DỰ ĐOÁN TÍNH CHẤT MÔI TRƯỜNG CỦA MỘT NHÓM CÁC CHẤT HỮU CƠ SỬ DỤNG MÔ HÌNH ĐỊNH LƯỢNG CẤU TRÚC VÀ ĐỘ TAN

Lê Thị Đào – Phạm Văn Tất

Trường Đại học Thủ Dầu Một

TÓM TẮT

Trong công trình này, giá trị độ tan của 27 hợp chất hữu cơ được tính toán bằng việc sử dụng các tham số mô tả phân tử khác nhau. Quan hệ định lượng cấu trúc độ tan QSSRs được xây dựng bằng cách kết hợp kĩ thuật hồi qui bội và giải thuật di truyền. Các tham số phân tử quan trọng logP, SsCH3_acnt, ABSQ, nelem, nrings, SHBa, Gmax, Gmin, Xvp6 và Xvpc4 được chọn để xây dựng mô hình QSSRs tuyến tính bằng giải thuật di truyền. Mô hình QSSR tuyến tính 4 biến tốt nhất nhận được từ các tham số mô tả. Chất lượng của mô hình QSSR tuyến tính này thể hiện ở giá trị thống kê $R^2_{luyện}$ = 96,600; sai số chuẩn ước tính SE = 0,2961; F-stat = 156,0; giá trị P = 0,0; R^2_{test} = 95,020 và giá trị RSS đánh giá chéo là 2,823. Mô hình mạng nơron I(4)-HL(4)-O(1) với $R^2_{luyện}$ = 99,030 được xây dựng bằng các tham số trong mô hình QSSR tuyến tính 4 biến số. Các giá trị độ tan dự đoán của các hợp chất hữu cơ nhận được từ các mô hình phù hợp tốt với các giá trị từ tài liệu.

Từ khóa: *quan hệ định lượng cấu trúc độ tan (QSSRs), hồi qui bội, mạng nơron*

*

1. GIỚI THIỆU

Độ tan của các hợp chất hữu cơ trong nước là một trong các tính chất môi trường quan trọng nhất để giám sát và đánh giá môi trường. Tính chất này là căn cứ để xử lí các chất ô nhiễm môi trường trong các nguồn nước thải của nhà máy hóa chất. Độ tan thể hiện khả năng phân tán của một chất ô nhiễm đi vào nước. Vì vậy, tham số này là một trong những chỉ số có giá trị để đánh giá mức độ phân bố và độc tính của hóa chất. Các tham số COD và BOD cũng liên quan một phần đến độ tan của hóa chất hữu cơ. Cả hai tham số này đã được sử dụng để đánh giá chất lượng nước. Điều này cũng quyết định việc sử dụng hóa chất trong công nghiệp và các quá trình tách các chất trong tự nhiên.

Quan hệ định lượng cấu trúc và tính chất (QSPR) được thành lập bằng kĩ thuật hồi qui bội và các đánh giá thống kê khác nhau [2, 3]. Mạng thần kinh nhân tạo ngày nay đang sử dụng trong nghiên cứu quan hệ định lượng cấu trúc hoạt tính QSAR đã đưa ra trong tài liệu [4, 5]. Kĩ thuật trí tuệ nhân tạo kết hợp mạng thần kinh, logic mờ và giải thuật di truyền thể hiện tính chất mềm dẻo khi tìm kiếm các mối quan hệ phức tạp và tinh vi trong quá trình khai thác dữ liệu [5].

Trong công trình này, chúng tôi đưa ra kĩ thuật sử dụng hồi qui tuyến tính bội và mạng thần kinh để xây dựng mối quan

hệ định lượng cấu trúc và độ tan QSSR khác nhau. Các tham số mô tả cấu trúc phân tử 2D và 3D của các hợp chất hữu cơ được tính toán khi sử dụng kết hợp cơ học phân tử MM+ và hóa học lượng tử bán kinh nghiệm SCF PM3. Các mô hình QSSR tuyến tính và QSSR nơron được xây dựng từ các tham số cấu trúc với sự hỗ trợ của giải thuật di truyền. Giá trị độ tan của các hợp chất hữu cơ dự đoán bằng mô hình QSSR tuyến tính và QSSR nơron được so sánh với dữ liệu thực nghiệm.

2. PHƯƠNG PHÁP TÍNH

2.1. Dữ liệu và phần mềm

Giá trị độ tan thực nghiệm của các hợp chất hữu cơ nhận được từ một nguồn [1], đưa ra trong Bảng 1. Các tính chất mô tả phân tử 2D, 3D và các mô hình QSSR tuyến tính xây dựng bằng Regress và QSARIS [7, 11]. Các mô hình QSSR nơron được xây dựng bằng INForm [9].

Bảng 1. *Độ tan thực nghiệm của các hợp chất hữu cơ ở 25°C [1]*

STT	Hợp chất	logS	STT	Hợp chất	logS
1	Isooctan	-3,699	15	o-dicloro benzen	-1,796
2	Pentan	-1,398	16	n- butyl acetat	-0,168
3	Cyclohexan	-2,222	17	Etyl ete	0,838
4	Cyclopentan	-2,000	18	Metyl isoamyl xeton	0,231
5	Heptan	-3,523	19	Metyl t-butyl ete	0,681
6	Hexan	-1,854	20	Metyl isobutyl xeton	-0,268
7	1,1,2-tricloro trifluoro etan	-1,770	21	Etyl acetat	0,940
8	1,2,4-tricloro benzen	-2,600	22	Metyl n-propyl xeton	0,775
9	Toluen	-1,284	23	Trietyl amin	0,740
10	Cloro benzen	-1,300	24	Propylen cacbonat	1,243
11	Cloroform	-0,089	25	Metyl etyl xeton	1,380
12	n-butyl clorua	-0,959	26	isobutyl ancol	0,930
13	Etylen diclorua	-0,092	27	n-butyl ancol	0,893
14	Dicloro metan	0,204			

Quá trình thực hiện xây dựng và đánh giá mô hình qua các giai đoạn:

- Tất cả các trường hợp, trừ trường hợp thứ nhất được sử dụng để khớp hoặc luyện mô hình. Giá trị quan sát thứ nhất được dự đoán bằng mô hình QSSR tuyến tính hoặc mô hình QSSR nơron phù hợp, giá trị lệch Y_1- $\hat{Y}_1$ được xác định.

- Tất cả các trường hợp, trừ trường hợp thứ hai được sử dụng để khớp hoặc luyện mô hình. Giá trị quan sát thứ hai được dự đoán bằng mô hình QSSR tuyến tính hoặc mô hình QSSR nơron phù hợp, giá trị lệch Y_2- $\hat{Y}_2$ được xác định.

- Quá trình thực hiện tiếp tục như thế, mỗi giá trị quan sát được dự đoán bằng mô hình từ các trường hợp còn lại.

- Các giá trị R^2_{test} trung bình toàn cục nhận được từ các mô hình ở trên.

Thực hiện đánh giá chéo, tập dữ liệu được chia thành 2 tập dữ liệu nhỏ gồm: nhóm dữ liệu luyện và nhóm dữ liệu kiểm tra. Mỗi mô hình QSSR được thành lập từ nhóm luyện sử dụng để dự đoán độ tan các hợp chất hữu cơ trong nhóm kiểm tra. Sự phù hợp tốt nhất của mô hình QSSR tuyến tính và QSSR nơron được thể hiện ở giá trị $R^2_{luyện}$ và R^2_{adj} hiệu chỉnh tương

ứng; khả năng dự đoán của các mô hình được đánh giá chéo và thể hiện ở giá trị R^2_{test} kiểm tra:

- Y: giá trị quan sát; $\hat{Y}$: giá trị dự đoán; $\overline{Y}$: giá trị trung bình;

- Nhóm luyện: $R^2_{luyện}$ (mô hình tuyến tính và mô hình nơron);

- Nhóm kiểm tra: R^2_{test} (mô hình tuyến tính và nơron);

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Tính toán các tham số phân tử

Các hợp chất hữu cơ được xây dựng, tối ưu hóa và tính toán các tham số mô tả tính chất cấu trúc đặc trưng phân tử bằng cơ học phân tử trong HyperChem [1]. Các tham số cấu trúc 2D và 3D, tham số hình học, tham số thế tĩnh điện phân tử, tham số phụ thuộc điện tích và hệ số phân tán octanol/nước nhận được từ hệ thống QSARIS [7, 11].

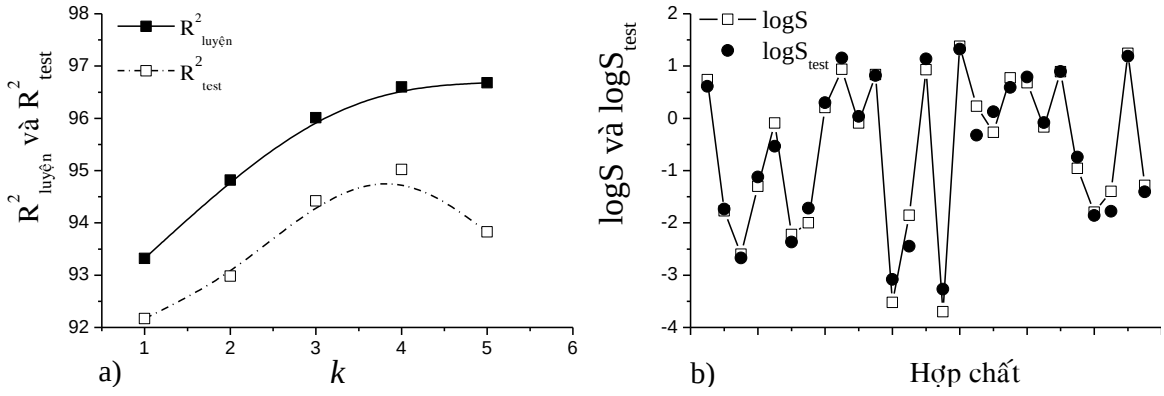
3.2. Xây dựng quan hệ QSSR tuyến tính

Mô hình QSSR tuyến tính được thành lập bằng hệ thống Regress [6, 8] và QSARIS [7], các tham số cấu trúc phân tử được lựa chọn đưa vào mô hình bằng giải thuật di truyền theo kĩ thuật tiến hóa vi phân. Tất cả các quá trình chọn lựa tham số cấu trúc phân tử dựa vào các giá trị thống kê mô hình: $R^2_{luyện}$, sai số chuẩn SE, R^2_{adj}, R^2_{test} và giá trị F-stat. Các mô hình QSSR tuyến tính tốt nhất nhận được dẫn ra Bảng 2.

Bảng 2. *Các mô hình QSSR tuyến tính (số tham số k = 1 đến 5) và các giá trị thống kê*

Tham số thống kê và tham số mô tả cấu trúc phân tử	Mô hình QSSR tuyến tính				
	A (k = 1)	B (k = 2)	C (k = 3)	D (k = 4)	E (k = 5)
$R^2_{luyện}$	93,320	94,820	96,010	96,600	96,680
R^2_{adj}	93,050	94,390	95,490	95,980	95,890
Sai số, SE	0,3890	0,3495	0,3136	0,2961	0,2994
F-stat	349,2283	219,8180	184,3612	156,0465	122,1842
R^2_{test}	92,170	92,980	94,420	95,020	93,830
Hằng số	0,9217	1,5831	2,1581	1,8666	0,3449
logP	-1,1566	-1,1350	-1,1926	-1,2251	-0,9714
SsCH3_acnt	-	0,1503	0,1931	-	0,1933
ABSQ	-	-	-0,5721	-	-
nelem	-	-	-	-	0,4477
nrings	-	-	-	-0,5465	-
Gmax	-	-	-	-	-0,0469
Gmin	-	-	-	0,3202	
Xvp6	-	-	-	-	-2,9653
Xvpc4	-	-	-	0,5461	-

Trong Bảng 2, các mô hình QSSR tuyến tính phù hợp nhất được chọn với số lượng tham số cấu trúc trong các mô hình dao động từ k = 1 đến k = 5. Sự thay đổi số lượng tham số cấu trúc dẫn đến thay đổi giá trị $R^2_{luyện}$ và R^2_{test} tương ứng như mô tả ở Hình 1.



Hình 1. *a) Biểu diễn sự thay đổi độ lớn giá trị $R^2_{luyện}$ và R^2_{test} theo số biến k trong mô hình. b) So sánh giá trị độ tan thực nghiệm và độ tan dự đoán đối với mỗi hợp chất.*

Trong các mô hình nhận được, mô hình QSSR với k = 4 cho giá trị R^2_{test} đạt giá trị cao nhất, sau đó giảm khi k tăng. Như vậy, mô hình QSSR với k = 4 là phù hợp nhất so với các mô hình còn lại. Chất lượng của mô hình QSSR này được thể hiện ở giá trị R^2 = 96,600; sai số chuẩn ước tính, SE = 0,2961; F-stat = 156,0 và R^2_{test} = 95,020; mô hình QSSR (với k = 4) được kiểm tra bằng kĩ thuật đánh giá chéo loại dần từng trường hợp với giá trị thống kê tổng bình phương hồi qui RSS = 2,823. Mô hình hồi qui QSSR tuyến tính này có dạng:

logS = -1,225LogP + 0,5461xvpc4 + 0,3202Gmin – 0,5465nrings + 1,86663 (1)

Như vậy, tập dữ liệu luyện đáp ứng tốt và mô tả bằng mô hình QSSR (1) rất có ý nghĩa về mặt thống kê. Kĩ thuật đánh giá chéo cho thấy mô hình QSSR(1) có thể được sử dụng để dự đoán logS. Các giá trị thống kê kiểm tra tính có nghĩa của các hệ số trong mô hình QSSR(1) (với k = 4), được dẫn ra ở Bảng 3. Kiểm tra tính có nghĩa của tham số đã chọn trong mô hình, tiến hành lấy 100 lần ngẫu nhiên của các giá trị logS trong số các chất đưa ra. Giá trị R^2 - R^2_n với n = 1, 2..., 100 được tính cho mỗi mô hình QSSR trong các mô hình tương ứng. Giá trị trung bình của R^2_n = 0,1504; giá trị trung bình bình phương độ lệch là 0,09849. Khoảng các giá trị R^2_n từ 0,004609 đến 0,4679.

Bảng 3. *Giá trị thống kê, hệ số của mô hình QSSR (1) với k = 4 và kiểm định giả thuyết*

Tham số	Hệ số	Giá trị P	Sai số chuẩn	Thống kê t-stat	Kiểm định giả thuyết
Hằng số	1,8666	0,0000	0,1171	15,9421	Giá trị $P < \alpha = 0,05$
logP	-1,2251	0,0000	0,0575	-21,2943	Giá trị $P < \alpha = 0,05$
Xvpc4	0,5461	0,0419	0,2528	2,1603	Giá trị $P < \alpha = 0,05$
Gmin	0,3202	0,0019	0,0908	3,5260	Giá trị $P < \alpha = 0,05$
nrings	-0,5465	0,0010	0,1448	-3,7736	Giá trị $P < \alpha = 0,05$

Các giá trị phần trăm đóng góp, $P_m x_k$,% của các tham số độc lập trong mô hình QSSR (1) với k = 4 xác định qua sự đóng góp của các tham số bằng giá trị C_{total} được mô tả ở Bảng 4. Phần trăm đóng góp trung bình MPx_k,% của mỗi biến độc lập được xác định bằng công thức:

$$MPx_k,\% = \frac{1}{N}\sum_{j=1}^{N}\left(100.\left|b_{m,i}x_{m,i}\right| \Big/ \sum_{i=1}^{k}\left|b_{m,k}x_{m,k}\right|\right) = \frac{1}{N}\sum_{j=1}^{N}\left(100.\left|b_{m,i}x_{m,i}\right| \Big/ \mathrm{C}_{\mathrm{total}}\right) \quad (2)$$

Ở đây N = 27 là tổng số hợp chất, m - hợp chất cần tính $P_m x_k$,%.

Sự đóng góp mức độ quan trọng của các tham số cấu trúc phân tử trong mô hình được sắp xếp theo trật tự dựa vào MPx_k,%: logP > Gmin > nrings > xvpc4; trong khi độ lớn của các hệ số tương ứng mỗi tham số trong mô hình được sắp xếp theo trật tự: logP > nrings > xvpc4 > Gmin.

Bảng 4. *Giá trị $P_m x_k$,% và MPx_k,%, của mỗi tham số trong mô hình QSSR (1) với k = 4.*

Hợp chất, m = 1- 27	C_{total}	$P_m x_k$, %			
		xvpc4	Gmin	nrings	LogP
Isooctan	6,0274	2,6157	4,8036	0,0000	92,5807
Heptan	5,8206	0,0000	7,4877	0,0000	92,5123
1,2,4-tricloro benzen	5,8770	8,7359	2,6695	9,2981	79,2965
Cyclohexan	5,1939	0,0000	9,2474	10,5210	80,2315
Cyclopentan	4,5477	0,0000	10,5614	12,0160	77,4226
Hexan	5,1794	0,0000	8,3769	0,0000	91,6231
o-dicloro benzen	4,9338	8,2924	3,9310	11,0756	76,7010
1,1,2-tricloro trifluoro etan	5,4325	16,8278	24,9094	0,0000	58,2628
Pentan	4,5067	0,0000	9,5473	0,0000	90,4527
cloro benzen	3,7351	3,1908	6,8066	14,6301	75,3725
Toluene	4,3299	2,4274	9,7745	12,6203	75,1777
n-butyl chlorua	3,1319	0,0000	8,3425	0,0000	91,6575
metyl isobutyl xeton	2,2394	7,0106	4,1042	0,0000	88,8852
n- butyl acetat	2,0134	1,5984	2,8944	0,0000	95,5072
etylene diclorua	2,1878	0,0000	8,1536	0,0000	91,8464
Chloroform	2,4001	0,0000	10,0058	0,0000	89,9942
dicloro metan	1,6916	0,0000	3,6806	0,0000	96,3194
metyl isoamyl xeton	2,8129	7,5866	3,4646	0,0000	88,9488
metyl t-butyl ete	1,7736	18,8568	0,7522	0,0000	80,3910
triethyl amine	2,5315	10,2335	15,0203	0,0000	74,7462
metyl n-propyl keton	1,5725	3,5448	5,8921	0,0000	90,5630
etyl ete	1,5860	0,0000	17,0344	0,0000	82,9656
n-butyl alcol	1,1898	0,0000	9,2511	0,0000	90,7489
isobutyl alcol	1,1235	8,8750	8,7083	0,0000	82,4168
etyl acetat	0,7790	4,1313	8,6586	0,0000	87,2101
propylen cacbonat	0,8925	11,9375	19,6830	61,2287	7,1509
mety etyl xeton	0,8625	9,1396	9,4530	0,0000	81,4074
Giá trị MPx_k,%		***4,6298***	***8,6376***	***4,8663***	***81,8664***

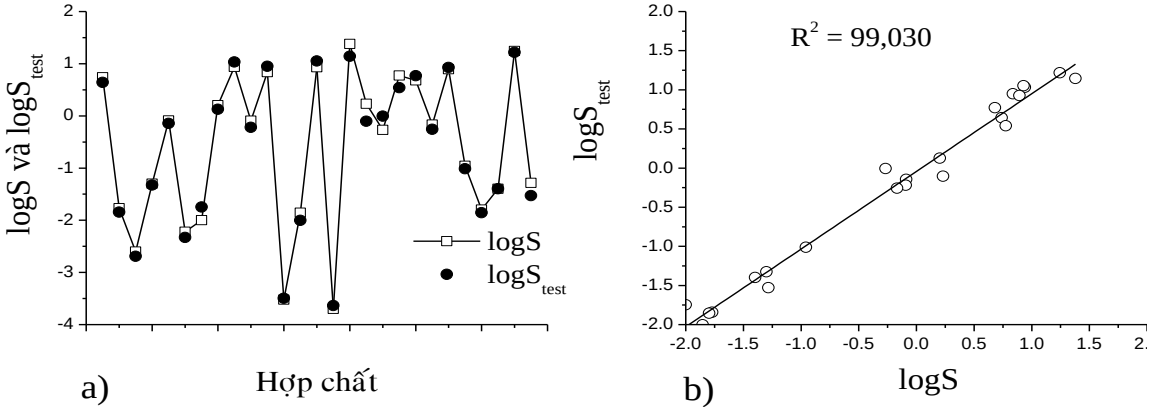
Từ kết quả Bảng 4, mức độ đóng góp của mỗi tham số trong mô hình QSSR (1) hay đúng hơn là đóng góp vào tính chất của chất; không thể dựa vào độ lớn của hệ số để đưa ra trật tự đóng góp quan trọng của tham số liên quan đến tính chất của hợp chất. Tham số logP liên quan mạnh đến độ tan của hợp chất hữu cơ. Như vậy độ tan

của chất hữu cơ gắn liền với khả năng phân tán của chất, thể hiện ở logP. Tham số Gmin thể hiện độ lớn thế tĩnh điện nguyên tử nhỏ nhất trong phân tử, tham số này có ảnh hưởng lớn đến độ tan hợp chất xếp sau tham số logP, điều này cũng thể hiện bản chất của thế tĩnh điện phân tử toàn cục. Ngoài ra tham số nrings cũng đóng góp vào độ tan, phụ thuộc ở số vòng trên phân tử mà được xác định từ R = p – (nvx – 1) với p là số cạnh liên kết vòng, nvx là số đỉnh trong phân tử không phải là các nguyên tử hydro.

3.3. Xây dựng mô hình QSSR nơron

Mô hình QSSR nơron được xây dựng trên cơ sở kĩ thuật thần kinh mờ với sự hỗ trợ của giải thuật di truyền trên hệ thống INForm [9]. Kiến trúc mạng thần kinh gồm 3 lớp I(4)-HL(4)-O(1); lớp nhập I(4) gồm 4 nơron là tham số logP, Gmin, nrings, xvpc4, lớp xuất O(1) gồm 1 nơron là tham số logS. Lớp ẩn HL(4) phía trong gồm 4 nơron. Thuật toán lan truyền ngược sai số được sử dụng để luyện mạng. Hàm truyền sigmoid đặt trên mỗi nút nơron của các lớp mạng; tham số luyện mạng gồm tốc độ học là 0,7; moment là 0,7. Sai số giám sát mục tiêu MSE = 0,000816 với 10.000 vòng lặp. Sau quá trình luyện mạng, giá trị $R^2_{luyện}$ = 99,030 trong khi mô hình QSSR (1) tuyến tính cho $R^2_{luyện}$ = 96,600.

Như vậy, mô hình QSSR nơron dựa trên kiến trúc mạng nơron I(4)-HL(4)-O(1) đạt được sự thích ứng tốt hơn so với mô hình QSSR (1) tuyến tính. Điều này có thể thấy ở Hình 1 và Hình 2, sự tương quan và tính phù hợp tốt giữa giá trị dự đoán và giá trị thực nghiệm.



***Hình 2**. a) So sánh giá trị logS và độ tan dự đoán $logS_{test}$ đối với mỗi hợp chất;*
b) Sự tương quan giữa giá trị thực nghiệm logS và giá trị dự đoán $logS_{test}$

3.4. Dự đoán độ tan của chất trong nhóm kiểm tra

Khả năng dự đoán của mô hình QSSR (1) và QSSR nơron đều được đánh giá cẩn thận bằng kĩ thuật loại bỏ dần từng trường hợp; kết quả dự đoán nhận được đối với 7 hợp chất chọn ngẫu nhiên từ Bảng 1, được dẫn ra ở Bảng 5.

Kết quả dự đoán của các mô hình QSSR được đánh giá bằng giá trị tuyệt đối của các sai số tương đối *ARE*,% tính bằng công thức:

$$ARE,\% = 100\left|(\log S - \log S_{test})/\log S\right| \quad (3)$$

***Bảng 5**. Độ tan của 7 chất chọn ngẫu nhiên được dự đoán từ QSSR (1) và QSSR nơron*

STT	Hợp chất	logS	QSSR nơron		QSSR tuyến tính	
			$logS_{test}$	*ARE,%*	$logS_{test}$	*ARE,%*
1	n-butyl clorua	-0,9586	-1,0117	5,5425	-0,7427	22,5235
2	etylen diclorua	-0,0920	-0,2191	138,1826	0,0356	138,7148
3	isobutyl alcol	0,9300	1,0523	13,1505	1,1382	22,3885
4	mety etyl xeton	1,3800	1,1438	17,1167	0,1973	85,7010
5	metyl t-butyl ete	0,6812	0,7703	13,0741	0,7886	15,7661
6	cyclohexan	-2,2220	-2,3304	4,8771	-2,3667	6,5100
7	o-dicloro benzen	-1,7960	-1,8548	3,2717	-1,8610	3,6210
	Giá trị MARE, %			***27,8879***		***42,1750***

Giá trị trung bình tuyệt đối của các sai số tương đối *MARE*, % được sử dụng để đánh giá tổng quát sai số của mô hình QSSR tính bằng công thức:

$$MARE,\% = \frac{100}{n}\left|\frac{(\log S - \log S_{test})}{\log S}\right| \quad (4)$$

Ở đây n = 7 là số hợp chất; logS là độ tan thực nghiệm, $logS_{test}$ độ tan dự đoán.

Như vậy, từ kết quả so sánh giữa hai mô hình QSSR (1) và QSSR nơron dựa vào các giá trị *MARE*,% được dẫn ra ở Bảng 5, cho thấy mô hình QSSR (1) có khả năng dự đoán kém hơn so với mô hình QSSR nơron. Kết quả dự đoán logS nhận được từ mô hình QSSR nơron gần với thực nghiệm hơn và mô hình QSSR nơron có khả năng thích ứng tốt hơn mô hình QSSR (1).

4. KẾT LUẬN

Công trình này đã xây dựng thành công mô hình QSSR tuyến tính với sự hỗ trợ của giải thuật di truyền. Kĩ thuật mới này cho phép xây dựng mô hình hồi qui đối với tập dữ liệu lớn. Giải thuật di truyền cho phép chọn lựa các tham số quan trọng đưa vào mô hình. Mô hình QSSR tuyến tính nhận được đạt yêu cầu về kiểm định thống kê. Ngoài ra kĩ thuật trí tuệ nhân tạo dựa trên quan hệ thần kinh mờ cũng được hỗ trợ bằng giải thuật di truyền để xây dựng kiến trúc mạng nơron I(4)-HL(4)-O(1) đáp ứng tốt với dữ liệu; mô hình QSSR nơron đã cho kết quả dự đoán tốt hơn nhiều so với mô hình QSSR tuyến tính. Giá trị *MARE*, % của mô hình QSSR tuyến tính lớn hơn 1,5 lần so với mô hình QSSR nơron.

Kết quả nhận được từ công trình này mở ra hướng nghiên cứu mới và có nhiều hứa hẹn ứng dụng trong lĩnh vực xử lí môi trường, thiết kế dược liệu và phẩm bào chế dược phẩm.

PREDICTION OF ENVIRONMENTAL PROPERTIES OF A SET OF ORGANIC COMPOUNDS USING QUANTITATIVE STRUCTURE SOLUBILITY RELATIONSHIPS QSSRs

Le Thi Dao – Pham Van Tat

Thu Dau Mot Universty

ABSTRACT

In this work the solubility values of 27 organic substances were calculated by using the different molecular descriptors. The quantitative structure-solubility relationships (QSSRs)

were constructed by incorporating the multiple regression technique and the genetic algorithm. The important molecular descriptors logP, SsCH3_acnt, ABSQ, nelem, nrings, SHBa, Gmax, Gmin, Xvp6 and Xvpc4 were selected for constructing the linear models QSSRs with the genetic algorithm. The best 4-variable linear model QSSR was derived from these descriptors. The quality of this linear model QSSR was pointed out in statistical values multiple R^2-training of 96.600, standard error of estimation, SE of 0.2961, F-statistic of 156.0, P-value of 0.0, multiple R^2-test of 95.020 and cross validation RSS of 2.823. The neural network model I(4)-HL(4)-O(1) with R^2-training of 99.030 was built by using descriptors in the 4-variable linear model. The predicted solubility values of organic substances resulting from these models were in good agreement with those from literature.

***Keywords*:** *quantitative structure-solubility relationships (QSSRs), multiple regression, neural network*

TÀI LIỆU THAM KHẢO

[1] Ian M. Smallwood., *Handbook of organic solvent properties,* John Wiley Inc (1996).

[2] Xiao-Lan Zeng, Hong-Jun Wang, Yan Wang, *QSPR models of n-octanol/water partition coefficients and aqueous solubility,* J.chemosphere. 10, 051, (2011).

[3] Darryl W. Hawker, Janet L. Cumming, Peta A. Neale, Michael E. Bartkow, Beate I. Escher, *A screening level fate model of organic contaminants from advanced water treatment in a potable water supply reservoir,* J. water research, 45, 768 - 780, (2011).

[4] Hongxia Zhao, Qing Xie, Feng Tan, Jingwen Chen, Xie Quan, Baocheng Qu, Xin Zhang, Xiaona Li, *Determination and prediction of octanol–air partition coefficients of hydroxylated and methoxylated polybrominated diphenyl ethers,* J. Chemosphere, 80, 660–664, (2010).

[5] Wen Zhou, Zhicai Zhai, Zunyao Wang, Liansheng Wang, *Estimation of n-octanol/water partition coefficients (Kow) of all PCB congeners by density functional theory,* J. Molecular Structure: THEOCHEM 755, 137–145, (2005).

[6] D. D. Steppan, J. Werner, P. R. Yeater, *Essential Regression and Experimental Design for Chemists and Engineers,* (2000).

[7] Phạm Văn Tất, *Phát triển mô hình quan hệ QSAR và QSPR,* NXB Khoa học tư nhiên và Công nghệ, Hà Nội, (2009).

[8] B. E. Joseph, *EXCEL for chemists,* Wiley-VCH, (2001).

[9] INForm v2.0, Intelligensys Ltd., UK (2000)

[10] HyperChem Release 8.05, Hypercube Inc., USA (2008).

[11] QSARIS 1.1, Statistical Solutions Ltd., USA (2001).