ХІМІЧНІ НАУКИ

УДК 541.6

С.В. Іванов, д-р хім. наук, проф.
В.В. Трачевський, канд. хім. наук, доц.
О.С. Тітова, канд. хім. наук, доц.
Н.В. Столярова, доц.

ВИВЧЕННЯ КІЛЬКІСНИХ СПІВВІДНОШЕНЬ СТРУКТУРА – ВЛАСТИВІСТЬ МЕТОДАМИ ЛІНІЙНОГО РЕГРЕСІЙНОГО АНАЛІЗУ І НЕЙРОННИХ МЕРЕЖ

Методами багатовимірного регресивного аналізу і нейронних мереж виконано моделювання залежності ентальпії протонізації від значень молекулярних дескрипторів молекул різних класів органічних сполук. Показано перевагу методу нейронних мереж для побудови кількісних співвідношень структура–властивість.

Modelation of protonisation dependence on the values of molecular discriptors of various classesorganic compounds is carried out by the methods of multydimensional regressive analysis and neuron nets. Advantage of neuron nets method for guantitive relationships structure-property description is shown.

Вступ

Штучні нейронні мережі (ШНМ) – відносно новий метод оброблення інформації і моделювання складних природних процесів [1; 2]. На відміну від відомих способів комп'ютерної обробки інформації за допомогою операцій числами і символами за жорстким алгоритмом, ШНМ підходять до обробки інформації як до процесу розпізнавання і класифікації образів за допомогою алгоритмів, що не формалізуються. Останнім часом ШНМ дедалі більше застосовують для прогнозування властивостей хімічних сполук з метою пошуку тих з них, які задовольняють заздалегідь задані параметри [3–8].

Постанова завдання

У цій роботі виконано порівняльний аналіз прогнозованої здатності методів лінійного регресивного аналізу і нейронних мереж на прикладі моделювання енергії протонування органічних сполук.

Молекулярне моделювання структури сполук провадили методом молекулярної механіки.

Квантово-хімічні розрахунки ряду теоретичних параметрів були виконані на напівемпіричному рівні методом AM1.

Експериментальна частина

Як об'єкти дослідження було вибрано 78 сполук (27 карбонових кислот, 21 спирт, 13 похідних аніліну і 17 силанолів).

За наслідками квантово-хімічних розрахунків для всіх молекул розраховували такі теоретичні молекулярні дескриптори:

Індекс основності…….0,30–[E(h)-(lw)]/100
QN Іонний індекс
основності…….0,30–[E(hw)-(l)]/100
A Ковалентний індекс
кислотності ……..max (-) заряд у молекулі
QP Іонний індекс
кислотності……...max (+) заряд атома H

Розраховані дані для анілінів наведено в табл. 1.

Таблиця 1

Молекулярні дескриптори для анілінів

Назва груп	V	P	B	QN	A	QP	ΔH
4-CH_3	1.1433	0.1226	0.1542	0.2220	0.1775	0.1012	1510
4-CH_3O	1.1219	0.1240	0.1569	0.2848	0.1769	0.0956	1509
3-CH_3	1.1395	0.1244	0.1581	0.2299	0.1758	0.1146	1508
Анілін	0.9693	0.1245	0.1580	0.2279	0.1748	0.1133	1505
4-F	1.0026	0.1293	0.1568	0.2270	0.1793	0.1171	1499
2-F	1.0084	0.1217	0.1551	0.2319	0.1791	0.1208	1495
3-CH_3S	1.3307	0.1177	0.1503	0.2226	0.1806	0.0998	1492
3-F	1.0044	0.1222	0.1551	0.2322	0.1791	0.1210	1489
2,4-F_2	1.0294	0.1223	0.1546	0.2263	0.1832	0.1318	I486
4-Cl	1.1304	0.1271	0.1557	0.2309	0.1795	0.1202	1486
3-Cl	1.1243	0.1272	0.1549	0.2289	0.1794	0.1192	1480
3-CF_3	1.2365	0.1116	0.1453	0.2339	0.1857	0.1025	1472

Для всіх класів сполук, окрім карбонових кислот, отримані статистично значущі рівняння ($p < 0,05$) з досить високими значеннями коефіцієнтів. Розраховані дані регресії для анілінів наведено в табл. 2.

Кількість нейронів у вхідному і вихідному шарах визначали відповідно до специфіки завдання: шість молекулярних дескрипторів на вході мережі і значення ентальпії протонування на виході мережі.

Таблиця 2

Параметри множинної регресії для анілінів

$R = 0,970$; $R^2 = 0,941$			F(6,6) - 16,09; p < 0,001. Статистична помилка = 5,36			
Дескриптори	Статистичний коефіцієнт	Статистична помилка	Коефіцієнт регресії	Статистична помилка коефіцієнта	*t*(71)	*p*-значення
Константа	–	–	483,89	540,04	0,89	0,40
V	-0,17	0,16	-24,48	21,77	-1,12	0,30
P	-0,31	0,22	-1090,66	774,10	-1,41	0,21
B	1,48	0,47	6466,86	2065,80	3,13	0,02
QN	-0,32	0,14	-309,42	137,57	-2,25	0,07
A	0,46	0,37	2268,66	1798,05	1,26	0,25
QP	-1,05	0,23	-1444,98	311,87	-4,63	< 0,05

Примітка: R – коефіцієнт кореляції; R^2 – коефіцієнт детермінації; F – критерій Фішера; t – критерій Стьюдента; p – статистичний рівень значущості.

З найбільшими внесками в усе рівняння регресії входять змінні, що описують розподіл заряду в молекулах, хоча у випадку спиртів статистично значущий внесок робить вандерваальсовий об'єм молекул, а в разі похідних аніліну − параметр, що характеризує основність з'єднань.

На рис. 1 показано двовимірну діаграму розсіювання, де по осі X відкладено прогнозовані значення ентальпії протонування, а по осі Y – їх досліджені значення.

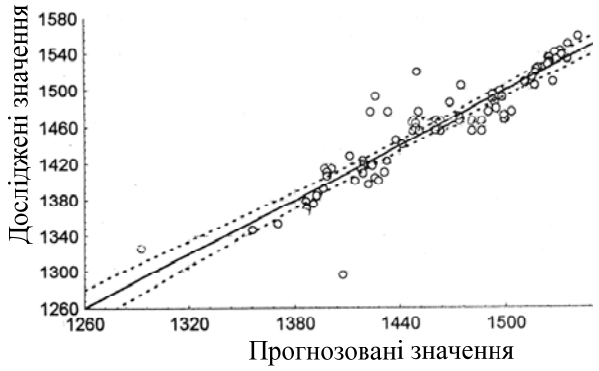

Рис. 1. Дані за ентальпіями для всіх молекул

На цій діаграмі добре видно тенденцію точок укладання на пряму лінію, хоча деякі точки виходять за межі 95% довірчого інтервалу, нанесеного штриховою лінією. У цій роботі для моделювання залежності ентальпії протонування від величин молекулярних дескрипторів використано ШНМ з архітектурою, показаною на рис. 2.

Нейрони організовано в три шари:

− вхідний;
− прихований;
− вихідний.

Її можна охарактеризувати як «вперед направлену (feed forward) з методом навчання типу «зворотного поширення помилок (back propagation)». Проте не існує точних правил, за якими можна було б визначати як кількість нейронів у прихованому шарі, так і загальну кількість прихованих шарів. Для вирішення цього питання існує два методи.

Перший метод пов'язаний з послідовним перебиранням кількості нейронів у прихованому шарі в деякому розумному інтервалі, наприклад, від одного до кількості вхідних нейронів. Потім вибирається така архітектура мережі, яка приводить до найбільш адекватного опису цільовою змінною.

Вибір здійснюється за будь-яким зі статистичних критеріїв:

− середньою абсолютною помилкою;
− середньоквадратичною помилкою;
− коефіцієнтом кореляції.

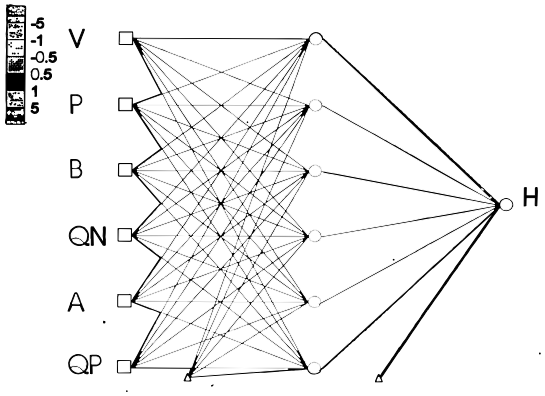

Рис. 2 Архітектура нейронної мережі

Але, незважаючи на те, що метод «зворотного поширення помилок» досить ефективний, процес навчання потребує дуже багато комп'ютерного часу навіть на сучасних швидкодійних ПК. Тому цей метод не завжди прийнятний.

Другий метод пов'язаний із заданням максимальної кількості прихованих нейронів з подальшим оцінюванням найбільш ефективної їх кількості за спеціальним алгоритмом. У цій роботі застосовували другий метод.

Зміна середньоквадратичної помилки від кількості інтерації зображено на рис. 3, *а*.

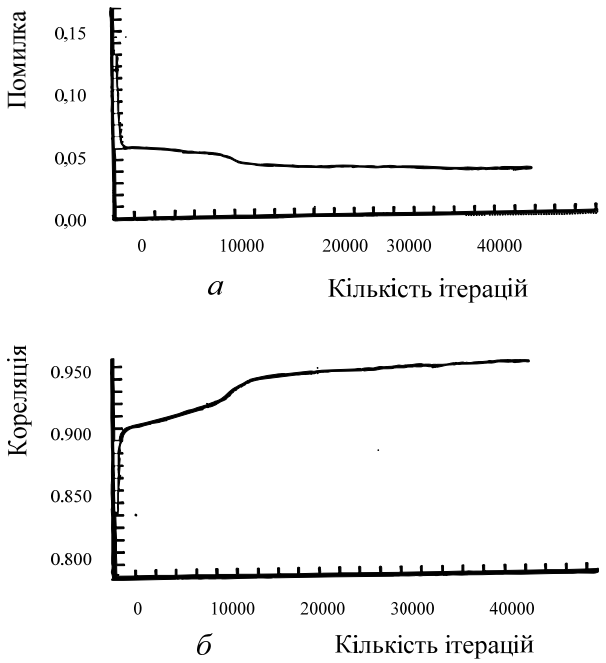

Рис. 3. Залежність середньоквадратичної помилки (*а*) і коефіцієнта кореляції (*б*) від кількості ітерацій

Під ітерацією розуміють один цикл подання нейронної мережі молекулярних дескрипторів усіх 78 молекул.

Середньоквадратична помилка являє собою квадратний корінь із суми різниць між експериментальними значеннями ентальпії протонування і значеннями на виході мережі. Алгоритм «зворотного поширення помилок» намагається мінімізувати цю помилку.

Після 10 000 ітерацій середньоквадратична помилка досягає мінімуму. Зміну коефіцієнта кореляції показано на рис. 3, *б*.

Після приблизно 10 000 ітерацій він набуває максимального значення.

На рис. 4, *а* показано, як нейронна мережа використовує приховані нейрони, у яких власне і відбувається основна обробка інформації, а також відсотковий внесок кожного нейрона в моделювання вихідного сигналу.

Як видно з рис. 4, внесок трьох нейронів (першого, п'ятого і шостого) помітно перевищує внески інших трьох. Звідси можна зробити висновок, що для побудови нейронної мережі можна було обмежитися тільки трьома прихованими нейронами. Відносні внески молекулярних дескрипторів в описі ентальпії протонування подано на рис. 4, *б*.

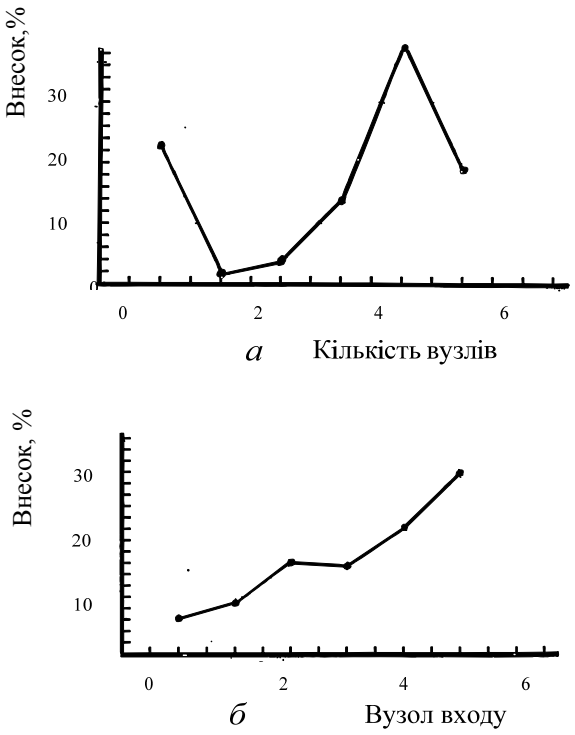

Рис. 4. Відносний вплив на ентальпію протонізації:
а – кожного молекулярного дискриптора;
б – кожного прихованого нейрона

Ці внески можна інтерпретувати так само, як і коефіцієнти при незалежних змінних у рівнянні лінійної регресії.

Найбільші значення мають заряди в молекулах і параметри, основність, що описують, і кислотність молекул.

Таким чином, дані, отримані методом нейронних мереж, узгоджуються з результатами регресійного аналізу.

На рис. 5 показано, як змінюються експериментальні і прогнозовані значення ентальпії протонування для всіх вивчених молекул. Криві майже збігаються, за винятком ділянки, в яку потрапили молекули з номерами від 60 до 70. Саме в цю ділянку входять перші члени гомологічного ряду карбонових кислот, в яких експериментальні значення ентальпії протонування майже однакові.

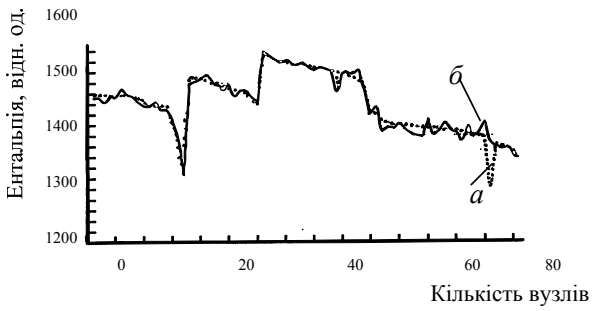

Рис. 5. Екпериментальні (*а*) і прогнозовані (*б*) значення ентальпії протонування для всього набору молекул

Діаграму розсіювання прогнозованих нейронною мережею значень ентальпії протонування залежно від експериментальних значень, зображено на рис. 6.

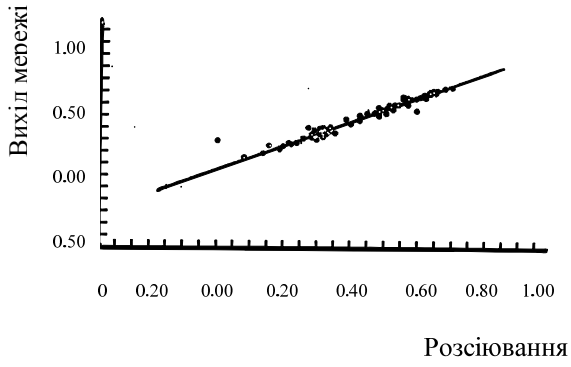

Рис. 6. Залежність розсіювання прогнозованих нейронною мережею значень ентальпії протонування від експериментальних

Чим ближче точки потрапляють на пряму лінію, яку можна описати рівнянням $Y = X$, тим краща якість моделі. Точки дуже щільно групуються навколо цієї прямої, тобто якість моделі дуже висока.

Становить інтерес порівняння статистичних показників математичних моделей, отриманих різними методами.

Для порівняння ми використовували такі статистичні характеристики:

− стандартна помилка оцінки;

− коефіцієнт кореляції.

Перша з них показує середню помилку в прогнозованому значенні, яке дає нейронна мережа, відносно експериментального значення, друга свідчить про якість моделі. Коефіцієнт кореляційної 0,919 відповідає статистичній помилці лінійної регресії 24,1, коефіцієнт кореляції 0,970 – нейронної мережі 14,1.

Висновки

Використовуючи метод ШНМ, можна отримати набагато кращі результати порівняно з регресійним аналізом. Тому його використання для вирішення різних хімічних завдань досить перспективне, оскільки дозволяє з більшою достовірністю моделювати хімічні сполуки із заздалегідь заданими параметрами.

Література

1. *Горбань А.Н., Россиев Д.А.* Нейронные сети на персональном компьютере. – НГУ: Новосибирск, 1996. – 166 с.

2. *Simon V., Gasteige I., Zupan I.* Neural Networks in Chemists // I. Amer. Chem. Soc. – 1993. − Vol. 115. – P. 9148.

3. *Gasteige I., Zupan I.*Foundations on Gomputing and Decision Sciences // Neural Networks in Chemistry, Angew. Chem. Ed. Engl. – 1993. – 105. № 4. – P. 503.

4. *Burns I.,* Whitesides G. Molecular modellinq and prediction of bioactivity // Chem. Rev. – 1993. – Vol. 93. – P. 2583.

5. *Баскин И.И., Гальберштам Н.М., Палюлин В.Н., Зефиров Н.С.* Нейрокомпьютинг и его применение // Информ. технол. – 1997. – № 9. – С. 27.

6. *Баскин И.И., Палюлин В.Н., Зефиров Н.С.* Проблема обучения распознавания образов // Нейрокомпьютер. – 1997. – № 3/4. – С. 17.

7. *Breindl A., Beck B., Clark T., Glen R.C.* Prediction of the n-octanol/water partition coefficient, logP, using a combination of semiempirical // MO-calculations and a neural network // J. Mol. Model. [Electronic Publ. − 1997. – 3(3). – P. 142-155.

8. *Дианов Р.С.* Разработка нейронной сети для прогнозирования показателей эффективности интенсификации притока газа // Пробл. развития газовой промышленности: Тез. докл. 13-й науч. конф. – Тюмень, 2004. – С. 83−85.

Стаття надійшла до редакції 05.09.07.