	3.581	.737	.660	103.861
CH(OH)Me	24	3.580	3.774	3.657	.798	.833	87.775
COEt	25	3.578	3.776	3.673	.791	.784	91.832
$CH=CHC_6H_4Me$-p	34	3.685	3.973	3.772	.741	.582	128.245
CH_2Ph	30	3.793	3.964	3.828	.780	.662	114.843
CH_2CH_2Ph	33	3.791	4.052	3.856	.764	.618	127.245
CH(OH)Et	27	3.770	3.971	3.856	.811	.817	104.117
COPh	29	3.880	3.969	3.909	.805	.680	113.373
COOPh	30	3.905	4.053	3.948	.805	.662	118.452
CH(OH)Ph	31	3.902	4.052	3.954	.798	.670	122.562
COBu	31	3.851	4.061	3.954	.798	.713	122.562
$C(OH)Ph_2$	41	3.930	4.138	4.004	.747	.539	164.152
OCOPh	30	4.006	4.139	4.057	.827	.692	121.698
CH=CHPh	31	4.077	4.217	4.131	.834	.694	128.070
$CH=CHC_6H_4Me$-m	34	4.163	4.370	4.219	.829	.745	143.463
CH=CHCOPh	33	4.213	4.363	4.271	.847	.682	140.956
COCH=CHPh	33	4.213	4.363	4.271	.847	.682	140.956
$CH=CHC_6H_4Me$-o	34	4.239	4.437	4.299	.844	.756	145.955
CH=CHCOCH=CHPh	37	4.296	4.495	4.358	.836	.626	161.239

Для каждого вектора $X = (x_1, x_2, \ldots, x_n)$ указанных типов индексов установлен вид эмпирического распределения. В соответствии с критериями $R/\overline{S}$, χ^2 и Колмогорова-Смирнова данное распределение является нормальным (см. рис. 3, где I и II - функции

плотности вероятности для индексных векторов H^b_{Yd} монозамещенных производных ферроцена и цимантрена соответственно).

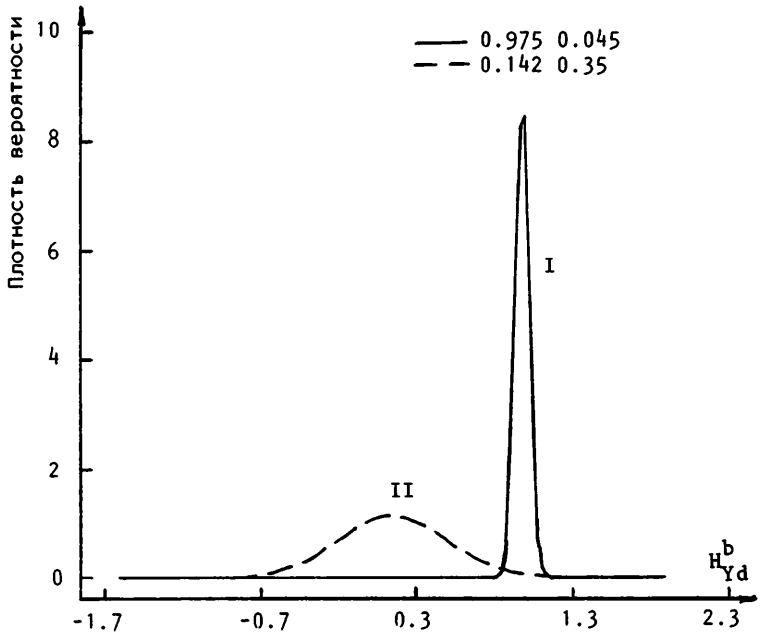


Рис. 3.

Важнейшие статистики распределения: эмпирическое среднее $\bar{x}$ (18), размах варьирования R (19), выборочные значения среднеквадратичного отклонения $\bar{S}$ (20) и коэффициента вариации ν (21), а также коэффициента вырождения α (22):

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad i = \overline{1,n}; \tag{18}$$

$$R = \max x_k - \min x_k, \quad k = \overline{1,n}; \tag{19}$$

$$\bar{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ ; \qquad (20)$$

$$\nu = \bar{S}/\bar{x}\ ; \qquad (21)$$

$$\alpha = 1 - \beta \qquad (22)$$

представлены в табл.2. В формулах (18)-(22) $\beta = \frac{1}{n}(n - \sum_k 1_k + K)$; n - число элементов индексного массива x_i, $i = \overline{1,n}$; K - число групп совпадающих индексов в массиве; $\sum_k 1_k$ - число индексов в K группах.

Анализ значений коэффициентов α (см. табл. 2, а также табл.3, в которой приведены значения коэффициента α на индексных векторах монозамещенных производных ферроцена, бензолхромтрикарбонила и цимантрена) показывает, что наименьшим вырождением обладает мультипликативный индекс H_{XY} и производные от него индексы H^n_{XY} и nH_{XY}, для которых α принимает значение равное 0.048. К маловырожденным можно отнести также все индексы кратности H_Y ($0.071 \leq \alpha \leq 0.143$), за исключением H^b_Y, и бинарные мультипликативные дистанционные индексы ($0.095 \leq \alpha \leq 0.190$). Наибольшим вырождением обладают дистанционный бинарный индекс кратности H^b_Y ($\alpha > 0.83$) и эксцентриситетный бинарный индекс основания ($\alpha = 1.0$).

Интересно отметить (табл.2), что средние значения нормированных информационных индексов располагаются внутри интервала неопределенности $Z \subset [0;1]$, содержащего отрезки $Z_1 \subset [0; 0.382]$ и $Z_2 \subset [0.382;1]$, отношения длин которых определяются гармоническим соотношением золотого сечения: $z_1/z_2 = z_2/Z$. При этом нормированные эксцентриситетные индексы группируются в окрестности точки $x_1 = 0.382 \in [0,1]$, а нормированные дистанционные индексы - в окрестности точки $x_2 = 0.618 \in [0,1]$, что, вероятно, свидетельствует о дополнительности эксцентриситетных и дистанционных индексов.

Таблица 2

№	Индекс	α	$\bar{x}$	$\bar{S}$	R	ν
1	H_{XYe}	0.048	1.787	0.321	1.345	0.179
2	H_{Xe}	0.762	2.134	0.334	1.248	0.156
3	H_{Ye}	0.071	1.894	0.327	1.381	0.172
4	H_{XYd}	0.048	3.122	0.517	2.814	0.165
5	H_{Xd}	0.048	3.562	0.492	2.833	0.138
6	H_{Yd}	0.143	3.167	0.499	2.745	0.157
7	H^{n}_{XYe}	0.048	0.361	0.049	0.204	0.136
8	H^{n}_{Xe}	0.214	0.431	0.048	0.179	0.112
9	H^{n}_{Ye}	0.071	0.383	0.049	0.207	0.129
10	H^{n}_{XYd}	0.048	0.631	0.081	0.479	0.129
11	H^{n}_{Xd}	0.048	0.721	0.074	0.471	0.103
12	H^{n}_{Yd}	0.143	0.640	0.077	0.462	0.120
13	H^{b}_{XYe}	0.643	0.973	0.028	0.082	0.029
14	H^{b}_{Xe}	1.000	0.000	0.000	0.000	0.000
15	H^{b}_{Ye}	0.690	0.949	0.057	0.278	0.060
16	H^{b}_{XYd}	0.190	0.650	0.062	0.233	0.096
17	H^{b}_{Xd}	0.248	0.580	0.052	0.227	0.089
18	H^{b}_{Yd}	0.857	0.975	0.045	0.278	0.046
19	nH_{XYe}	0.048	56.631	19.478	77.969	0.344
20	nH_{Xe}	0.214	67.477	21.748	82.474	0.322
21	nH_{Ye}	0.071	59.990	20.201	81.189	0.337
22	nH_{XYd}	0.048	98.702	31.388	138.996	0.318
23	nH_{Xd}	0.048	112.111	32.692	146.611	0.292
24	nH_{Yd}	0.119	100.054	31.250	140.029	0.312

Таблица 3

Класс	Тип индекса	Характеристика					
		e			d		
		XY	X	Y	XY	X	Y
I	H	0.048	0.762	0.071	0.048	0.048	0.143
	H^n	0.048	0.214	0.071	0.048	0.048	0.143
	nH	0.048	0.214	0.071	0.048	0.048	0.119
	H^b	0.643	1.0	0.690	0.190	0.238	0.857
III	H	0.048	0.786	0.071	0.048	0.048	0.095
	H^n	0.048	0.214	0.071	0.048	0.048	0.095
	nH	0.048	0.214	0.071	0.048	0.048	0.095
	H^b	0.571	1.0	0.742	0.095	0.190	0.881
II	H	0.048	0.762	0.095	0.048	0.048	0.095
	H^n	0.048	0.214	0.095	0.048	0.048	0.095
	nH	0.048	0.214	0.095	0.048	0.048	0.095
	H^b	0.643	1.0	0.738	0.095	0.048	0.833

В табл.4 приведены диапазоны изменения R_{min} и R_{max} индексных векторов и рассчитанные значения коэффициентов взаимного вложения K^i_j, i,j = I,II,III, одноименных пар этих векторов для трех исследованных классов металлокомплексов: бензолхромтрикарбонила (a), цимантрена (b) и ферроцена (c). Коэффициент K рассчитывался по схеме, учитывающей все случаи взаимного расположения на координатной прямой значений элементов одноименных индексных векторов и изменяется от 0 до 100%. Он отражает степень перекрывания диапазонов индексных векторов и является простейшей характеристикой, идентифицирующей способности индекса. Однозначное отнесение, по величине индекса H, исследуемого вещества к одному из двух классов соединений (например к I или II) воз -

Таблица 4

Индекс	R^{III}_{min}	R^{III}_{max}	K^{III}_{II}%	R^{II}_{min}	R^{II}_{max}	K^{II}_{I}%	R^{I}_{min}	R^{I}_{max}	K^{I}_{III}%
	a			b			c		
H_{XYe}	1.15	2.52	97	1.17	2.55	93	1.14	2.49	97
H_{Xe}	1.53	2.78	83	1.28	2.78	83	1.53	2.78	100
H_{Ye}	1.24	2.64	97	1.26	2.66	94	1.23	2.61	97
H_{XYd}	2.05	4.36	95	2.10	4.30	58	1.13	3.95	59
H_{Xd}	2.27	4.55	97	2.27	4.50	70	1.52	4.36	69
H_{Yd}	2.11	4.41	96	2.16	4.36	58	1.23	3.97	58
H^{n}_{XYe}	0.27	0.48	87	0.29	0.49	77	0.26	0.46	89
H^{n}_{Xe}	0.27	0.53	100	0.27	0.53	69	0.34	0.52	69
H^{n}_{Ye}	0.29	0.50	87	0.31	0.51	77	0.28	0.49	89
H^{n}_{XYd}	0.48	0.82	90	0.51	0.84	40	0.26	0.74	45
H^{n}_{Xd}	0.53	0.86	91	0.56	0.87	50	0.35	0.82	55
H^{n}_{Yd}	0.50	0.83	87	0.53	0.85	37	0.28	0.74	44
H^{b}_{XYe}	0.86	1.00	62	0.78	1.00	37	0.92	1.00	60
H^{b}_{Xe}	0.00	0.00	-	0.00	0.00	-	0.00	0.00	-
H^{b}_{Ye}	0.76	1.00	80	0.81	1.00	68	0.72	1.00	85
H^{b}_{XYd}	0.58	0.86	78	0.60	0.91	38	0.52	0.75	50
H^{b}_{Xd}	0.52	0.78	75	0.54	0.84	32	0.44	0.67	44
H^{b}_{Yd}	0.00	1.00	99	0.00	0.99	27	0.72	1.00	27
nH_{XYe}	21.91	98.28	92	19.85	94.45	86	23.96	101.93	92
nH_{Xe}	29.08	109.52	90	26.02	104.43	82	32.14	114.61	90
nH_{Ye}	23.65	102.77	92	21.49	98.37	85	25.80	106.99	92
nH_{XYd}	38.97	169.85	91	35.62	161.14	90	23.78	162.77	85
nH_{Xd}	43.14	177.86	91	38.61	169.64	89	31.99	178.60	92
nH_{Yd}	40.18	172.31	91	2.16	4.36	58	25.80	165.83	86

можно, если коэффициент K_{II}^{I} для этого индекса равен нулю. Более строгим критерием идентифицирующей способности индекса является степень перекрывания функций плотности вероятности индексных векторов. На рис.3 приведены графики таких функций для случая минимального взаимного вложения исследованных пар одноименных индексных векторов H_{Yd}^{b} для I и II классов. Из рисунка видно, что графики функций перекрываются (с K_{I}^{II} = 27%) и, следовательно, в качестве классифицирующего индекса, т.е индекса, позволяющего однозначно отнести данное соединение к одному из трех рассматриваемых классов, например, по критерию неперекрывания размахов варьирования рассматриваемых индексных векторов или непересечения их функций плотности вероятности, нельзя принять ни один из типов представленных в табл.4 индексов.

Вместе с тем, рассмотренные индексы могут служить основой для построения классифицирующего супериндекса. Для этого необходимо отобрать базовый набор нескоррелированных индексов. С этой целью проведен регрессионный анализ всего массива полученных индексных векторов.

По степени линейной связи между рассмотренными парами индексных векторов можно выделить три группы индексов. Высокий коэффициент корреляции ($r > 0.95$) и, следовательно, близкая к линейной связь (первая группа) существует в следующих группах индексов:

$$H_{XYe} = -0.068 + 0.979H_{Ye} = -0.162 + 0.913H_{Xe};$$

$$H_{XYe}^{n} = -0.019 + 0.992H_{Ye}^{n} = -0.038 + 0.924H_{Xe}^{n};$$

$$nH_{XYe} = -1.185 + 0.964nH_{Ye} = -3.057 + 0.885nH_{Xe};$$

$$H_{Ye}^{n} = 0.102 + 0.148H_{Ye} = 0.251 + 0.002nH_{Ye};$$

$$H_{XYd} = -0.151 + 1.033H_{Yd} = -0.592 + 1.043H_{Xd};$$

$$H^{n}_{XYd} = -0.046 + 1.057H^{n}_{Yd} = -0.144 + 1.075H^{n}_{Xd};$$

$$nH_{XYd} = -1.725 + 1.004nH_{Yd} = -8.750 + 0.958nH_{Xd};$$

$$H^{n}_{Yd} = 0.168 + 0.149H_{Yd}.$$

Вторая группа индексов характеризуется слабой линейной связью с коэффициентами корреляции 0.90 - 0.95:

$$H^{n}_{Ye} = -0.026 + 0.948H^{n}_{Xe};$$

$$H_{Yd} = 1.670 + 0.015nH_{Yd};$$

$$H_{Yd} = 0.553 + 1.380H_{Ye}.$$

Нескоррелированные и слабоскоррелированные между собой пары индексных векторов ($|r| \leq 0.90$) составляют третью группу. Сюда входят пары индексных векторов: H_i, H^{n}_i, nH_i, где i = Y, X или XY, с одной стороны, и все бинарные - с другой. К этой же группе относятся пары индексных векторов состоящих из бинарных индексов.

В табл.5 даны значения коэффициента корреляции r, полученные из уравнения $Y = a+bX$ при рассмотрении парных зависимостей между одноименными индексными векторами рассматриваемых классов соединений. При установлении связи между классами (I)-(II) и (I)-(III) в качестве независимой переменной (X) рассматривались индексные векторы класса (I), а в случае связи (II)-(III) - индексные векторы класса (II). Значения индексов в каждом массиве являются фиксированными переменными и представляют собой выборки из двумерного нормального распределения.

Из табл.5 следует, что большинство зависимостей между одноименными индексными векторами разных классов соединений характеризуются высокими коэффициентами корреляции от 0.95 до 1.00. При этом можно отметить следующие отчетливо просматриваемые тенденции:

Таблица 5

Вид связи	Тип индекса	Характеристика					
		e			d		
		XY	X	Y	XY	X	Y
I-II	H	1.00	0.90	1.00	0.90	0.93	0.89
	H^n	1.00	0.80	1.00	0.75	0.82	0.72
	nH	1.00	0.98	1.00	0.98	0.99	0.98
	H^b	0.83	-	0.84	0.80	0.76	-0.28
I-III	H	1.00	1.00	1.00	0.94	0.97	0.94
	H^n	1.00	1.00	1.00	0.84	0.91	0.83
	nH	1.00	1.00	1.00	0.99	0.99	0.99
	H^b	0.72	-	0.92	0.88	0.85	-0.31
II-III	H	1.00	0.90	1.00	0.95	0.97	0.95
	H^n	1.00	0.80	1.00	0.90	0.92	0.90
	nH	1.00	0.98	1.00	0.99	0.99	0.99
	H^b	0.87	-	0.86	0.94	0.90	0.58

1) эксцентриситетные индексы характеризуются более высокими r, по сравнению с соответствующими дистанционными индексами;

2) в группе индексов H, H^n, nH и H^b наиболее высокие r имеют индексы nH, характеризующие полное информационное содержание спектра метрической характеристики графа, наименьшее значения r имеют бинарные информационные индексы H^b;

3) среди эксцентриситетных индексов основания (X), кратности (Y) и мультипликативного (XY) наименьшим r характеризуются индексы основания H_{Xe}, H^n_{Xe} и nH_{Xe}, тогда как аналогичные дистанционные индексы H_{Xd}, H^n_{Xd} и nH_{Xd}, напротив, характеризуются высокими значениями r.

Наблюдаемые закономерности свидетельствуют о том, что эксцентриситетные индексы хорошо отражают строение заместите - лей и слабо зависят от структуры общего фрагмента молекулы. Напротив, значения дистанционных индексов в большей степени от - ражают строение ядра молекулы и его взаимосвязь со структурой заместителей R.

Проведенный статистический анализ всего множества исследованных индексных векторов позволяет выделить такие пары векторов (например H_{Yd} и H^{b}_{Xd}), значения вариант в каждой из компо - нент которых нормально распределены, имеют низкий уровень вырождения α (табл.3) и которые, в силу некоррелируемости между собой, могут рассматриваться как реализации независимых случайных величин. Это позволяет использовать такие индексы в каче - стве базовых для формирования множества супериндексных векто - ров и априорного алфавита классов на основе классифицирующего правила в виде линейной комбинации выделенной пары индексных векторов каждого из исследованных классов металлокомплексов:

$$(H^{SI})_i = \sum_{j=1}^{2} a_j \cdot H^i_j, \quad i = \overline{1,m}, \qquad (23)$$

где m - число супериндексных классов, a_j - оцениваемые коэффициенты, $H^i_1 = (H_{Yd})_i$ $H^i_2 = (H^b_{Xd})_i$ - дистанционный индекс кратности и бинарный дистанционный индекс r-го класса соответственно.

Для случая $a_1 = a_2 = [(R_{H^b_{Xd}})_i]^6 \cdot (R_{H_{Yd}})_i \cdot 100$, $i = \overline{1,3}$, в табл.6 приведены рассчитанные по формуле (23) значения статистик супериндексных векторов. Здесь $(R_{H^b_{Xd}})_i$ и $(R_{H_{Yd}})_i$ размахи варьирования индексных векторов H^b_{Xd} и H_{Yd} i-го класса.

На рис.4 даны в одном масштабе графики функций плотности вероятности этих векторов.

Таблица 6

Переменная	$(H^{SI})_I$	$(H^{SI})_{III}$	$(H^{SI})_{II}$
Размер	42	42	42
Среднее	0.141	0.353	0.735
Станд.отклонение	0.019	0.034	0.077
Минимум	0.063	0.224	0.482
Максимум	0.170	0.411	0.861
Размах	0.107	0.187	0.379

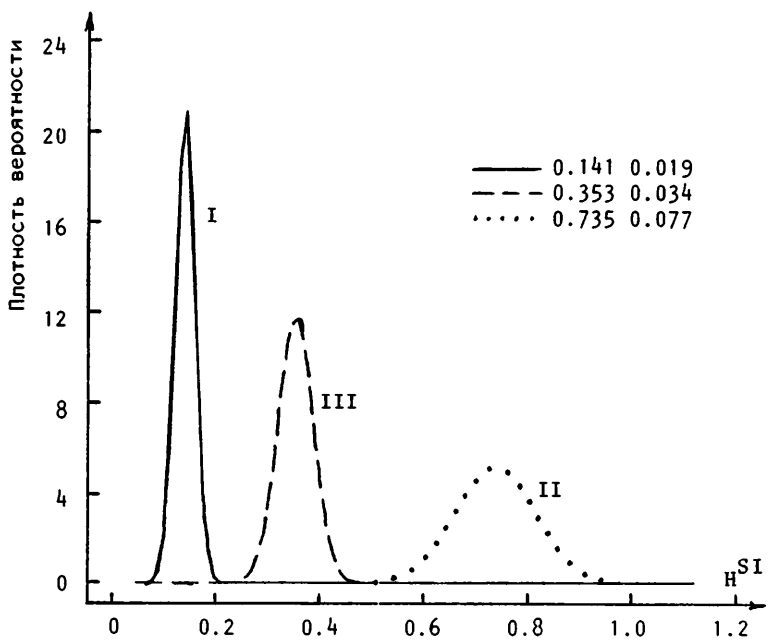


Рис. 4

Из табл.6 и рис.4 видно, что рассчитанный таким образом супериндекс может использоваться для однозначного отнесения неизвестного соединения к определенному классу веществ.

В заключение следует отметить, что теоретико-множественный анализ спектров метрических характеристик молекулярных графов и предложенная на их основе схема формирования информационных индексов и классифицирующих супериндексов уже применены нами при установлении спектро-структурных корреляций на примере масс-спектрометрии [9,11,22,23] и могут быть использованы в распознающих системах искусственного интеллекта для создания базы знаний по корреляциям структура - масс-спектр при решении прямой и обратной аналитических задач.

Литература

1. SHANNON C.E., WEAVER W. The mathematical theory of communications. - Urbana: University of Illinois, 1949.

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

3. RASHEVSKY M. Life, information theory and topology // Bull. Math. Biophys. - 1955. - Vol. 17. - P. 229-235.

4. TRUCCO E. A note of the information content of graphs //Bull. Math. Biophys. - 1956. - Vol. 18. - P.129-135.

5. BONCHEV D. Information Indices for Atoms and Molecules //Math. Chem. (MATCH). - 1979. - N 7. - P. 65-113.

6. BONCHEV D. Information theoretic indices for characterization of chemical structures. - Chichester U.K.: Researsh Studies Press. - 1983.

7. СТАНКЕВИЧ М.И., СТАНКЕВИЧ И.В., ЗЕФИРОВ Н.С. Топологические индексы в органической химии //Успехи химии. - 1988. - Т. 57. - Вып.3. - С. 337-366.

8. МАГНУСОН В., ХАРРИС Д., БЕЙСАК С. Топологические индексы, основанные на симметрии окресностей: химические и биохимические приложения // Химические приложения топологии и теории графов. /Под ред. Р.Кинг к - М.: Мир, 1987. - С.206-221.

9. SKOROBOGATOV V.A., KONSTANTINOVA E.V., NEKRASOV Yu.S., SUKHAREV Yu.N., TEPFER E.E. On the correlation between the molecular information topological and mass-spectra indices of organometallic compounds //Comm. Math. Chem. (MATCH). - 1991. - N 26. - P. 215-228.

10. МЖЕЛЬСКАЯ Е.В., СКОРОБОГАТОВ В.А. Метрические спектры молекулярных графов //Математические вопросы химической информатики. - Новосибирск, 1989. - Вып. 130: Вычислительные системы. - С. 68-83.

11. СУХАРЕВ Ю.Н., НЕКРАСОВ Ю.С.,МОЛГАЧЕВА Н.С.,ТЕПФЕР Э.Э. Идентификация веществ на основе масс-спектральных индексов //Тезисы докладов на IX Всесоюзной конференции "Химическая информатика". - Черноголовка, 1992. - С. 243.

12. ENTRIGER R.C., JACKSON D.E., SNYDER D.A. Distance in graphs //Czech. math. J. - 1976. - Vol.26.-N 2. -P. 283-296.

13. СКОРОБОГАТОВ В.А., ХВОРОСТОВ П.В. Анализ метрических свойств графов //Методы обнаружения закономерностей с помощью ЭВМ. - Новосибирск, 1981. - Вып. 91: Вычислительные системы. - С. 3-20.

14. BALABAN A.T. Numerical modelling of Chemical structu - res: local graph invariants and topological indices //Craph Theory and Topology in Chemistry (Studies in physical and theoretical chemistry, vol. 51) /Eds.King R.B. and Rouvray D.H. - Elsevier, 1987. - P. 159-176.

15. SKOROBOGATOV V.A., DOBRYNIN A.A. Metric analysis of graphs //Comm. Math. Chem. (MATCH). - 1988. - N 23.-P.105-151.

16. ДОБРЫНИН А.А., СКОРОБОГАТОВ В.А. Метрические инварианты подграфов молекулярных графов //Математические методы в химической информатике. - Новосибирск, 1991. - Вып. 104: Вычислительные системы. - С. 3-62.

17. TRINAJSTIC N. Chemical graph theory. - Florida: CRC Press, Boca Raton, 1983. - 313 p.

18. ХАРАРИ Ф. Теория графов. - М.: Мир, 1973. - 300 с.

19. ROUVRAY D.H., RANDAY R.B. The fractal nature, graph invariants, and physicochemical properties of normal alkanes //J. Chem. Phys. - 1986. - Vol.85. - N 4. - P. 2286-2290.

20. КОНСТАНТИНОВА Е.В., СКОРОБОГАТОВ В.А. Структурные и численные инварианты обыкновенных и молекулярных графов //Математические методы в химической информатике. - Новосибирск, 1991. - Вып. 140: Вычислительные системы. - С.87-129.

21. HALMOS P.R. Measure Theory. - New York: Van Nostrand, 1950.

22. НЕКРАСОВ Ю.С., ТЕПФЕР Э.Э., СУХАРЕВ Ю.С. О взаимосвязи масс-спектральных и структурных индексов арилсиланов //Изв. РАН. Сер. хим. - 1993. - №2. - С. 381-384.

23. SUKHAREV Yu.N., NEKRASOV Yu.S., MOLGACHOVA N.S., TEPFER E.E. Computer Processing and Interpretation of Mass Spectral Information. Part IX - Generalized Characteristics of Mass Spectra //Org. Mass-Spectrom. - 1993. - Vol. 28.-P.1555-1561.

Поступила в ред.-изд.отд.

30 сентярбя 1994 года