UDK 544.022.244

N.V.Vityuk, A.N.Vityuk

COMBINATORIAL SIMILARITY MEASURE FOR SOLVING THE PROBLEM «QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP»

Description of the molecular structure is the important stage in solving the QSAR-problem. A list of substituents in the molecule is only a partial model of the molecule and reflects not all integral properties of the prototype. We propose to add the list of substituents by measures of similarity, which are expressed by numbers of similar (different) combinations of substructures in the molecules under study.

***Keywords**: QSAR − «Quantitative Structure − Activity Relationship», descriptors of the structural formula, measures of similarity.*

Описание молекулярной структуры является важнейшим этапом в решении задачи QSAR – нахождении количественной связи структура – свойство. Наиболее четкий физико-химический смысл имеет перечисление субструктур в молекуле. Такое описание молекулярной структуры используется в модели Фри-Вильсона. Эвристичность этой модели обусловлено тем, что она не использует никакой информации о механизме действия данного вещества на живой организм, но предполагает, что активность молекулы является суммой активностей субструктур в молекуле. Однако, перечисление молекулярных субструктур является неполной моделью молекулы и не описывает целостные интегральные свойства прототипа. В данной работе предлагается список молекулярных субструктур дополнить перечнем комбинаторных мер сходства, определяемых числом схожих (различных) субструктур у молекулы-эталона и рассматриваемой молекулы.

***Ключевые слова:** QSAR − количественная связь «структура − активность», дескрипторы структурной формулы, меры подобия.*

Опис молекулярної структури є найважливішим етапом в рішенні задачі QSAR − знаходженні кількісного зв'язку «структура – активність». Найбільш чіткий фізико-хімічний сенс має перерахування субструктур в молекулі. Такий опис молекулярної структури використовується в моделі Фрі-Вільсона. Евристичність цієї моделі обумовлена тим, що вона не використовує ніякої інформації про механізм дії речовини на живий організм, але припускає, що активність молекули є сумою активностей субструктур в молекулі. Проте, перелік молекулярних субструктур є неповною моделлю молекули і не описує цілісні інтегральні властивості прототипу. У цій роботі пропонується список молекулярних субструктур доповнити переліком комбінаторних мір схожості, визначених _____

числом схожих (різних) субструктур у молекули-еталону та даної молекули.

Ключові слова: *QSAR – кількісний зв'язок «структура – активність», дескриптори структурної формули, міры подібності.*

The task of predicting the properties of a system based on the composition and structure of the system was originally formulated by chemists as a problem of «QSAR» - «Quantitative Structure - Activity Relationship». The solution of this problem allows to carry out purposefully synthesis of substances with desired chemical and pharmaceutical properties. At present the «QSAR»-problem has purchased all typical features of theory of patterns recognition [1].

The mathematical methods used for the decision of «QSAR»-problem found application in the decision of many technical tasks [2-4]. However, the basic paradigm of solving QSAR-problem — structurally similar objects have similar properties — has not lost its applicability.

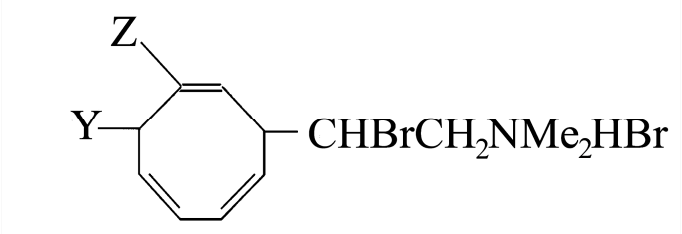
There are descriptors of different levels [5] for description of structure of molecule. The most common numerical characteristics of molecular structure have a clear physicochemical interpretation. Descriptors of the structural formula are the simplest ones among them.

Such descriptors have been proposed by Free and Wilson [6] for a relatively identical group of compounds whose structure is described by the homogeneous core and variable substituents. One of the reasons «longevity» Free-Wilson model is that this method does not use any information about the mechanism of action of chemical substances, and the predictive ability follows from the assumption that the biological activity A of the molecule is the sum of the contributions of certain activities of the substituents in the molecule, i.e.

$$A = a_o + \sum_j a_j X_j ,$$

here a_j – contribution of j-th substituent to the total activity A, X_j - binary variable, which in the presence or absence of the j-th substituent takes respectively the values 1 or 0. The regression coefficients are found by solving the system of equations for the set of molecules under study.

The sample of 22 derivatives of *N*,*N*-Dimethyl-2-bromophenethylamines



(hypotensive activity of these compounds described in [7]) has been used as a classic "ground" to test various methods of solving the QSAR-problem based on the Free-Wilson model.

Duewer [8] carried out a comprehensive statistical QSAR-analysis of these compounds, called attention to the fact that the structure of the covariance matrix in the Free-Wilson model had several features that require careful use of direct methods of regression analysis. Duewer also showed that different solutions QSAR-problem based on the additive idea of Free-Wilson differ only by overcoming the mathematical difficulty in solving multidimensional QSAR-equation – the need to consider the possible influence of the substituents on each other [9].

We can assume that the mathematical difficulties in solving QSAR-problems ensues from the fact that the list of local substructures (enumeration of substituents) is only a partial model of the molecule, and reflects not all integral properties of the prototype.

We propose to complete additionally the list of local substructures by measures of similarity, expressed as a number of similar (different) combinations of substructures. Using combinatorial measure of similarity which are abstracted from the chemical nature of the substructure (substituents) provides a more extended integral description of the molecule and develops the metho-dology of applying the principle of similarity.

The idea to supplement the list of local substructures with the list of combinatorial similarity measures was used for solving QSAR-problem on the above-mentioned 22 derivatives of *N*,*N*-Dimethyl-2-bromophenethylamines (table 1).

Number 1 in table 1 was assigned to the most active molecule – ("the leader": Y = CH_3, Z = *Br*), number 22 was assigned to the most not active molecule ("the anti-leader": Y = Z = *H*). This numbering of the molecules according to their activity is equivalent to using the rank scale activity measurement.

For determination the measure of similarity of an *i*-th molecule with an *j*-th molecule we determined the number of $p(1)$ unit, $p(0)$ zero features; number of coinciding $q(1,1)$, $q(0,0)$ and number of mismatched features $q(0.1)$, $q(1.0)$ in *i*-th and *j*-th molecules. In this way, we make full use of the original data matrix, because QSAR-problem is usually characterized by a small number of overlapping features and a large number of zero signs.

Procedure involving "zero events" from a complete dictionary of features describing the phenomenon expands the field of applicability of the principle of similarity. It allows to define in multidimensional space of different features the "closeness" of an object to the most active (leader) molecule and "remoteness " from the not-active molecule (anti-leader).

As noted above, mathematical algorithm of regression analysis imposes certain constraints on the original digital data. Therefore, to implement the heuristic Free-Wilson idea we earlier offered non-regression approaches such as method of the Barycentric coordinates [10] and the Trend vector [11].

In this article we used cluster analysis to establish a relationship "structure - activity" because cluster analysis does not impose any restrictions on the kind of the objects, and allows us to consider the set of initial data almost arbitrary nature.

We shared *N* objects into two clusters. Cluster, consisting of molecules with the arithmetic mean value of their numbers in Table 1 less than *N*/2 (that is cluster consisting of molecules with "associated" activity rank less than *N*/2) was attributed as “active”.

Analysis of the original 22 compounds in the space of binary variables X*j* in Free-Wilson model (Table) showed that the cluster №1 ("active") consists of five compounds (№ 3, 6, 8,12,15) with an average rank (that corresponds to the "associated” rank of activity of the compounds 8,8), cluster № 2 (non-active) includes the remaining 17 compounds with " associated" rank of acti-vity 12,3. The difference between the "related " ranks of activity is 12,3 – 8,8 = 3,5.

Spearman correlation coefficient ρ between the experimentally obtained ranks of activity and "associated" ranks assigned by cluster analysis was ρ = 0,231. This value is less than ρ_{crit} - the critical value of Spearman correlation coefficient (for *N* = 22 objects and α = 0,05 critical Spearman correlation coefficient is ρ_{crit} = 0,360 [12]).

In the second stage initial database was supplemented by combinatorial similarity measures considered relative molecular compounds "leader» and "anti-leader». In this the molecules themselves № 1 and № 22, have been eliminated from consideration. Therefore, a set of *N* = 20 compounds has been subjected to the cluster analysis.

Analysis of this sample in the "advanced" feature space has led to the emergence of two clusters consisting of the same number of objects (10). Cluster № 1 includes compounds № 1, 2, 5, 6, 7, 8, 9, 11, 13, 17 (“associated” rank of activity is 7,9), and other objects formed cluster № 2 (“associated” rank of activity is 13.1. The difference between the "associated" ranks of activity of formed two clusters is equal to 13,1 – 7,9 = 5,2. This value exceeds difference between clusters № 1 and № 2 in enumerative feature space (ΔR = 3,5). It demonstrates efficiency the proposed approach for the selection of active subsample of the objects. Spearman correlation coefficient ρ between the experimentally obtained ranks of activity and “associated” ranks assigned by cluster analysis was significant: ρ = 0,451> ρ_{crit} (ρ_{crit} = 0,379 for *N* = 20 objects and α = 0,05 [12]).

Thus, when we replenish the feature space by combinatorial measures of similarity of the molecules with the molecules of "leader" and "anti-leader" we achieve statistically significant excretion of subsample of "active" compounds for further solving QSAR-problem via more advanced methods.

Table

Hypotensive activity of some N,N-Dimethyl-2-bromophenethylamines [4]

№	Y-site						Z-site						Activity
	Substituent variations												
	H	*F*	*Cl*	*Br*	*I*	CH_3	*H*	*F*	*Cl*	*Br*	*I*	CH_3	
1	0	0	0	0	0	1	0	0	0	1	0	0	9,520
2	0	0	0	1	0	0	0	0	0	1	0	0	9,350
3	0	0	0	0	0	1	0	0	0	0	0	1	9,300
4	0	0	0	0	0	1	1	0	0	0	0	0	9,300
5	0	0	0	0	1	0	1	0	0	0	0	0	9,250
6	0	0	0	1	0	0	0	0	0	0	0	1	9,220
7	0	0	0	1	0	0	0	0	1	0	0	0	9,000
8	0	0	1	0	0	0	0	0	0	0	0	1	8,960
9	0	0	1	0	0	0	0	0	0	1	0	0	8,920
10	0	0	1	0	0	0	0	0	1	0	0	0	8,890
11	0	0	0	1	0	0	1	0	0	0	0	0	8,890
12	0	1	0	0	0	0	0	0	0	0	0	1	8,820
13	0	0	1	0	0	0	1	0	0	0	0	0	8,680
14	0	1	0	0	0	0	0	0	0	1	0	0	8,570
15	1	0	0	0	0	0	0	0	0	0	0	1	8,460
16	1	0	0	0	0	0	0	0	0	0	1	0	8,400
17	1	0	0	0	0	0	0	0	0	1	0	0	8,300
18	0	1	0	0	0	0	0	0	1	0	0	0	8,190
19	1	0	0	0	0	0	0	0	1	0	0	0	8,150
20	0	1	0	0	0	0	1	0	0	0	0	0	8,150
21	1	0	0	0	0	0	0	1	0	0	0	0	7,520
22	1	0	0	0	0	0	1	0	0	0	0	0	7,460

REFERENCES

1. *Витюк Н.В. Анализ связи "структура – свойство (активность)" на основе теории распознавания образов // Вісник Одеського державного морського університету. – 2001. – № 6. – С.220-231.*
2. *Вітюк М.В., Матоліков Д.П., Немчук О.О. Метод аналізу параметрів систем автоматизації приводів перевантажувачів на основі мір схожості // Теорія і практика будівництва, 2012. – № 7. – С.33-39.*
3. *Вітюк М.В., Яхнік Д.П. Выбор технологии перегрузочных работ, используя математические модели. Зб. тез IV Міжнародної науково-практичної конференції «Сучасні порти – проблеми та рішення» (Одеса (Україна) - Польща - Німеччина. 26.04-3.05.2012. Одеса: ОНМУ, 2012. – С.130-131.*

4. *Витюк Н.В., Яхник Д.П. Оптимизация технологии перегрузочных работ с помощью тренд-вектора: Тези доп. науково-практичної конференції "Інформаційні управляючі системи та технології". – Одеса: ОНМУ, 2012. – С.110-113.*
5. *Raevsky O.A. Molecular structure descriptors in the computer-aided design of biologically active compounds // Russ. Chem. Rev. – 1999. –V .68. − № 6. − C.505-524.*
6. *Free S.M., Wilson J.W. A mathematical contribution to structure-activity studies // Journal of Medicinal Chemistry. – 1964. − V. 7. − P.395-399.*
7. *Graham J.D.P., Karrar M.A. Structure-Action Relations in N,N-Dimethyl-2-halogenophenethylamines // Journal of Medicinal Chemistry. −1963. − V.6. − P.103-107.*
8. *Duewer D.L. The Free-Wilson paradigm redux: Significance of Free-Wilson coefficients, insignificance of coefficient "uncertainties" and statistical sins // Journal of Chemometrics.− 1990. − V.4. − № 4. −P. 299-321.*
9. *Antonov N.S., Gevenyan M.I., Tseirova L.T., Quantitative structure-activity relationships (QSAR) with variable contribution of substituents to the biological activity of chemical compounds // Pharmaceutical Chemistry Journal. − 1982. − V.16. − № 3. − P. 225-228.*
10. *Vityuk N.V., Kuz'min V.E. Synergism of QSAR models: The Free-Wilson method in a baricentric coordinate system // Pharmaceutical Chemistry Journal. − 1995. − V.29. − № 8. −P.543-546.*
11. *Vityuk N.V., Voskresenskaya E.B., Kuz`min V.E. Synergism of the Methods of Baricentric Coordinates and of Trend Vector in the Solution of the "Structure – Activity (Property) Problems // "Pattern Recognition and Image Analisys". − V.9. − № 3. − 1999. − P.529-538.*
12. *Поллард Дж. Справочник по вычислительным методам статистики. − М.: Финансы и статистика, 1982. − С. 342. (Pollard J.H. A handbook of Numerical and Statistical Techniques, Cambridge University Press, 1977.*

Стаття надійшла до редакції 12.03.2014

Рецензенти:

доктор технічних наук, професор, завідувач кафедри «Інформаційні технології» Одеського національного морського університету **В.В. Вичужанін**

доктор фізико-математичних наук, професор, завідувач кафедри фізичного і математичного моделювання Південноукраїнського національного педагогічного університету **А.Ю. Ків**