

## ● Nghiên cứu - Kỹ thuật

thu được, nhóm nghiên cứu đề xuất bổ sung chỉ tiêu định lượng SC và chỉ tiêu định tính với các chất đánh dấu (SC, XA, AF) trong các tiêu chuẩn dược liệu, sản phẩm từ ĐQDT nhằm nâng cao công tác kiểm tra chất lượng dược liệu ĐQDT ở Việt Nam.

### Tài liệu tham khảo

1. Võ Văn Chi (1997), *Từ điển Cây thuốc Việt Nam*, NXB Y học, tr. 500 - 501.
2. *Dược điển Việt Nam IV*, NXB Y học, Hà Nội (2009), tr. 760 - 762.
3. Fukuda Kozo (2009), "Fibrinolytic activity of ligustilide and pharmaceutical comparison of *Angelica acutiloba* roots before and after processing in hot water", *Journal of Traditional Medicines*, 26(5+6), pp. 210-218.
4. Fujioka T. (1999), "Antiproliferative constituents from *Umbelliferae* plants. V. A new furanocoumarin and falcariindiol furanocoumarin ethers from the root of *Angelica japonica*", *Chemical & Pharmaceutical*

*Bulletin*, 47(1), pp. 96-100.

5. Haiqing Yu (2008), "Antiinflammatory constituents in noni (*Morinda citrifolia*) fruits. In *dietary supplements*", *American Chemical Society*, Vol 987, pp 179-190.
6. Kang S.Y. (2001), "Coumarins isolated from *Angelica gigas* inhibit acetylcholinesterase: structure-activity relationships", *Journal of Natural Products*, 64(5), pp. 683-685.
7. Viện Dược liệu (2012), Báo cáo tổng kết Dự án sản xuất thử nghiệm "Hoàn thiện quy trình công nghệ sản xuất Angabin và Somamim từ cây đương quy di thực Nhật Bản".
8. Manthey John A. (2013), "The HPLC-Fluorescence detection of coumarins in 'Hamlin' Sweet orange and 'Marsh' grapefruit leaf cankers", *Proc. Fla. State Hort. Soc.*, 126, pp. 217-219.
9. Nguyễn Thị Hà Ly (2013), "Xây dựng phương pháp định lượng Z-Ligustilid trong dược liệu đương qui di thực Nhật Bản bằng kỹ thuật sắc ký lỏng hiệu năng cao", *Tạp chí Dược liệu*, tập 18, số 1, trang 55-61.

(Ngày nhận bài: 09/09/2015 - Ngày duyệt đăng: 01/10/2015)

# Xây dựng mô hình toán học mới nhằm phát hiện hợp chất ức chế hệ thống ubiquitin-proteasome chỉ từ thông tin cấu trúc phân tử

Lê Thị Thu Hương<sup>1</sup>, Vũ Đức Lợi<sup>1</sup>, Bùi Thanh Tùng<sup>2</sup>, Phạm Thế Hải<sup>3</sup>

<sup>1</sup>BM Dược liệu & Dược học Cổ truyền, Khoa Y Dược, ĐH Quốc gia Hà Nội

<sup>2</sup>BM Dược lý & Dược lâm sàng, Khoa Y Dược, ĐH Quốc gia Hà Nội

<sup>3</sup>Bộ môn Dược lực, Trường ĐH Dược Hà Nội

\*E-mail: lthuong1017@gmail.com

### Summary

As Quantitative Structure-Activity Relationships (QSAR) approaches are promising fields with several applications in drug development, especially in relation to the ubiquitin-proteasome system (UPS) that is responsible for the selective degradation of the majority of the intracellular proteins in eukaryotic cells and regulates nearly all cellular processes, dysfunction of the ubiquitination machinery, and proteasome inhibitors effective for various associated disorders, three QSAR models for identification of new lead-like UPS inhibitors were developed using TOMOCOMD molecular descriptors on a database extracted from PubChem BioAssay (AID 2486). Linear discriminant analysis (LDA), Multi-layer perceptrons (MLP) and Support vector machine (SVM) were used to adjust the models. OECD principles were used to validate the QSAR models. These models achieved good global accuracy in the training set and prediction set with value superior to 85.9 % and 81.2 %, respectively, and fulfilled OECD principles. These models

## ● Nghiên cứu - Kỹ thuật

*proved time-saving and cost-reducing in early screening phase to identify new lead-like UPS-inhibitory compounds.*

**Keywords:** Ubiquitin-proteasome system (UPS), Inhibitor, QSAR model, Linear Discriminant Analysis, Multi-layer perceptrons, Support vector machine.

### Đặt vấn đề

Quá trình nghiên cứu và phát triển thuốc thường tốn kém thời gian, công sức và tiền bạc. Trong đó, tốn kém nhất là quá trình phát hiện hợp chất dẫn đường (*lead compounds*). Hàng loạt phương pháp mới ra đời giúp sàng lọc nhanh và hiệu quả các hợp chất có hoạt tính sinh học. Theo hướng này, các phương pháp lý thuyết đã được phát triển và ứng dụng rộng rãi, từ đó, ra đời khái niệm *in silico* bổ sung cho các khái niệm *in vivo*, *in vitro*, *in situ* đã tồn tại từ lâu. Phương pháp *in silico* phát triển các nghiên cứu dựa trên các lý thuyết về hóa học, toán học, công nghệ thông tin, xác suất thống kê... để phát triển các mô hình dự đoán lý thuyết. Việc phát triển các mô hình về mối quan hệ định lượng giữa cấu trúc và hoạt tính của các hợp chất (*Quantitative Structure-Activity Relationships* - QSAR), là một bước tiến quan trọng trong các nghiên cứu này. Các mô hình QSAR giúp phát hiện các hợp chất mới có hoạt tính sinh học một cách nhanh chóng. Hơn nữa chúng lại rất hiệu quả và tiết kiệm do chỉ cần thông tin về cấu trúc của hợp chất. Do đó, chúng có thể giúp phát hiện các hợp chất ngay cả khi chúng chưa được tổng hợp hay phân lập. QSAR thực sự bắt đầu được nghiên cứu từ những năm 60 của thế kỷ XX bởi Corwin Hansch và các đồng sự<sup>[1]</sup>. Hiện nay phương pháp này đang được ứng dụng rất rộng rãi trên thế giới do các ưu điểm của chúng so với các phương pháp truyền thống.

Proteasomes là phức hợp protein bên trong tất cả các tế bào nhân chuẩn với chức năng chính là tiêu hủy các protein không cần thiết hoặc bị hỏng trong quá trình phân giải protein. Quá trình phân giải protein này được thực hiện trong hệ thống ubiquitin-proteasome (UPS). Rối loạn chức năng bộ máy ubiquitin hoặc các hoạt động phân giải protein của proteasome gây ra nhiều bệnh lý khác nhau, như ung thư, viêm, nhiễm virus, rối loạn hệ thống thần kinh trung ương, rối loạn chức năng trao đổi chất. Các chất ức chế UPS (UPS-I) đã được phát triển thành thuốc và sử dụng hiệu quả đối trong một số bệnh lý,

nhưng hầu hết chúng đều có tác phụ và nhiều bệnh nhân có biểu hiện kháng thuốc. Do đó, việc tìm kiếm các hợp chất mới là cần thiết, đặc biệt cho việc phát triển các thuốc chữa ung thư thể hệ mới. Các mô hình QSAR được áp dụng trong quá trình này giúp tăng hiệu quả, tiết kiệm chi phí và thời gian. Các mô hình này là công cụ để dự đoán hoạt tính của các chất và để xác định các đặc điểm cấu trúc quan trọng cho hoạt tính. Do đó, nghiên cứu này được thực hiện nhằm mục tiêu xây dựng và đánh giá các mô hình toán học QSAR mới để sàng lọc và phát hiện các hợp chất UPS-I chỉ từ thông tin cấu trúc phân tử.

### Nguyên liệu và phương pháp nghiên cứu

#### Nguyên liệu

Cơ sở dữ liệu (CSDL) được thu thập từ các nghiên cứu sử dụng thí nghiệm được mô tả trong PubChem BioAssay (AID 2486), với thông tin về tên, cấu trúc, số hiệu hợp chất (CID). Sau quá trình làm sạch, loại trừ các muối và các chất vô cơ, CSDL được sử dụng để xây dựng các mô hình gồm 705 hợp chất trong đó: 258 hợp chất UPS-I (*active*) và 447 hợp chất không ức chế UPS (*inactive*).

#### Phương pháp nghiên cứu

##### Xây dựng các mô hình QSAR

##### Định lượng hóa cấu trúc và tác dụng sinh học

Mô hình QSAR biểu diễn mối liên hệ định lượng giữa cấu trúc và tác dụng sinh học của các chất. Để xây dựng được các mô hình này thì cả cấu trúc và tác dụng sinh học đều phải được định lượng hóa. Các cấu trúc được mã hóa thành các tham số phân tử đặc trưng cho từng hợp chất, còn tác dụng sinh học có thể là các giá trị định lượng thu được từ các thí nghiệm thực nghiệm hay được mã hóa thành 1 và -1 trong các mô hình phân loại có hay không có hoạt tính sinh học.

*Tính toán các tham số phân tử đặc trưng cho cấu trúc (tham số cấu trúc)*

## ● Nghiên cứu - Kỹ thuật

Tham số cấu trúc là một số thu được từ quá trình toán học và logic mạng thông tin được mã hóa trong cấu trúc hóa học [2]. Trong nghiên cứu này, các tham số được tính toán sử dụng phần mềm TOMOCOMD [3]. Các nguyên tử, được đặc trưng bằng các tính chất nguyên tử như khối lượng nguyên tử (M), độ phân cực (P), thế tích Van der Waals (V), độ âm điện Mulliken (K).

### Thiết kế tập huấn luyện và tập kiểm tra

CSDL ban đầu được phân chia thành tập huấn luyện (*training set* - TS) để xây dựng các mô hình và tập kiểm tra (*prediction set* - PS) để đánh giá khả năng ngoại suy của mô hình bằng kỹ thuật phân tích cụm (*cluster analysis* - CA) [4] của chương trình STATISTICA 9.0 [5]. Phương pháp này chia mỗi tập các hợp chất *active* và *inactive* thành 11 cụm. Từ từng cụm trên, lấy ra 75% các hợp chất để xây dựng TS và 25% các hợp chất để xây dựng PS.

### Triển khai các mô hình QSAR

Các kỹ thuật phân tích sự khác biệt tuyến tính (LDA) [6] của chương trình STATISTICA 9.0 [5] và học máy (*machine learning*) của WEKA [7] được sử dụng để phát triển các mô hình QSAR. Các kỹ thuật học máy bao gồm: mạng nơ-ron nhân tạo lan truyền thẳng nhiều lớp (MLP) [8] và máy vector hỗ trợ (SVM) [9]. Để triển khai các kỹ thuật này, bước đầu tiên và tương đối quan trọng đó là lựa chọn các tham số phân tử.

Lựa chọn tham số phân tử: Đối với LDA, nghiên cứu sử dụng phương pháp tập hợp con tốt nhất (*best subset*) để lựa chọn các tham số phân tử cho việc phát triển mô hình. Đối với các phương pháp học máy, phương pháp bộ lọc (*filter*) và đóng gói (*wrapper*) của WEKA [7] được sử dụng; đồng thời tập hợp các tham số phân tử được lựa chọn cho mô hình LDA cũng được sử dụng để phát triển mô hình học máy. Nhiều mô hình MLP và SVM được xây dựng với nhiều tập hợp tham số phân tử khác nhau và những mô hình tốt nhất được giữ lại. Các mô hình được lựa chọn theo nguyên lý parsimony "*Occam's Razor*" (đơn giản nhưng tốt) và dựa vào các thông số thống kê là: độ chính xác hay tỉ lệ phần trăm các chẩn đoán chính xác (*Accuracy* - Q), độ nhạy (*Sensitivity* - Se), độ đặc hiệu (*Specificity* - Sp) và tỉ lệ dương tính giả (*False Positive Rate* -  $FP_{rate}$ ). Trong đó:

$Độ\ nhạy = \frac{số\ lượng\ dương\ tính\ thật}{(số\ lượng\ dương\ tính\ thật + số\ lượng\ âm\ tính\ giả)}$

$Độ\ đặc\ hiệu = \frac{số\ âm\ tính\ thật}{(số\ âm\ tính\ thật + số\ dương\ tính\ giả)}$

Các hợp chất được phân loại theo kết quả của xác suất hậu nghiệm (*posterior probability*). Trong trường hợp này, hiệu của các xác suất hậu nghiệm được tính toán theo phương trình:

$$\Delta P\% = [P(\text{có hoạt tính}) - P(\text{không hoạt tính})] \times 100$$

Trong đó:

*P* (có hoạt tính) là xác suất mà mô hình phân loại một hợp chất là có hoạt tính ức chế UPS, còn *P* (không hoạt tính) là xác suất mà mô hình phân loại một hợp chất là không có hoạt tính.

Như vậy, nếu một hợp chất  $\Delta P\% > 0$ , sẽ được dự đoán là có hoạt tính và  $\Delta P\% < 0$ , sẽ được dự đoán là không có hoạt tính.

### Đánh giá các mô hình QSAR

Các nguyên tắc của Tổ chức Hợp tác và Phát triển Kinh tế Thế giới (*Organization for Economic Cooperation and Development* - OECD) [10] (quy định về thực hành tốt xây dựng các mô hình QSAR) được sử dụng để đánh giá các mô hình. Theo đó, một mô hình được đánh giá là đủ tiêu chuẩn và được phép dùng để sử dụng trong các nghiên cứu khi thỏa mãn: 1) Có đích xác định (*defined endpoint*); 2) Sử dụng một thuật toán rõ ràng; 3) Có miền cấu trúc ứng dụng rõ ràng; 4) Thỏa mãn các đánh giá nội trên tập huấn luyện để kiểm tra mức độ khớp (*goodness-of-fit*), độ mạnh và ổn định của mô hình (*robustness*) và các đánh giá ngoại, tức là có khả năng dự đoán cao (đánh giá khả năng dự đoán của mô hình trên tập kiểm tra).

## Kết quả và thảo luận

### Mô hình toán học phân loại dự đoán khả năng ức chế UPS

Sử dụng kỹ thuật phân tích cụm CA, CSDL ban đầu được chia ra làm TS gồm 529 hợp chất và PS gồm 176 hợp chất. Ba mô hình phân loại được xây dựng trên TS sử dụng các kỹ thuật xác suất thống kê và trí tuệ nhân tạo.

### Mô hình LDA

Sử dụng phương pháp tập hợp con tốt nhất của STATISTICA, 10 tham số phân tử TOMOCOMD-CARDD (chỉ số đôi, chỉ số tuyến tính, chỉ số song tuyến tính) được lựa chọn để phát triển mô hình. Mô hình phân loại chính xác 85,82% các hợp chất của TS với tỉ lệ dương tính giả ( $FP_{rate}$ ) là 11,93%.

## ● Nghiên cứu - Kỹ thuật

$$\begin{aligned} \text{Hoạt tính UPS-I} = & 0.187 + 0.015^{\text{P}} q_{\text{B}}(x) - 6.238 \times 10^{-6\text{V}} q_{\text{SL}}^{\text{H}}(x_{\text{E,H}}) - 5.378 \times 10^{-4\text{Mg}} q_{\text{S}}(x) \\ & - 1.117 \times 10^{-11\text{V}} f_{\text{EL}}^{\text{H}}(x_{\text{E,H}}) + 2.714 \times 10^{-8\text{K}} f_{\text{S}}(x) + 4.523 \times 10^{-6\text{P}} f_{\text{SL}}(x_{\text{E}}) \\ & + 3.429 \times 10^{-13\text{VP}} b_{\text{SL}}^{\text{H}}(x_{\text{E}}, y_{\text{E}}) - 5.072^{\text{MVs}} b_{\text{AL}}^{\text{H}}(x_{\text{E,H}}, y_{\text{E,H}}) \\ & + 1.322^{\text{VKs}} b_{\text{AL}}(x_{\text{E}}, y_{\text{E}}) - 1.981 \times 10^{-9} \text{PKs} b_{\text{AL}}(x_{\text{E}}, y_{\text{E}}) \end{aligned}$$

### Mô hình MLP

Nghiên cứu sử dụng các mạng nơ ron nhân tạo gồm 3 lớp (một lớp đầu vào, một lớp đầu ra và một lớp ẩn). Mạng nơ ron được huấn luyện sử dụng hàm sigmoid là hàm kích hoạt. Số lượng các node ẩn được lựa chọn theo phương pháp "thử và lỗi", sau đó so sánh chất lượng các mô hình thu được. Các phương pháp lựa chọn tham số phân tử theo mô tả trong phần phương pháp nghiên cứu được sử dụng. Mô hình MLP thu được có cấu hình 9-7-2, với 7 node ở lớp ẩn. Mô hình này phân loại chính xác 88,06 % các hợp chất của TS với tỉ lệ dương tính giả ( $FP_{\text{rate}}$ ) là 8,44%.

### Mô hình c-SVM

Tất cả các hàm nhân (*kernel*) có trong WEKA đều được thử. Đối với từng hàm, từng thông số được kiểm tra sử dụng phương pháp "thử và lỗi", sau đó so sánh chất lượng các mô hình thu được. Kết quả là hàm nhân *BF* sử dụng bán kính (*radial basis function - RBF*) cho kết quả tốt nhất với 10 tham số phân tử với  $\gamma = 2$  và  $C = 1000$ . Mô hình SVM sử dụng 198 véc tơ hỗ trợ, trong đó 105 thuộc phân lớp các hợp chất UPS-I và 93 thuộc phân lớp còn lại (các hợp chất không ức chế UPS). Mô hình này phân loại chính xác 89,4% các hợp chất của TS với tỉ lệ dương tính giả là 5,8%.

**Đánh giá các mô hình QSAR xây dựng được**

Để các mô hình QSAR được thông qua bởi các cơ quan quản lý, chúng phải thỏa mãn các nguyên tắc của OECD (tổ chức kinh tế thế giới) [11].

### Tiêu chí 1 (có đích xác định)

Trong nghiên cứu này, đích là sự phân loại các hợp chất thành 2 nhóm: chất UPS-I và chất không ức chế UPS sử dụng thí nghiệm được mô tả theo PubChem BioAssay (AID 2486). Do đó, đích ở đây là hoàn toàn xác định.

### Tiêu chí 2 (thuật toán rõ ràng)

Phương pháp xây dựng các mô hình trong nghiên cứu này được mô tả ở trên. Mô hình

toán học được phát triển sử dụng kỹ thuật LDA, MLP và SVM để biểu diễn mối quan hệ giữa đích (ức chế UPS) và các tham số phân tử TOMOCOMD. Các kỹ thuật này đã được mô tả khá kỹ trong các tài liệu tham khảo, các chỉ số TOMOCOMD cũng đã được định nghĩa rõ ràng [12-14]. Vì thế mô hình có thể được xây dựng lại và được sử dụng để phát hiện các hợp chất UPS-I mới.

### Tiêu chí 3 (miền cấu trúc xác định)

Sử dụng kết quả của chương trình *AMBITDiscovery* [15], có 34 hợp chất của TS nằm ngoài miền cấu trúc của các mô hình. Kết quả này có thể chấp nhận được theo như công bố của các nghiên cứu trước đây [16].

### Tiêu chí 4

#### Đánh giá nội

Kết quả của quá trình đánh giá nội sử dụng phương pháp đánh giá chéo (*cross validation - CV*). Trong trường hợp này, 10% các hợp chất của TS được để ở ngoài, sau đó mô hình được xây dựng lại với 90% hợp chất còn lại, rồi sử dụng mô hình này dự đoán 10% các hợp chất bị bỏ ra ngoài đó. Kết quả các đánh giá chéo được biểu diễn trong bảng 1.

**Bảng 1: Các thông số thống kê của 3 mô hình trong CV 10%**

Mô hình	Q %	Se %	Sp %	$FP_{\text{rate}}$ %
LDA	85,26	84,63	85,28	14,13
MLP	87,03	84,06	88,95	10,09
SVM	88,90	86,34	90,64	8,62

Theo như các kết quả này, mô hình tương đối ổn định và không bị phụ thuộc vào bất kỳ hợp chất hay nhóm hợp chất nào của TS.

#### Đánh giá ngoại

Sử dụng mô hình để dự đoán các hợp chất trên tập kiểm nghiệm. Tập kiểm nghiệm bao gồm những phân tử mà mô hình chưa bao giờ nhìn thấy trong quá trình huấn luyện. Kết quả của quá trình đánh giá ngoại được mô tả trong bảng 2.

## ● Nghiên cứu - Kỹ thuật

**Bảng 2:** Các thông số thống kê của 3 mô hình trên tập PS

Mô hình	Q %	Se %	Sp %	FP <sub>min</sub> %
LDA	81,62	80,08	82,10	16,88
MLP	85,17	82,92	86,36	12,66
SVM	83,49	82,16	83,91	15,23

Theo như quan sát trên bảng 2, các mô hình có khả năng dự đoán tương đối cao (mô hình MLP dự đoán chính xác 85,17%), và khả năng nhầm lẫn khi kết luận một chất không có hoạt tính là có hoạt tính tương đối thấp (đối với mô hình MLP là 12,66%).

### Tiêu chí 5

Giải thích cơ chế (nếu có thể). Trong trường hợp này, do đặc thù của CSDL và các tham số phân tử được sử dụng nên nghiên cứu này không tiến hành giải thích cơ chế.

Trong nghiên cứu, các mô hình QSAR mới được xây dựng có khả năng dự đoán tương đối cao. Các kết quả trên tập huấn luyện là 81,62%, 83,49% và 85,17% lần lượt cho các mô hình LDA, MLP và SVM, và khả năng nhầm lẫn một chất không có khả năng ức chế thành có khả năng cao nhất chỉ có 16,88%. Điều này, chứng tỏ mô hình có khả năng phát hiện/dự đoán tốt. Các mô hình sử dụng LDA có chất lượng kém hơn các mô hình sử dụng các kỹ thuật học máy. Tuy nhiên, các mô hình LDA có ưu thế là chỉ rõ trọng số của từng tham số phân tử cho tác dụng sinh học, điều mà các mô hình học máy không làm được. Cuối cùng, các mô hình mới được đánh giá cẩn thận theo các tiêu chí của tổ chức OECD. Tùy từng mục đích, các mô hình này có thể được sử dụng đơn lẻ hay kết hợp để sàng lọc hoạt tính ức chế UPS cho các hợp chất.

### Kết luận

Trong nghiên cứu này, 3 mô hình QSAR mới đã được xây dựng thành công để biểu diễn mối quan hệ định lượng giữa cấu trúc và tác dụng ức chế UPS sử dụng các kỹ thuật xác xuất thống kê (LDA) và trí tuệ nhân tạo (MLP, SVM). Các mô hình xây dựng được đạt độ chính xác cao và đều thỏa mãn các đánh giá theo các nguyên tắc của OECD. Do đó, các mô hình này có thể được sử dụng bởi bất kỳ trung tâm hay viên nghiên cứu nào trong quá trình tìm kiếm hợp chất mới UPS-I

mà chỉ cần thông tin về cấu trúc phân tử. Việc sử dụng những các mô hình này nói riêng cũng như các phương pháp trong nghiên cứu này nói chung có thể giúp tiết kiệm thời gian và tiền bạc trong việc sàng lọc nhằm phát hiện các hợp chất dẫn đường mới trong quá trình nghiên cứu và phát triển các thuốc chữa ung thư thế hệ mới.

### Tài liệu tham khảo

1. Hansch C. and T. Fujita (1964), "Rho Sigma pi analysis: A method for the correlation of biological activity and chemical structure.", *J. Am. Chem. Soc.*, 86, p. 1616-1626.
2. Todeschini R. and V. Consonni (2000), *Handbook of Molecular Descriptors 1st ed. Methods and Principles in Medicinal Chemistry*, ed. R. Mannhold, H. Kubinyi, and H. Timmerman Vol. 11. 2000, D-69469 Weinheim, Federal Republic of Germany: WILEY-VCH Verlag GmbH. 667.
3. Marrero-Ponce Y. and V. Romero (2002), *TOMOCOMD software (TOPOlogical MOlecular COmputer Design) for Windows*, Central University of Las Villas, Villa Clara. Cuba. p. (<http://tomocomd.com/>).
4. Mc Farland J. W. and D. J. Gans (1995), *Cluster Significance Analysis, in Chemometric Methods in Molecular Design*, H. Waterbeemd, Editor, VCH Publishers: Weinheim, Ger. p. 295-307.
5. *STATISTICA (data analysis software system) V 9. 2009*, Tulsa, OK. StatSoft Inc.
6. H. v.d.W. (1995), *Discriminant Analysis for Activity Prediction, in Chemometric Methods in Molecular Design*, v.d.W. H. Editor, VCH Publishers: Weinheim. p. 265-288
7. H. Witten I. and E. Frank (2005), *Data Mining. Practical Machine Learning Tools and Techniques*. Second ed. Morgan Kaufmann Series in Data Management Systems, ed. J. Gray and M. Research, Morgan Kaufmann Publishers.
8. Svozil D., V. Kvasnicka and J. Pospichal (1997), "Introduction to multi-layer feed-forward neural network", *Chemom. Intell. Lab. Syst.*, 39(1), p. 43-62.
9. Cortes C. and V. Vapnik (1995), "Support-Vector Networks", *Mach. Learn.*, 20, p. 273-297.
10. *OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*. in 37<sup>th</sup> Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology. 17-19 November (2004) Paris.
11. OECD, in *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, Series on Testing and Assessment No. 69 (2007) Paris p. 154pp.
12. Marrero-Ponce Y. (2004), "Total and local