© Коллектив авторов, 2017

Е. П. Андреева, А. Н. Прошин, И.В. Серков, Л. Н. Петрова, С. О. Бачурин

ИСПОЛЬЗОВАНИЕ МОЛЕКУЛЯРНЫХ ТОПОЛОГИЧЕСКИХ ДЕСКРИПТОРОВ ДЛЯ КЛАСТЕРИЗАЦИИ БАЗЫ ДАННЫХ ПРОИЗВОДНЫХ ИЗОТИОМОЧЕВИНЫ В РАМКАХ ИЗУЧЕНИЯ ЗАВИСИМОСТИ СТРУКТУРА — АКТИВНОСТЬ

Институт физиологически активных веществ РАН, Россия, 142432, Московская обл., Черноголовка, Северный пр-д, д. 1; e-mail: admiandrew@yandex.ru

Произведена успешная кластеризация базы данных производных S,N,N,N′-тетразамещенных изотиомочевин, обладающими нейропротекторными свойствами, с целью изучения количественной связи структура — активность. Кластеризация методом *к*-средних проводилась в факторном пространстве топологических дескрипторов с последующим объединением выделенных кластеров на основе анализа внутри- и межкластерных расстояний. Первоначальное число кластеров в методе *к*-средних определялось, исходя из количества итераций, за которое было получено решение. Для оценки однородности базы данных и выделенных кластеров использовался коэффициент разнообразия. Графическое представление соединений, входящих в базу данных, в виде точек факторного пространства позволяет сделать вывод об успешности данного метода.

Ключевые слова: кластеризация; топологические дескрипторы; производные изомочевины.

В настоящее время разработано много разных подходов к анализу количественной связи структура — свойство или структура — активность (КССС/КССА) [1 – 6]. Мы остановили свой выбор на методе кластерного анализа, так как простые модели, характерные для ранних стадий развития КССС, не показали эффективности решения поставленной задачи [7, 8]. Кластерный анализ включает в себя различные математические алгоритмы и правила классификации исходной базы данных (БД), что позволяет в дальнейшем делать выводы о зависимости между изучаемым свойством и описателями молекулярной структуры (дескрипторами) или их сочетаниями [9, 10]. При этом набор дескрипторов, значимых для выделения групп соединений, и правила кластеризации специфичны для различных БД.

В настоящей работе представлена кластеризация ряда производных изотиомочевины, а именно S,N,N,N′-тетразамещенные изотиомочевины, которые обладают нейропротекторными свойствами [11]. Нейропротекторную активность (Са %) синтезированных соединений оценивали по их способности ингибировать глутамат-индуцированный захват $^{45}Ca^{2+}$ в синаптосомы коры мозга крыс [12]. Исходная БД состояла из 69 соединений, не содержащих атомы кислорода и галогенов. В качестве дескрипторов молекулярной структуры взяты топологические индексы (87 дескрипторов), которые количественно описывают молекулярную топологию на основе представления структуры в виде графа [13].

В качестве оценки успешности кластеризации мы взяли значение коэффициента разнообразия (div), который рассчитывали с помощью программы CheD [10] как для БД в целом, так и для получаемых в процессе исследования кластеров. Чем ближе к 1 этот коэффициент, тем неоднородней (разнообразней) БД. И наоборот, чем коэффициент разнообразия ближе к 0, тем более близки молекулярные структуры в данной выборке. Для исходной БД он был равен 0,455.

Обработка БД велась с помощью программного пакета STATISTICA. Дескрипторы с нулевой вариацией для всех соединений были удалены. Если 2 или более дескриптора имели взаимную корреляцию 0,99 и выше, то оставлялся только один из таких дескрипторов. После проведения этих процедур число дескрипторов сократилось с 87 до 57, но все равно было велико для того, чтобы сделать какие-либо выводы. Для того чтобы уменьшить число переменных в БД, применялся факторный анализ, который является статистическим методом обработки данных и позволяет сократить число переменных, основываясь на методе выделения главных компонент (факторов) [8]. Каждый фактор является линейной комбинацией исходных дескрипторов. Для определения числа факторов использовался критерий каменистой осыпи. Исходя из этого критерия, выделено 4 фактора (главные компоненты), которые описывали 83 % информации для данного пространства. В результате каждая структура рассматривалась в дальнейшем как точка $X_j = (x_{1j}, x_{2j}, x_{3j}, x_{4j})$ в 4-мерном факторном пространстве. В качестве метрики пространства взято евклидово расстояние, вычисляемое по формуле:

$$D_{jk} = \sqrt{\sum_{i=1}^{n} (x_{ji} - x_{ki})^2}. \qquad (1)$$

Здесь x_{ij} — i-тая координата соединения X_j. Близость точек X_j и X_k в пространстве факторов измеряется величиной D_{jk}: чем меньше эта величина, тем ближе точки в пространстве, при $D_{jk} = 0$ точки совпадают.

На рис. 1 представлены проекции точек нашей БД на плоскость факторов 1 – 2. Отчетливо видны 4 кла-

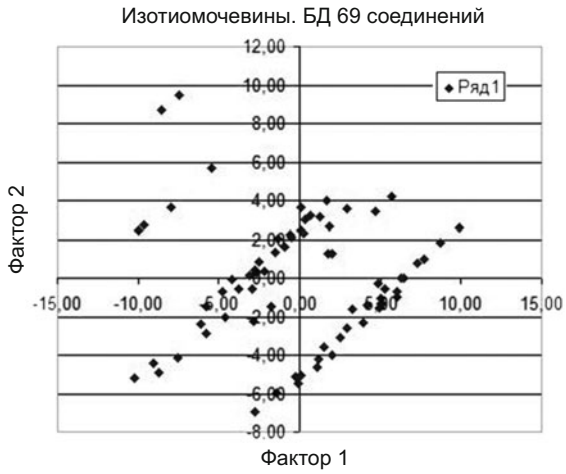


Рис. 1. Проекции точек в подпространстве факторов 1 – 2 топологических дескрипторов.

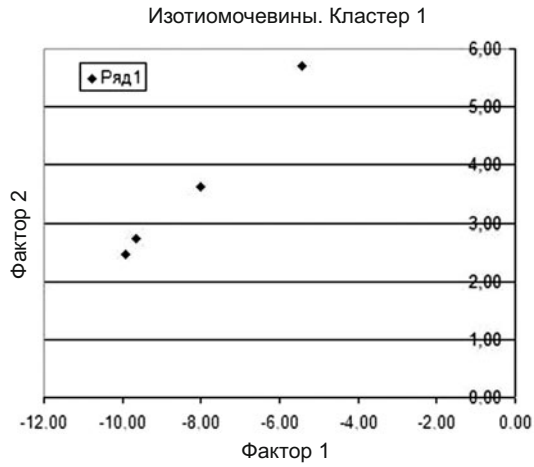


Рис. 2. Кластер 1. Проекции точек на плоскость факторов 1 – 2.

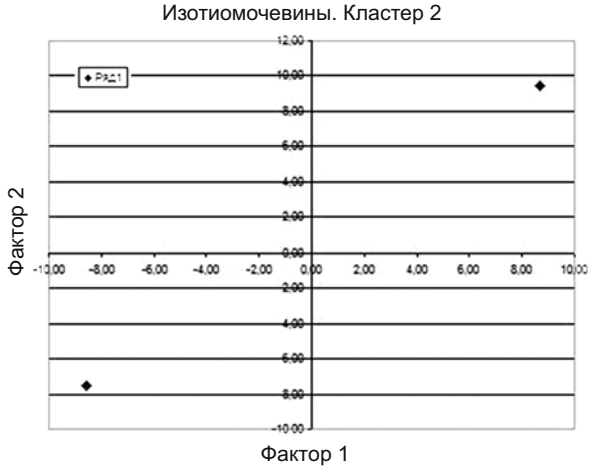


Рис. 3. Кластер 2. Проекции точек на плоскость факторов 1 – 2.

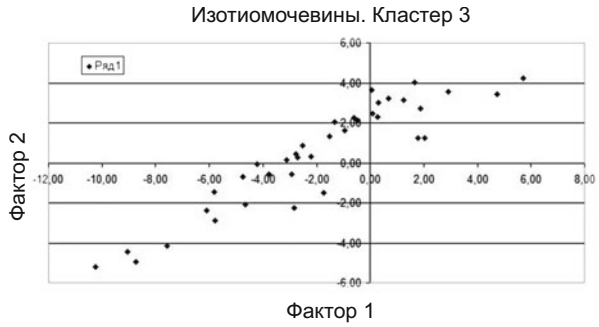


Рис. 4. Кластер 3. Проекции точек на плоскость факторов 1 – 2.

стера, достаточно изолированные друг от друга. Поскольку кластеры вытянуты в пространстве, и расстояние между точками, входящими в один кластер, может быть больше расстояния между центрами соседних кластеров, стандартный метод выделения кластеров в данном случае не может иметь успеха. Поэтому нами предложен следующий алгоритм. На первом этапе методом *к*-средних [14, 15] выделялись компактные множества точек, которые затем объединялись в более крупные кластеры на основе сравнения внутри- и межкластерных расстояний.

Таблица 1

Характеристики кластеров в факторном пространстве топологических дескрипторов

Номер кластера	Коэффициент разнообразия	Число соединений
1	0,344	4
2	0,159	2
3	0,452	36
4	0,300	27

Метод *к*-средних выделяет заданное число кластеров так, что внутрикластерные расстояния (расстояние между точками одного кластера) существенно меньше межкластерных расстояний (расстояния между ближайшими точками соседних кластеров). Для определения числа кластеров использовали следующий подход. Первоначальное число кластеров задано равным 69 (по числу соединений в базе данных). В этом случае кластеры содержали только 1 точку. Затем мы уменьшали число кластеров, следя за количеством точек в кластерах и числом итераций для получения разбиения. Более естественным признавалось разбиение на кластеры за меньшее число итераций. Для 51 кластера решение получено за 1 итерацию. При дальней-

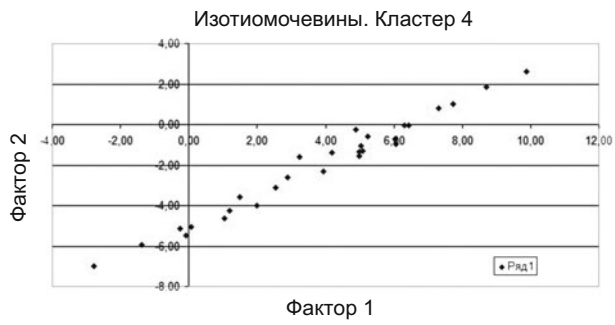
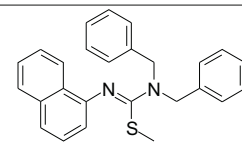
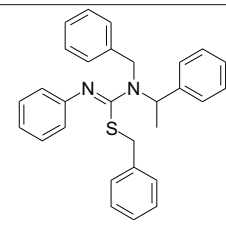
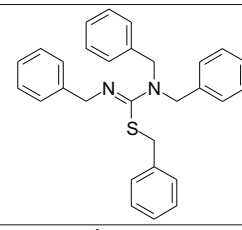
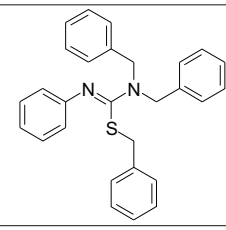
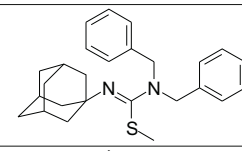
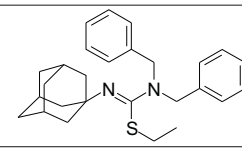
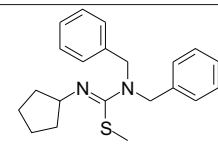
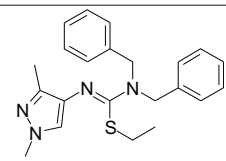
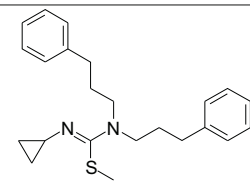
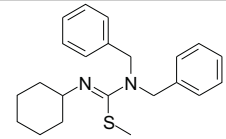
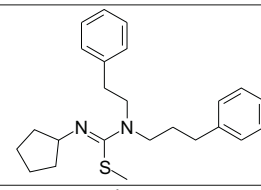
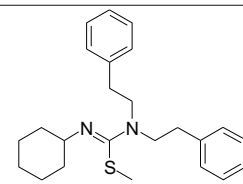
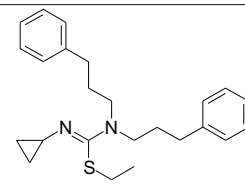
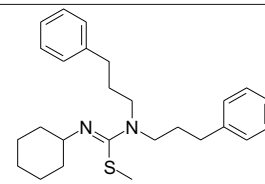
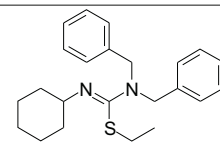
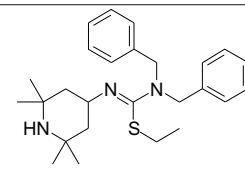
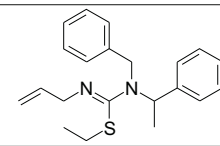
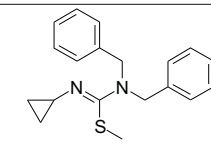
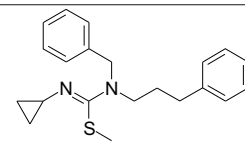
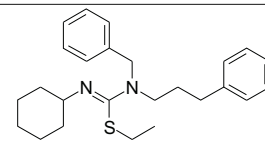
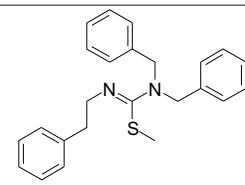


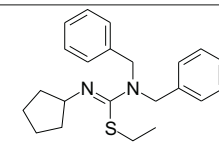
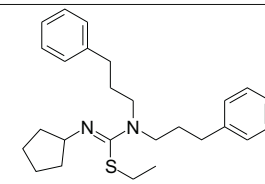
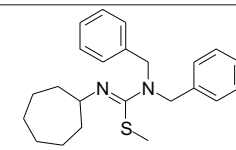
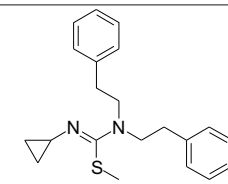
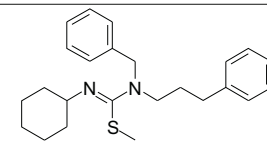
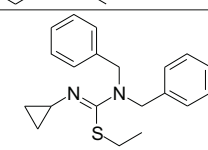
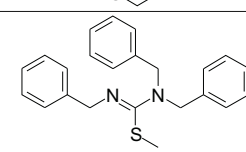
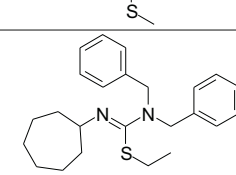
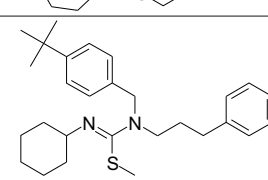
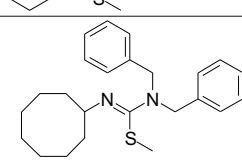
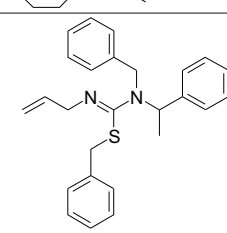
Рис. 5. Кластер 4. Проекции точек на плоскость факторов 1 – 2.

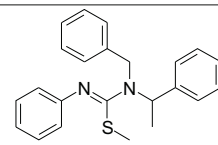
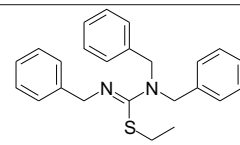
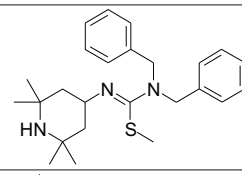
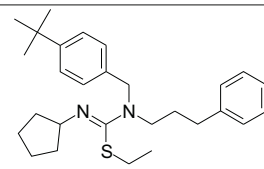
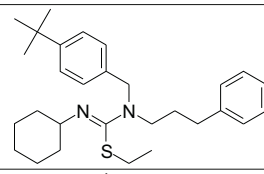
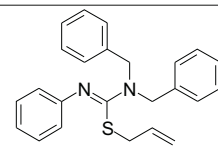
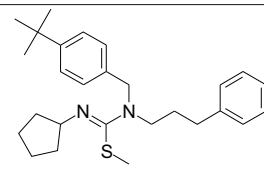
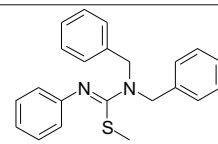
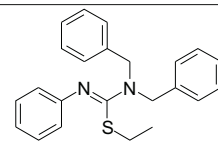
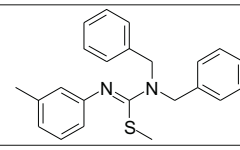
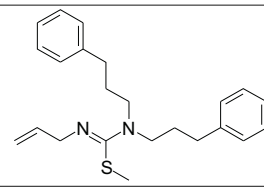
Таблица 2

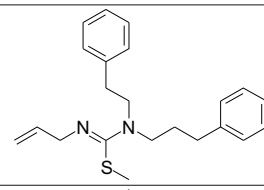
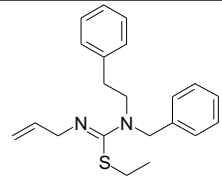
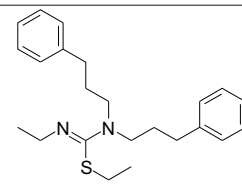
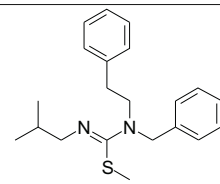
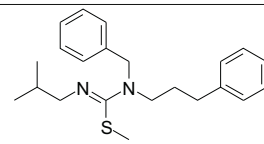
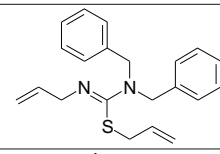
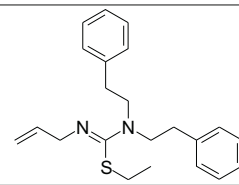
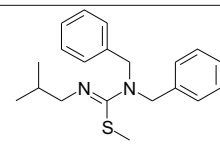
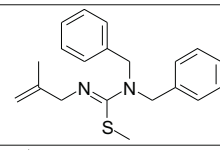
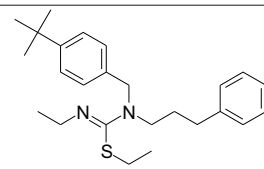
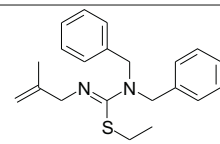
Результат кластеризации БД производных изотиомочевин

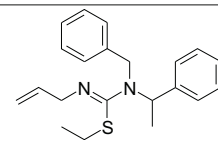
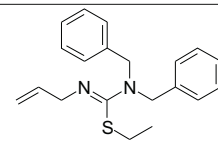
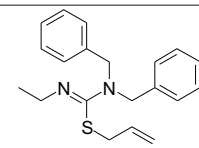
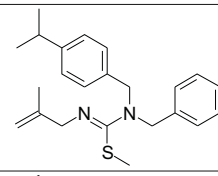
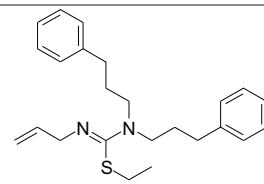
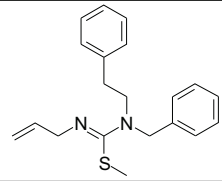
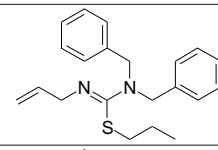
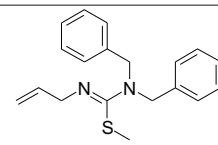
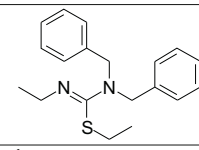
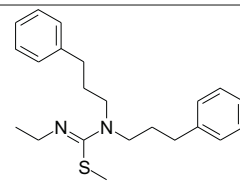
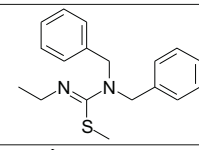
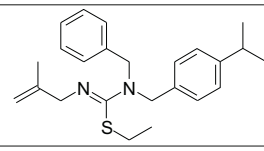
Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
Кластер 1 (4 соединения) div = 0,344				
1	$C_{26}H_{24}N_2S$	396,56		87,1 ± 5,2
2	$C_{29}H_{28}N_2S$	436,62		96,5 ± 0,4
3	$C_{29}H_{28}N_2S$	436,62		108,3 ± 9,3
4	$C_{28}H_{26}N_2S$	422,59		132,2 ± 6,8
Кластер 2 (2 соединения) div = 0,159				
5	$C_{26}H_{32}N_2S$	404,62		9,2 ± 1,4
6	$C_{27}H_{34}N_2S$	418,65		42,5 ± 7,4
Кластер 3 (36 соединений) div = 0,452				
7	$C_{21}H_{26}N_2S$	338,52		0,1 ± 0,1
8	$C_{22}H_{26}N_4S$	378,54		0,1 ± 0,1
9	$C_{23}H_{30}N_2S$	366,57		0,1 ± 0,1
10	$C_{22}H_{28}N_2S$	352,54		0,2 ± 0,2

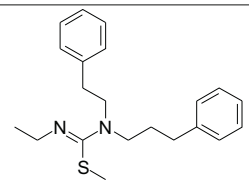
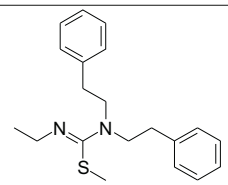
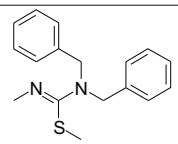
Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
11	$C_{24}H_{32}N_2S$	380,60		1,5 ± 1,5
12	$C_{24}H_{32}N_2S$	380,60		2,1 ± 2,1
13	$C_{24}H_{32}N_2S$	380,60		2,9 ± 2,9
14	$C_{26}H_{36}N_2S$	408,65		3,3 ± 0,4
15	$C_{23}H_{30}N_2S$	366,57		3,7 ± 3,7
16	$C_{26}H_{37}N_3S$	423,67		4,7 ± 2,5
17	$C_{21}H_{26}N_2S$	338,52		5,95 ± 1,55
18	$C_{19}H_{22}N_2S$	310,46		6,2 ± 1,9
19	$C_{21}H_{26}N_2S$	338,52		6,3 ± 6,3
20	$C_{25}H_{34}N_2S$	394,62		6,8 ± 6,8
21	$C_{24}H_{26}N_2S$	374,55		7,1 ± 0,8

Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
22	$C_{22}H_{28}N_2S$	352,54		7,5 ± 1,0
23	$C_{26}H_{36}N_2S$	408,65		7,8 ± 1,8
24	$C_{23}H_{30}N_2S$	366,57		8,8 ± 1,3
25	$C_{21}H_{26}N_2S$	338,52		8,8 ± 5,4
26	$C_{24}H_{32}N_2S$	380,60		10,1 ± 4,7
27	$C_{20}H_{24}N_2S$	324,49		11,8 ± 3,4
28	$C_{23}H_{24}N_2S$	360,52		12,3 ± 5,6
29	$C_{24}H_{32}N_2S$	380,60		12,9 ± 6,4
30	$C_{28}H_{40}N_2S$	436,71		27,1 ± 10,0
31	$C_{24}H_{32}N_2S$	380,60		28,7 ± 3,9
32	$C_{26}H_{28}N_2S$	400,59		29,9 ± 5,9

Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
33	$C_{23}H_{24}N_2S$	360,52		30,1 ± 8,5
34	$C_{24}H_{26}N_2S$	374,55		31,0 ± 1,6
35	$C_{25}H_{35}N_3S$	409,64		51,2 ± 7,1
36	$C_{28}H_{40}N_2S$	436,71		55,4 ± 6,2
37	$C_{29}H_{42}N_2S$	450,73		56,7 ± 3,1
38	$C_{24}H_{24}N_2S$	372,53		59,3 ± 12,1
39	$C_{27}H_{38}N_2S$	422,68		62,2 ± 5,7
40	$C_{22}H_{22}N_2S$	346,50		70,2 ± 3,5
41	$C_{23}H_{24}N_2S$	360,52		73,6 ± 2,8
42	$C_{23}H_{24}N_2S$	360,52		74,1 ± 0,8
Кластер 4 (27 соединений), div = 0,300				
43	$C_{23}H_{30}N_2S$	366,57		0,1 ± 0,1

Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
44	$C_{22}H_{28}N_2S$	352,55		0,1 ± 0,1
45	$C_{21}H_{26}N_2S$	338,52		0,1 ± 0,1
46	$C_{23}H_{32}N_2S$	368,59		0,1 ± 0,1
47	$C_{21}H_{28}N_2S$	340,53		0,9 ± 0,9
48	$C_{22}H_{30}N_2S$	354,56		1,7 ± 1,7
49	$C_{21}H_{24}N_2S$	336,50		3,2 ± 1,3
50	$C_{22}H_{28}N_2S$	352,55		3,2 ± 3,2
51	$C_{20}H_{26}N_2S$	326,51		3,4 ± 1,2
52	$C_{20}H_{24}N_2S$	324,49		3,6 ± 3,6
53	$C_{25}H_{36}N_2S$	396,64		3,9 ± 3,9
54	$C_{21}H_{26}N_2S$	338,52		5,6 ± 5,4

Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
55	$C_{21}H_{26}N_2S$	338,52		5,95 ± 1,55
56	$C_{20}H_{24}N_2S$	324,49		6,7 ± 2,9
57	$C_{20}H_{24}N_2S$	324,49		10,8 ± 0,6
58	$C_{23}H_{30}N_2S$	366,57		11,2 ± 6,3
59	$C_{24}H_{32}N_2S$	380,60		11,6 ± 2,7
60	$C_{20}H_{24}N_2S$	324,49		11,6 ± 5,0
61	$C_{21}H_{26}N_2S$	338,52		13 ± 4,3
62	$C_{19}H_{22}N_2S$	310,46		14,9 ± 0,6
63	$C_{19}H_{24}N_2S$	312,48		23,6 ± 1,3
64	$C_{22}H_{30}N_2S$	354,56		25,7 ± 3,6
65	$C_{18}H_{22}N_2S$	298,45		49,6 ± 7,1
66	$C_{24}H_{32}N_2S$	380,60		49,7 ± 1,1

Соединение	Брутто-формула	Молекулярная масса	Структура	Ca %
67	$C_{21}H_{28}N_2S$	340,53		54,1 ± 8,1
68	$C_{20}H_{26}N_2S$	326,51		74,1 ± 0,5
69	$C_{17}H_{20}N_2S$	284,43		84,7 ± 3,2

шем уменьшении числа кластеров количество итераций, за которое было получено решение, возрастало. Кластеры содержали от 1 до 6 точек. Межкластерные расстояния принимали значения от 0,65 до 11,73. Максимальное внутрикластерное расстояние равнялось 0,41. Таким образом, выделены компактные множества точек. На следующем этапе кластеры, находящиеся на расстоянии меньше 2, объединялись в один. Получившиеся в результате цепочки точек (рис. 2 – 5) полностью совпадают с кластерами на рис. 1.

Данные о числе соединений и коэффициенте разнообразия (div) представлены в табл. 1, структуры соединений и величины их активности (Ca %) — в табл. 2.

Величина Ca % отражает способность соединения ингибировать специфический глутамат-индуцированный захват $^{45}Ca^{2+}$ в синаптосомы коры мозга крыс. Чем выше активность соединения, тем меньше $^{45}Ca^{2+}$ захватывается в синаптосомы. Поэтому соединение считается активным, если Ca % < 50, в противном случае соединение неактивно. Таким образом, в БД содержится 52 активных и 14 неактивных соединений. Кластер 1 включает в себя 4 соединения (1 – 4) (div = 0,344), причем все они являются неактивными. Соединения 1, 2, 4, входящие в этот кластер, содержат арильный заместитель у иминного атома азота. Кластер 2 состоит из 2 точек, которые изолированы от остальных в факторном пространстве (рис. 1), и обе являются активными. Структуры 5, 6, входящие в него, имеют адамантановый фрагмент. Кластер 3, в который входят 28 активных (7 – 34) и 8 неактивных (35 – 42) соединений, по разнообразию сравним с полной БД (div = 0,452). Стоит отметить, что он распадается на 2 отдельных кластера (6 точек ниже основного множества), но в рамках изложенного метода не представляется возможным сделать это разделение. Наиболее активные соединения этого кластера (7, 9, 10) имеют циклоалкильные заместители у иминного атома азота. Неактивные же соединения (40 – 42), как и в случае кластера 1, содержат арильный фрагмент у иминного атома азота. Кластер 4 содержит 3 неактивных (67 – 69) и 24 активных (43 – 66) соединения и является более однородным (div = 0,300) по сравнению со всей БД. В этот кластер попали соединения с алкильными или аллильным (наиболее активные 43 – 45) заместителями у иминного атома азота.

Таким образом, произведена успешная кластеризация базы данных производных изотиомочевины. Предложен алгоритм двухэтапной кластеризации, основанный на кластеризации методом *к*-средних и процедуры объединения первичных кластеров в более крупные, исходя из сравнения внутри- и межкластерных расстояний. Первоначальное число кластеров для метода *к*-средних выбрано в результате анализа числа итераций, за которое получено решение. Графическое представление точек в проекциях на оси факторного пространства топологических дескрипторов полностью подтверждает успешность данного подхода.

Работа выполнена за счет средств Российского научного фонда, проект № 14-23-00160.

В работе использовано оборудование Центра коллективного пользования ИФАВ РАН (соглашение № 14.621.21.0008, идентификатор работ REMEFI 62114X0008).

ЛИТЕРАТУРА

1. D. K. Agrafiotis, D. Bandyopadhyay, J. K. Wegner, and H. van Vlijmen, *J. Chem. Inf. Model.*, **47**, 1279 – 1293 (2007).
2. R. Guha and P. C. Jurs, *J. Chem. Inf. Model*, **45**, 800 – 806 (2005).
3. D. M. Hawkins, S. C. Basak, and X. Shi, *J. Chem. Inf. Comput. Sci.*, **41**, 663 – 670 (2001).
4. I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *J. Chem. Inf. Comp. Sci.*, **37**(4), 715 – 721 (1997).
5. M. Kumar, K. Thurow, N. Stoll, R. Stoll, *Eur. J. Med. Chem.*, **42**, 675 – 685 (2007).
6. G. Downs and J. Barnard, in: *Reviews in Computational Chemistry*, vol. 18, Lipkowitz K. and Boyd D. (eds.), Wiley-VCH, New York (2002), pp. 1 – 40.

7. T. Varin, R. Bureau, Ch. Mueller and P. Willett, *J. Mol. Graph. Model.*, **28**(2), 187 – 195 (2009).
8. G. J. Niemi, in: *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, W. Karcher and J. Devillers (eds.), Brussels and Luxembourg, (1990), pp. 153 – 169.
9. A. Smellie, *J. Chem. Inf. Comput. Sci.* **44**, 1929 – 1935 (2004).
10. S. V. Trepalin and A. V. Yarkov, *J. Chem. Inf. Comput. Sci.*, **41**, 100 – 107 (2001).
11. G. L. Perlovich, A. N. Proshin, T. V. Volkova, et al., *J. Med. Chem.*, **2**(7), 1845 – 1852 (2009).
12. S. Bachurin, I. Baskin, V. Grigoriev, et al., *Drug Fut.*, **29**, Supp. A., 188 (2004).
13. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH Publishers, Weinheim (2000).
14. J. D. Holliday, S. L. Rodgers, P. Willett, et al., *J. Chem. Inf. Comput. Sci.*, **44**, 894 V. 902 (2004).
15. R. Stanforth, E. Kolossov and B. Mirkin, in: *Selected Contributions in Data Analysis and Classification*, P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (eds)., Springer, (2007), part II, pp. 225 – 233.

Поступила 13.10.14

APPLICATION OF MOLECULAR TOPOLOGICAL DESCRIPTORS FOR CLUSTERING A DATABASE OF ISOTHIOUREA DERIVATIVES IN STUDYING STRUCTURE – ACTIVITY RELATIONSHIPS

E. P. Andreeva*, A. N. Proshin, I. V. Serkov, L. N. Petrova, and S. O. Bachurin

Institute of Physiologically Active Compounds, Russian Academy of Sciences, Chernogolovka, Moscow oblast, 142432 Russia

* e-mail: admiandrew@yandex.ru

Clustering of a database on some S,N,N,N′-tetrasubstituted isothiourea derivatives possessing neuroprotective properties was realized for studying their structure – activity relationships. Combination of PCA, *k*-means clustering, and joining methods based of within- and between-cluster distances analysis have been applied for clustering in a factor space of molecular topological descriptors. The initial number of clusters in *k*-means clustering was determined from the number of iterations for which a solution was obtained. Homogeneity of the whole database and clusters was estimated by calculating the coefficients of molecular diversity. Graphical representation of compounds as points in the factor space leads to a conclusion about successful applicability of the proposed clustering approach.

Keywords: clustering; topological descriptors; isothiourea derivatives; structure – activity relationships.