Управление, вычислительная техника и информатика

УДК 54.022;519.722;519.17

НОВЫЙ ФУНКЦИОНАЛ ИНФОРМАТИВНОСТИ ДЛЯ АНАЛИЗА СТРУКТУРЫ ХИМИЧЕСКИХ ГРАФОВ[1]

М. Дэмер, Ф. Эммерт-Штрайб*, Ю.Р. Цой**, К. Вармуза***

Институт биоинформатики и исследования процессов трансляции при Университете изучения здоровья, медицинской информатики и технологии, г. Халль, Тироль, Австрия
E-mail: Matthias.Dehmer@umit.at
*Центр исследования раковых заболеваний и клеточной биологии, Школа медицины, стоматологии и биомедицинских наук Королевского университета, г. Белфаст, Сев. Ирландия, Великобритания
E-mail: v@bio-complexity.com
**Томский политехнический университет, г. Томск, Россия
E-mail: yurytsoy@gmail.com
***Лаборатория хемометрии Институт химии Венского технологического университета, г. Вена, Австрия
E-mail: kvarmuza@email.tuwien.ac.at

Предлагается функционал информативности, основанный на степенных ассоциациях графов. Такой подход позволяет получить параметрическую меру энтропии графа, необходимую для оценки информативности структуры графа. Приведен пример, демонстрирующий вычисление предлагаемой меры.

Ключевые слова:
Меры информативности, энтропия, молекулярные графы, топологические дескрипторы, статистическое моделирование сетей, сложные сети, хемометрия.

Key words:
Information Measures; Entropy; Molecular Graphs; Topological Descriptors; Statistical Modeling of Networks; Complex Networks, Chemometrics.

Статистические и теоретико-информационные методы оценки параметров сетей представляют в настоящее время значительный интерес [1, 2] и находят применение в развитии статистических оценок корреляции, мер информативности, таких как энтропия, условная энтропия и совместная информация для структурного анализа сетей [3–7]. Классические подходы для определения структурной сложности графов химических соединений в большинстве своем основаны на использовании формулы энтропии Шеннона [3] для вывода конечного распределения вероятности, зависящего от выбранного критерия эквивалентности [8–14].

Отметим, что предлагаемые меры энтропии графов могут быть применены не только для графов химических соединений [5, 6], но и для сложных сетей произвольной природы, благодаря полиномиальной сложности вычисления этих мер, что может быть доказано аналогично подходу, описанному в [4].

1. Топологические дескрипторы и хемометрия

Развитие и эффективное использование формального представления структур химических соединений представляет собой одну из важнейших

[1] Статья основана на частичном переводе главы монографии: Dehmer M., Emmert-Streib F., Tsoy Y., Varmuza K. Quantifying Structural Complexity of Graphs: Information Measures in Mathematical Chemistry / In: M. Putz (Ed.): Quantum Frontiers of Atoms and Molecules in Physics, Chemistry, and Biology. – Hauppauge, NY: Nova Science Publishers, 2010. – P. 467–485.

Перевод, а также публикация рисунков и иллюстраций выполнены с разрешения издательского дома Nova Science Publishers, Inc. Перевел: Ю.Р. Цой

задач химии. Типичная молекула в органической химии (раздел химии, связанный с описанием и анализом углеродосодержащих молекул) состоит из атомов элементов, таких как углерод (C), водород (H), азот (N) и кислород (O), но могут встречаться и другие элементы. В органической молекуле могут встречаться одинарные, двойные, тройные и ароматические связи. Органическая молекула представляет собой трехмерную структуру и рассмотрения отдельных атомов и связей между ними вполне достаточно для описания молекулы [15].

Теория графов представляет мощный математический аппарат для представления молекулярных структур. В общем случае, вершины графа соответствуют атомам, а ребра – межатомным связям. Двойные и тройные связи между атомами представлены двойными и тройными ребрами между соответствующими вершинами, а ароматические связи могут быть заменены чередующимися одинарными и двойными связями в ароматическом кольце. При таком представлении атомы водорода часто не рассматриваются (безводородные структуры). Как правило, при использовании теории графов для описания структур химических веществ рассматриваются только скелеты молекул, в которых все вершины (атомы) считаются идентичными, а все ребра (связи) равными. К примеру, на рис. 1 показаны разные формы записи брутто-формулы уксусной кислоты, а также межатомные связи с участием атомов водорода, соответствующая безводородная структура и скелет молекулы.

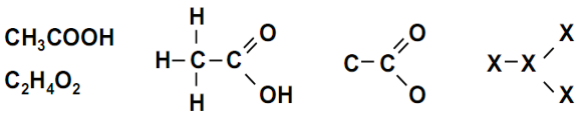

Рис. 1. *Различные представления молекулы уксусной кислоты*

Топологический дескриптор обычно является инвариантной характеристикой графа, описывающей некоторое его свойство и соответствующей структуре рассматриваемого химического вещества [16]. В течение последних десятилетий в химии предложено несколько сотен топологических дескрипторов, часть из которых достаточно абстрактны (и поэтому меньше поддержаны со стороны химического сообщества), а другие в определенной степени основаны на химических понятиях. Топологический показатель не зависит от нумерации вершин и от любого 2-мерного представления графа и может быть представлен одним числом (которое называется молекулярным топологическим показателем), либо множеством чисел. Как правило, топологические показатели вычисляются по безводородным молекулярным графам; в ряде случаев принимаются во внимание различия между атомами и связями, но также возможно и исключение влияния скелета молекулы.

Топологические показатели характеризуют структурные свойства, такие как ветвление, симметрию, форму или размер графа. Количество топологических показателей зависит от топологического расстояния между атомами (соответствующего количеству связей). Более сложные меры основаны на подсчете числа вершин и ребер или на определении изоморфизмов подграфов. Глобальные топологические показатели описывают всю структуру соединения, в то время как локальные топологические показатели учитывают отдельные атомы и соединения. Топологические информационные показатели являются особым типом топологических дескрипторов [5, 8]. При анализе молекулярного графа можно вывести распределение вероятности, с использованием определенного критерия эквивалентности, для разбиения элементов графа (например, вершин) на классы эквивалентности. Результатом будет являться мера информативности, представляющая энтропию топологии рассматриваемого графа. Такие меры несут в себе информацию о структуре графа.

В общем случае молекулярные дескрипторы – это конечный результат применения логической и математической процедуры, которая преобразует химическое символьное представление молекулы в число, либо в результат некоторых стандартизированных экспериментов [16]. Помимо топологических дескрипторов предложено большое количество других дескрипторов, характеризующих различные свойства молекулярных структур. Однако, не все дескрипторы одинаково «полезны», поскольку многие из них сильно коррелируют, а некоторые идентичны. В коммерческих приложениях [9] есть возможность вычисления более 2000 дескрипторов молекулярных структур, если даны трехмерные координаты атомов, включая все атомы водорода. В зависимости от используемых структурных данных будем различать 0-мерные дескрипторы (например, количество атомов); 1-мерные (например, величины электрических зарядов молекулы); 2-мерные (например, топологические показатели, включая меры информативности); 3-мерные (например, сумма геометрических расстояний между выбранными атомами).

Таким образом, задача описания структур химических соединений с использованием численных показателей (молекулярных дескрипторов) является одной из основных для решения базовых проблем химии, таких как создание баз данных химических формул, поиск идентичных или подобных структур химических веществ и, в особенности, для конструирования математических моделей, описывающих взаимосвязь между структурой, физическими, химическими и биологическими свойствами химических соединений.

Еще одной важной задачей химической информатики, помимо задачи описания графов химических соединений с использованием глобальных показателей, является задача сравнения структур химических соединений с применением молекулярных дескрипторов. Подобие структуры химических веществ часто выражается через сходство векторов,

состоящих из бинарных компонентов [10]. Каждый такой компонент обозначает наличие («1») или отсутствие («0») соответствующей подструктуры и называется *бинарным дескриптором подструктуры* (*binary substructure descriptor*). Примерами подструктур являются бензольное кольцо, группа сложного метилового эфира $-CH_2-COOCH_3$, либо просто определенное количество атомов азота. К настоящему времени определено большое количество подструктур (около 1000), встречающихся в структуре многих химических соединений и используемых для быстрого вычисления порядка 1000 дескрипторов для приблизительно 10000 структур химических веществ [11]. Удачной и потому распространенной мерой сравнения схожести векторов бинарных дескрипторов является показатель Танимото, известный также как коэффициент подобия Джаккарда (*Jaccard similarity coefficient*) [10, 12]. Пусть x_A и x_B являются m-компонентными векторами двух структур химических веществ A и B, соответственно, тогда показатель Танимото t вычисляется как

$$t = \frac{\sum AND(x_{A_j}, x_{B_j})}{\sum OR(x_{A_j}, x_{B_j})}, j = 1, 2, ..., m,$$

где $\Sigma AND()$ обозначает суммарное количество соответствующих единичных дескрипторов обоих векторов, а $\Sigma OR()$ – суммарное количество дескрипторов обоих векторов, когда хотя бы один из дескрипторов равен единице. Показатель Танимото принимает максимальное значение, равное 1, если все компоненты векторов дескрипторов попарно равны. В этом случае структуры A и B считаются похожими, а в некоторых случаях и идентичными. Для измерения разнообразия баз данных структур химических соединений было предложено определять распределение значений t для случайно выбранных пар структур [13].

Количественные отношения между данными структуры и физико-химическими свойствами (QSPR или QSAR модели) представляют большое значение для химической информатики [14] и разработки лекарств (*drug design*) [17, 18]. Для разработки QSPR/QSAR моделей большую роль играет хемометрика, которая использует методы анализа многомерных данных [19]. Многие разделы хемометрики рассматриваются в аналитической химии [12], основной задачей которой является «извлечение максимального объема информации из данных химического эксперимента» [20]. Основные математические и статистические методы, применяемые в хемометрике, принадлежат к анализу многомерных данных.

Одной из успешных стратегий разработки QSPR/QSAR моделей является описание структур химических веществ набором молекулярных дескрипторов (x_1, ..., x_m), представленных вектором x, и создание эмпирической регрессионной модели $\hat{y}=b_0+x^Tb$, где $\hat{y}$ – прогнозируемое значение свойства y, b – вектор коэффициентов регрессии, а b_0 – коэффициент смещения. Такие модели зависят от имеющихся данных (т. е. являются эмпирическими), поскольку для разработки и тестирования модели используется набор из n (обычно находящегося в пределах от 30 до 300) структурных единиц молекул с известными свойствами. В хемометрике широко используются методы множественной регрессии, например, PLS регрессия (регрессия по методу частичных наименьших квадратов), благодаря тому, что этот метод дает возможность настраивать сложность модели (что позволяет избежать эффекта переобучения (*overfitting*)), может работать в случае, когда количество (m) переменных превышает количество (n) объектов, и нечувствителен к сильнокоррелирующим переменным [19].

Как правило, набор дескрипторов, необходимый для построения наилучшей QSPR или QSAR модели, заранее неизвестен. В силу этого, иногда вначале выбирают большое количество (несколько сотен) потенциальных дескрипторов, а затем применяют методы отбора переменных, например, генетический алгоритм [21]. Одним из основных требований при построении моделей является «аккуратная» работа с не встречавшимися ранее данными. Для достижения этой цели используются различные методы, например, двойная кросс-валидация [22] или методы бутстраппинга [23].

Отношения между описанием структуры химических соединений и их свойствами носят очень сложный характер, и поэтому в настоящее время отсутствует общая теория, применимая во всех случаях. В силу этого разработка новых молекулярных дескрипторов представляет интерес, несмотря на то, что к настоящему времени разработано уже большое их количество. Структуры химических веществ весьма разнообразны, даже если они описываются довольно простыми графами, и потому требуют для своего описания большого количества различных молекулярных дескрипторов. Информационные показатели характеризуют внутреннюю симметрию графов (скелеты молекул), которая представляет большую важность для ряда физико-химических свойств веществ.

В больших химических базах данных, включающих значительное количество структур реальных химических соединений, топологические дескрипторы обычно не рассматриваются. Однако эти дескрипторы часто рассматриваются при анализе графов синтетических соединений, например, искусственных изомеров. В наших предыдущих работах было показано, что информационные топологические дескрипторы для реальных структур химических веществ (по данным спектрографических баз данных) существенно отличаются от таковых для искусственных изомеров [8].

2. Математические основы

Прежде чем перейти к основным определениям представим кратко известные математические понятия, более подробно освещенные в публикациях

[16, 17, 24–26]. Будем называть $G=(V,E)$[1], $|V|<\infty$, конечным неориентированным графом, если $E\subseteq C_2^V$. Граф G называется связанным, если для любых двух вершин v_i и v_j существует ненаправленный маршрут, соединяющий эти вершины. В противном случае граф называют несвязанным. Обозначим через Γ_{UC} множество всех конечных связанных графов. Степень вершины $v\in V$ графа G обозначается как $\delta(v)$ и равна количеству ребер инцидентных данной вершине.

Будем называть множество Γ_R^k классом k-регулярных графов, причем граф $G\in\Gamma_R^k$ тогда и только тогда, когда $\forall v\in V{:}\delta(v)=k$. Для данного графа $G=(V,E)\in\Gamma_{UC}$ величина $\sigma(v)=\max_{v\in V}d(u,v)$, где $d(u.v)$ обозначает кратчайшее расстояние между вершинами u и v, называется эксцентриситетом вершины $v\in V$. $d(u,v)$ должно представлять целочисленную метрику. $\rho(G)=\max_{v\in V}\sigma(v)$ называется диаметром графа. Далее, для графа $G=(V,E)\in\Gamma_{UC}$ можно определить следующий набор вершин

$$S_j(v_i,G):=\{v\in V\mid d(v_i,v)=j,\ j\geq 1\},\qquad(1)$$

называемый j-сферой вершины v_i по отношению к графу G.

Для вычисления энтропии Шеннона [3, 27] положим, что X является дискретной случайной величиной на множестве A и $p(x_i)=P(X=x_i)$ является функцией плотности вероятности X. Тогда энтропия по Шеннону определяется согласно следующему выражению:

$$H(X):=-\sum_{x_i\in A}p(x_i)\log(p(x_i)).$$

3. Энтропия графа на основе информационных функционалов

Кратко опишем основные подходы к определению энтропии графа [4]. Пусть $G\in\Gamma_{UC}$ и S является некоторым множеством, например, множеством вершин или путей. Будем называть информационным функционалом графа G отображение $f{:}S\to R_+$. Отображение f всегда полагается монотонным. Теперь, для графа $G\in\Gamma_{UC}$ определим для каждой вершины $v_i\in V$ величины

$$P_f(v_i):=\frac{f(v_i)}{\sum_{j=1}^{|V|}f(v_j)},\qquad(2)$$

где f представляет произвольный информационный функционал. Очевидно, что величины $P_f(v_i)$ могут рассматриваться как вероятности, поскольку:

$$P_f(v_1)+P_f(v_2)+\ldots+P_f(v_{|V|})=1.$$

Тогда соответствующее распределение вероятности имеет вид:

$$P_f^G(V):=(P_f^G(v_1),P_f^G(v_1),\ldots,P_f^G(v_{|V|})).$$

Принимая во внимание ур. (2) энтропия топологии графа G может быть определена как [4]:

$$I_f(G):=-\sum_{i=1}^{|V|}P_f(v_i)\log(P_f(v_i))=$$

$$=-\sum_{i=1}^{|V|}\frac{f(v_i)}{\sum_{j=1}^{|V|}f(v_j)}\log\left(\frac{f(v_i)}{\sum_{j=1}^{|V|}f(v_j)}\right).$$

4. Новый функционал информативности: Степенные ассоциации

Далее будет представлена оригинальная мера энтропии для конечных, неориентированных, связанных графов, основанная на специальном функционале информативности. Общая идея вычисления энтропии графов с использованием вершинных вероятностей (*vertex probabilities*) в зависимости от функционала информативности была представлена в статье [1]. Подобная процедура позволяет избежать проблемы определения вершинных разбиений для вычисления конечного распределения плотности вероятности. Для построения рассматриваемого функционала информативности будем использовать степенные ассоциации для кратчайших путей графа. Введем необходимые определения.

Определение 4.1. *Пусть граф $G\in\Gamma_{UC}$. Обозначим $S_j(v_j,G):=\{v_{a_j},v_{b_j},\ldots,v_{z_j}\}$, $1\leq j\leq\rho(G)$, $1\leq i\leq|V|$. Тогда для $v_i\in V$ определим наборы кратчайших путей:*

$$P_1^j(v_i):=(v_i,v_{a_1}^j,v_{a_2}^j,\ldots,v_{a_j}^j),$$
$$P_2^j(v_i):=(v_i,v_{b_1}^j,v_{b_2}^j,\ldots,v_{b_j}^j),$$
$$\vdots\qquad\qquad\vdots$$
$$P_{k_j}^j(v_i):=(v_i,v_{z_1}^j,v_{z_2}^j,\ldots,v_{z_j}^j).$$

Определение 4.2. *Пусть граф $G\in\Gamma_{UC}$. Определим следующую последовательность степеней:*

$$s_1^j(v_i):=(\delta(v_i),\delta(v_{a_1}^j),\delta(v_{a_2}^j),\ldots,\delta(v_{a_j}^j)),$$
$$s_2^j(v_i):=(\delta(v_i),\delta(v_{b_1}^j),\delta(v_{b_2}^j),\ldots,\delta(v_{b_j}^j)),$$
$$\vdots\qquad\qquad\vdots$$
$$s_{k_j}^j(v_i):=(\delta(v_i),\delta(v_{z_1}^j),\delta(v_{z_1}^j),\ldots,\delta(v_{z_j}^j)).$$

Будем называть эти последовательности строками свойств (property strings) от вершины v_i до всех других вершин графа G.

Отметим, что описанные строки свойств содержат структурную информацию о графе G.

Определение 4.3. *Пусть граф $G\in\Gamma_{UC}$. Для $v_i\in V$ определим:*

[1] В теории графов V традиционно обозначает множество вершин, а E – множество ребер, соединяющих некоторые из вершин из V. – Прим. перев.

$$\Delta^G(v_i,1) := |\,\delta(v_i)-\delta(v_{a_1}^1)\,|+\cdots+|\,\delta(v_i)-\delta(v_{z_1}^1)\,|,$$
$$\Delta^G(v_i,2) := |\,\delta(v_i)-\delta(v_{a_1}^2)\,|+\cdots+|\,\delta(v_i)-\delta(v_{z_1}^2)\,|$$
$$+\cdots+|\,\delta(v_{z_1}^1)-\delta(v_{z_2}^2)\,|,$$
$$\vdots \qquad\qquad \vdots$$
$$\Delta^G(v_i,\rho(G)) := |\,\delta(v_i)-\delta(v_{a_1}^{\rho(G)})\,|$$
$$+\cdots+|\,\delta(v_{a_{\rho(G)-1}}^{\rho(G)})-\delta(v_{a_{\rho(G)}}^{\rho(G)})\,|;$$
$$+|\,\delta(v_i)-\delta(v_{z_1}^{\rho(G)})\,|+\cdots+|\,\delta(v_{z_{\rho(G)-1}}^{\rho(G)})-\delta(v_{z_{\rho(G)}}^{\rho(G)})\,|.$$

В определении (4.3) разности вида $|\delta(x)-\delta(y)|$ называются степенными ассоциациями (*degree-degree associations*). Используя определение (4.3), включающее степенные ассоциации набора кратчайших путей, можем теперь определить параметризованный функционал информативности графа G.

Определение 4.4. *Пусть граф* $G\in\Gamma_{UC}$. *Определим функционал* $f^{\Delta}(v_i)$ *информативности:*

$$f^{\Delta}(v_i) := \alpha^{c_1\Delta^G(v_i,1)+c_2\Delta^G(v_i,2)+\cdots+c_{\rho(G)}\Delta^G(v_i,\rho(G))},$$
$$c_k>0,\ 1\le k\le\rho(G),\ \alpha>0.$$

Заметим, что при решении практических задач можно принять $\alpha=e$. Поэтому следующий функционал информативности представляет экспоненциальную функцию:

Определение 4.5. *Пусть граф* $G\in\Gamma_{UC}$ *и пусть:*

$$P_G^{f^{\Delta}}(V) := (P_{f^{\Delta}}^G(v_1), P_{f^{\Delta}}^G(v_2),\ldots,P_{f^{\Delta}}^G(v_{|V|})),$$

распределение вероятности для f^{Δ}. *Тогда энтропию графа* G *можно определить как:*

$$I_{f^{\Delta}}(G) := -\sum_{i=1}^{|V|}\frac{f^{\Delta}(v_1)}{\sum_{j=1}^{|V|}f^{\Delta}(v_j)}\log\left(\frac{f^{\Delta}(v_1)}{\sum_{j=1}^{|V|}f^{\Delta}(v_j)}\right).$$

В результате получим новое семейство мер энтропии графа. Отметим, что, варьируя значения параметров c_i и α, можно изменять веса различных структурных параметров, например, вершин с большими степенями (хабов, *hubs*), для вычисления энтропии графа G. Для более ясного понимания смысла представленной меры энтропии графа сформулируем следующую теорему и ее следствие:

Теорема 4.1: *Пусть граф* $G\in\Gamma_{UC}$. *Распределение вероятности*

$$P_G^{f^{\Delta}}(V) := (P_{f^{\Delta}}^G(v_1), P_{f^{\Delta}}^G(v_2),\ldots,P_{f^{\Delta}}^G(v_{|V|})),$$

максимизирует значение энтропии $I_{f^{\Delta}}(G)$.

Доказательство: Рассмотрим k-регулярный граф. В соответствии с определением (см. Раздел 2), условие $G\in\Gamma_{UC}$ выполняется тогда и только тогда, когда $\delta(v)=k$, $\forall v\in V$. Исходя из этого имеем:

$$s_1^1(v_i) := (k,k),$$
$$s_2^1(v_i) := (k,k),$$
$$\vdots$$
$$s_{k_1}^1(v_i) := (k,k),$$
$$s_1^j(v_i) := (k,k,\ldots,k),$$
$$s_2^j(v_i) := (k,k,\ldots,k),$$
$$\vdots$$
$$s_{k_j}^j(v_i) := (k,k,\ldots,k),$$

где $1\le j\le\rho(G)$, а $s_{\mu}^j(v_i) := (k,k,\ldots,k)$ содержит $j+1$ переменную. Тогда $\Delta^G(v_i,1)=0,\ldots,\Delta^G(v_i,j)=0$, откуда следует, что $f^{\Delta}(v_i)=\alpha^0=1$, и, окончательно,

$$P_G^{f^{\Delta}}(V) := (P_{f^{\Delta}}^G(v_1), P_{f^{\Delta}}^G(v_2),\ldots,P_{f^{\Delta}}^G(v_{|V|})) =$$
$$=\left(\frac{1}{|V|},\frac{1}{|V|},\ldots,\frac{1}{|V|}\right),$$

что и является доказательством теоремы.

Следствие 4.2: *Пусть* $K_{|И|,|И|}$ *является полносвязным графом с* $|И|$ вершинами. *Распределение вероятности*

$$P_{K_{|И|,|И|}}^{f^{\Delta}}(V) := (P_{f^{\Delta}}^{K_{|И|,|И|}}(v_1), P_{f^{\Delta}}^{K_{|И|,|И|}}(v_2),\ldots,P_{f^{\Delta}}^{K_{|И|,|И|}}(v_{|V|})),$$

максимизирует значение энтропии $I_{f^{\Delta}}(K_{|И|,|И|})$.

Интерпретация теоремы 4.1 приводит к следующему наблюдению: данное распределение вероятностей максимизирует энтропию графа G, если все его вершины топологически эквивалентны. Отсюда следует, что энтропия графа G уменьшается с ростом разнообразия характеристик соседства вершин. Отметим, что данный вывод зависит от рассматриваемой меры энтропии, к примеру, применение классической меры I_{ORB} [28] энтропии графа:

$$I_{ORB}(G) = |\,V\,|\log(|\,V\,|)-\sum_{i=1}^{k}N_i\log(N_i),$$

приводит к $I_{ORB}(K_{|И|,|И|})=0$.

5. Численный пример

Для иллюстрации данных выше определений рассмотрим пример структуры, рис. 2. В целях демонстрации произведем вычисления введенных показателей для вершины v_1, поскольку расчеты для других вершин можно сделать аналогично.

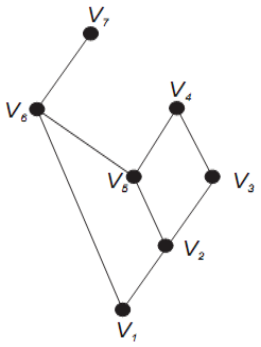

Рис. 2. *Неориентированный связанный граф*

Имеем:

$$P_1^1(v_1) := (v_1, v_2),$$
$$P_2^1(v_1) := (v_1, v_6),$$
$$P_1^2(v_1) := (v_1, v_2, v_3),$$
$$P_2^2(v_1) := (v_1, v_2, v_5),$$
$$P_3^2(v_1) := (v_1, v_6, v_5),$$
$$P_4^2(v_1) := (v_1, v_6, v_7),$$
$$P_1^3(v_1) := (v_1, v_2, v_3, v_4),$$
$$P_2^3(v_1) := (v_1, v_2, v_5, v_4),$$
$$P_3^3(v_1) := (v_1, v_6, v_5, v_4),$$
$$s_1^1(v_1) := (2,3),$$
$$s_2^1(v_1) := (2,3),$$
$$s_1^2(v_1) := (2,3,2),$$
$$s_2^2(v_1) := (2,3,3),$$
$$s_3^2(v_1) := (2,3,3),$$
$$s_4^2(v_1) := (2,3,1),$$
$$s_1^3(v_1) := (2,3,2,2),$$
$$s_2^3(v_1) := (2,3,3,2),$$
$$s_3^3(v_1) := (2,3,3,2)$$

и

$$\Delta^G(v_1,1) = 1+1 = 2,$$
$$\Delta^G(v_1,2) = 1+1+1+0+1+0+1+2 = 7,$$
$$\Delta^G(v_1,3) = 1+1+0+1+0+1+1+0+1 = 6,$$
$$\Delta^G(v_1,4) = 0.$$

В результате получим,

$$f^{\Delta}(v_1) := \alpha^{2c_1+7c_2+6c_3}.$$

Выполнив указанные шаги для всех вершин графа G, можно вычислить оригинальную меру энтропии:

$$I_{f^{\Delta}}(G) = -\frac{\alpha^{2c_1+7c_2+6c_3}}{D_G}\log\left(\frac{\alpha^{2c_1+7c_2+6c_3}}{D_G}\right) -$$
$$-\frac{\alpha^{2c_1+4c_2+6c_3}}{D_G}\log\left(\frac{\alpha^{2c_1+4c_2+6c_3}}{D_G}\right) -$$
$$-\frac{\alpha^{c_1+4c_2+5c_3+11c_4}}{D_G}\log\left(\frac{\alpha^{c_1+4c_2+5c_3+11c_4}}{D_G}\right) -$$
$$-\frac{\alpha^{c_1+3c_2+9c_3}}{D_G}\log\left(\frac{\alpha^{c_1+3c_2+9c_3}}{D_G}\right) -$$
$$-\frac{\alpha^{c_1+6c_2}}{D_G}\log\left(\frac{\alpha^{c_1+6c_2}}{D_G}\right) - \frac{\alpha^{3c_1+3c_2+5c_3}}{D_G}\log\left(\frac{\alpha^{3c_1+3c_2+5c_3}}{D_G}\right) -$$
$$-\frac{\alpha^{2c_1+5c_2+6c_3+10c_4}}{D_G}\log\left(\frac{\alpha^{2c_1+5c_2+6c_3+10c_4}}{D_G}\right),$$

где

$$D_G := \alpha^{2c_1+7c_2+6c_3} + \alpha^{2c_1+4c_2+6c_3} +$$
$$+\alpha^{c_1+4c_2+5c_3+11c_4} + \alpha^{c_1+3c_2+9c_3} +$$
$$+\alpha^{c_1+6c_2} + \alpha^{3c_1+3c_2+5c_3} + \alpha^{2c_1+5c_2+6c_3+10c_4}.$$

Будущие исследования будут направлены на исследование разработанной меры и ее применение для анализа структур химических соединений. Отметим, что объединение статистики, теории информации и теории графов обладает существенным потенциалом [29, 30], свойства которого пока мало исследованы.

Заключение

Определен новый информационный функционал, основанный на степенных ассоциациях, что привело к формулировке специальной меры информативности для графов (энтропия графа), для которой показано, что максимум энтропии достигается для полносвязной регулярной сети.

Авторы благодарят Д. Бончева и А. Моушовитц за плодотворные дискуссии. Работа поддержана COMET Center ONCOTYROL и грантами Federal Ministry for Transport Innovation and Technology (BMVIT), Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth, (BMWA/BMWFJ), Tiroler Zukunftsstiftung (TZS), а также Styrian Business Promotion Agency, (SFG), University for Health Sciences, Medical Informatics and Technology и BIOCRATES Life Sciences AG. Перевод выполнен при частичной поддержке гранта РФФИ, проект № 09-08-00309-а. Авторы благодарны Михаю Путцу, главному редактору монографии, выпущенной издательством NOVA Science Publishing, в которой опубликован оригинал статьи, за плодотворные дискуссии и поддержку при подготовке русскоязычного перевода. Авторы также выражают глубокую признательность И.А. Курзиной и М.М. Поповской за ценные замечания и комментарии к переводу.

СПИСОК ЛИТЕРАТУРЫ

1. Sole R.V., Valverde S. Information theory of complex networks: On evolution and architectural constraints // Lecture Notes in Physics. – 2004. – V. 650. – P. 189–207.
2. Schweitzer F., Ebeling W., Rose H., Weiss O. Network optimization using evolutionary strategies. // PPSN IV: Proc. of the 4th Intern. Conf. on Parallel Problem Solving from Nature. – London, UK, 1996. – Berlin: Springer-Verlag, 1996. – P. 940–949.
3. Shannon C. E., Weaver W. The Mathematical Theory of Communication. – Urbana, IL, USA: University of Illinois Press, 1997. – 144 p.
4. Dehmer M. Information processing in complex networks: Graph entropy and information functional // Applied Mathematics and Computation. – 2008. – V. 201. – P. 82–94.
5. Bonchev D. Information Theoretic Indices for Characterization of Chemical Structures. – Chichester: Research Studies Press, 1983. – 264 p.
6. Trinajstic N. Chemical Graph Theory. – Boca Raton, FL, USA: CRC Press, 1992. – 278 p.
7. Kier L.B., Hall L.H. Molecular connectivity in structure-activity analysis. – N.Y., NY, USA: Wiley, 1986. – 262 p.

8. Dehmer M., Varmuza K., Borgert S., Emmert-Streib F. On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures // Journal of Chemical Information and Modeling. – 2009. – V. 49. – № 7. – P. 1655–1663.
9. Todeschini R., Consonni V., Mauri A., Pavan M.. Dragon, software for calculation of molecular descriptors // Talette s.r.l. homepage. 2009. URL: http://www.talete.mi.it (дата обращения: 01.09.2009).
10. Willet P. Similarity and clustering in chemical information systems. – Letchworth, UK: Research Studies Press, 1987. – 266 p.
11. Varmuza K., Demuth W., Karlovits M., Scsibrany H. Binary substructure descriptors for organic compounds // Croat. Chem. Acta. – 2005. – V. 78. – P. 141–149.
12. Vandeginste B.G.M., Massart D.L.L. Buydens C.M., De Jong S., Smeyers-Verbeke J. Handbook of chemometrics and qualimetrics: Part B. – Amsterdam, The Netherlands: Elsevier, 1998. – 876 p.
13. Demuth W., Karlovits M., Varmuza K. Spectral similarity versus structural similarity: Mass spectrometry // Anal. Chim. Acta. – 2004. – V. 516. – P. 75–85.
14. Gasteiger J., Engel T. Chemoinformatics: A Textbook. – Weinheim, Germany: Wiley VCH, 2003. – 680 p.
15. Mowshowitz A. Entropy and the complexity of graphs IV: Entropy measures and graphical structure // Bull. Math. Biophys. – 1968. – V. 30. – P. 533–546.
16. Todeschini R., Consonni V., Mannhold R. Handbook of Molecular Descriptors. – Weinheim, Germany: Wiley-VCH, 2002. – 688 p.
17. Kubinyi H. Hansch Analysis and Related Approaches. – Weinheim, Germany: Wiley-VCH, 1993. – 240 p.
18. Zupan J., Gasteiger J. Neural networks in chemistry and drug design. – Weinheim, Germany: Wiley-VCH, 1987. – 305 p.
19. Varmuza K., Filzmoser P. Introduction to Multivariate Statistical Analysis in Chemometrics. – Boca Raton, FL, USA: Francis & Taylor, CRC Press, 2009. – 336 p.
20. Kowalski B.R. Chemometrics: Views and propositions // J. Chem. Inf. Comput. Sci. – 1975. – V. 5. – P. 201–203.
21. Leardi R., Norgaard L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions // J. Chemometr. – 2004. – V. 18. – P. 486–497.
22. Filzmoser P., Liebmann B., Varmuza K. Repeated couble crosss validation // J. Chemometr. – 2009. – V. 23. – P. 160–171.
23. Efron B., Tibshirani R. J. An introduction to the bootstrap. – London, UK: Chapman & Hall, 1993. – 456 p.
24. Bonchev D. Overall connectivities and topological complexities: A new powerful tool for QSPR/QSAR // J. Chem. Inf. Comput. Sci. – 2000. – V. 40. – № 4. – P. 934–941.
25. Bonchev D. The overall Wiener index – a new tool for characterization of molecular topology // J. Chem. Inf. Comput. Sci. – 2001. – V. 41. – P. 582–592.
26. Bonchev D., Trinajstic N. Overall molecular descriptors. 3. Overall zagreb indices // SAR QSAR Environ. Res. – 2001. – V. 12. – P. 213–236.
27. Cover T. M., Thomas J. A. Elements of Information Theory. – Hoboken, NJ, USA: Wiley & Sons, 2006. – 776 p.
28. Rashevsky N. Life, information theory, and topology // Bull. Math. Biophys. – 1955. – V. 17. – P. 229–235.
29. Emmert-Streib F., Dehmer M. Information Theory and Statistical Learning. – Berlin: Springer, 2008. – 389 p.
30. Pearl J. Probabilistic Reasoning in Intelligent Systems. – San Francisco, CA, USA: Morgan-Kaufmann, 1988. – 552 p.

Поступила 09.03.2010 г.