Bioorganic & Medicinal Chemistry 16 (2008) 6448–6459

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Chemometric and chemoinformatic analyses of anabolic and androgenic activities of testosterone and dihydrotestosterone analogues

Yoanna María Alvarez-Ginarte [a,b], Rachel Crespo-Otero [b], Yovani Marrero-Ponce [c,*,†], Pedro Noheda-Marin [d], Jose Manuel Garcia de la Vega [e], Luis Alberto Montero-Cabrera [b], José Alberto Ruiz García [a], José A. Caldera-Luzardo [f], Ysaias J. Alvarado [f]

[a] Pharmaceutical Chemistry Center, 16042 La Habana, Cuba
[b] Laboratory of Theoretical and Computational Chemistry, University of Havana, 10400 La Habana, Cuba
[c] Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, Poligon la Coma s/n (detras de Canal Nou), PO Box 22085, E-46071 Valencia, Spain
[d] Instituto de Química Orgánica (IQOG), Consejo Superior de Investigaciones Científicas (CSIC), 28006 Madrid, Spain
[e] Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain
[f] Laboratorio de Electrónica Molecular, Departamento de Química, Modulo II, Grano de Oro, Facultad Experimental de Ciencias, La Universidad del Zulia (LUZ), Venezuela

ARTICLE INFO

Article history:
Received 18 February 2007
Revised 29 March 2008
Accepted 1 April 2008
Available online 7 April 2008

Keywords:
QSAR model
Anabolic and androgenic activities
Testosterone and dihydrotestosterone steroid analogues
Genetic algorithm
Quantum and physicochemical molecular descriptor

ABSTRACT

Predictive quantitative structure–activity relationship (QSAR) models of anabolic and androgenic activities for the testosterone and dihydrotestosterone steroid analogues were obtained by means of multiple linear regression using quantum and physicochemical molecular descriptors (MD) as well as a genetic algorithm for the selection of the best subset of variables. Quantitative models found for describing the anabolic (androgenic) activity are significant from a statistical point of view: R^2 of 0.84 (0.72 and 0.70). A leave-one-out cross-validation procedure revealed that the regression models had a fairly good predictability [q^2 of 0.80 (0.60 and 0.59)]. In addition, other QSAR models were developed to predict anabolic/androgenic (A/A) ratios and the best regression equation explains 68% of the variance for the experimental values of AA ratio and has a rather adequate q^2 of 0.51. External validation, by using test sets, was also used in each experiment in order to evaluate the predictive power of the obtained models. The result shows that these QSARs have quite good predictive abilities (R^2 of 0.90, 0.72 (0.55), and 0.53) for anabolic activity, androgenic activity, and A/A ratios, respectively. Last, a Williams plot was used in order to define the domain of applicability of the models as a squared area within ±2 band for residuals and a leverage threshold of $h = 0.16$. No apparent outliers were detected and the models can be used with high accuracy in this applicability domain. MDs included in our QSAR models allow the structural interpretation of the biological process, evidencing the main role of the shape of molecules, hydrophobicity, and electronic properties. Attempts were made to include lipophilicity (octanol–water partition coefficient (log P)) and electronic (hardness (η)) values of the whole molecules in the multivariate relations. It was found from the study that the log P of molecules has positive contribution to the anabolic and androgenic activities and high values of η produce unfavorable effects. The found MDs can also be efficiently used in similarity studies based on cluster analysis. Our model for the anabolic/androgenic ratio (expressed by weight of levator ani muscle, LA, and seminal vesicle, SV, in mice) predicts that the 2-aminomethylene-17α-methyl-17β-hydroxy-5α-androstan-3-one (**43**) compound is the most potent anabolic steroid, and the 17α-methyl-2β,17β-dihydroxy-5α-androstane (**31**) compound is the least potent one of this series. The approach described in this report is an alternative for the discovery and optimization of leading anabolic compounds among steroids and analogues. It also gives an important role to electron exchange terms of molecular interactions to this kind of steroid activity.

1. Introduction

Testosterone is the primary male sex hormone. Anabolic and androgenic steroids are synthetics derived from testosterone, which is secreted by the testicles as well as, in a small quantity, by the ovaries and the suprarenal cortex. The masculine (androgenic) effects are coupled with an anabolic effect (tissue building). Testosterone is converted to dihydrotestosterone upon interaction

* Corresponding author. Tel.: +53 42 281192, +53 42 281473 [Cuba], +34 963544431 [Spain]; fax: +53 42 281130, +53 42 281455 [Cuba], +34 963543274 [Spain].
E-mail addresses: ymarrero77@yahoo.es, ymponce@gmail.com, yovanimp@uclv.edu.cu (Y. Marrero-Ponce).
URL: http://www.uv.es/yoma/ (Y. Marrero-Ponce).
† On leave from the Unit of Computer-Aided Molecular 'Biosilico' Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry–Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba.

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.04.001

with the 5-α reductase enzyme; more specifically, this enzyme removes the C_{4-5} double bond of testosterone by the addition of hydrogen atoms to its structure. The removal of this more labile π bond is important, as in this case it creates a steroid that binds to the androgen receptor much more avidly than testosterone.[1]

The general chemical structure of testosterone is based upon the androstane C19 steroid, consisting of the fused four-ring steroid nucleus (17 carbon atoms, rings A–D) and the two axial methyl groups (carbon 18 and 19) and the A/B and C/D ring junctions (see Fig. 1).[2]

Anabolic steroids cause retention of nitrogen, calcium, potassium, chloride, phosphate, and water, as well as the growth of bones.[3] These drugs are used in the fast recovery from protein-wasting disorders. In HIV patients, anabolic steroids are used to regain lean muscle mass, as well as to prevent organ failure and secondary immune dysfunction. These compounds have proven to be an effective oral therapy to promote weight gain after extensive surgery, chronic infections, and severe trauma.[4] They are indicated in the treatment of anemia caused by deficient red-cell production, chronic obstructive pulmonary disease (attributed to emphysema as well as bronchitis) and metastatic cancer.[5,6]

The goal of researchers in the anabolic steroid golden age (1935–1965) was to synthesize a compound that retained a high degree of anabolic activity coupled with a vastly diminished andro-

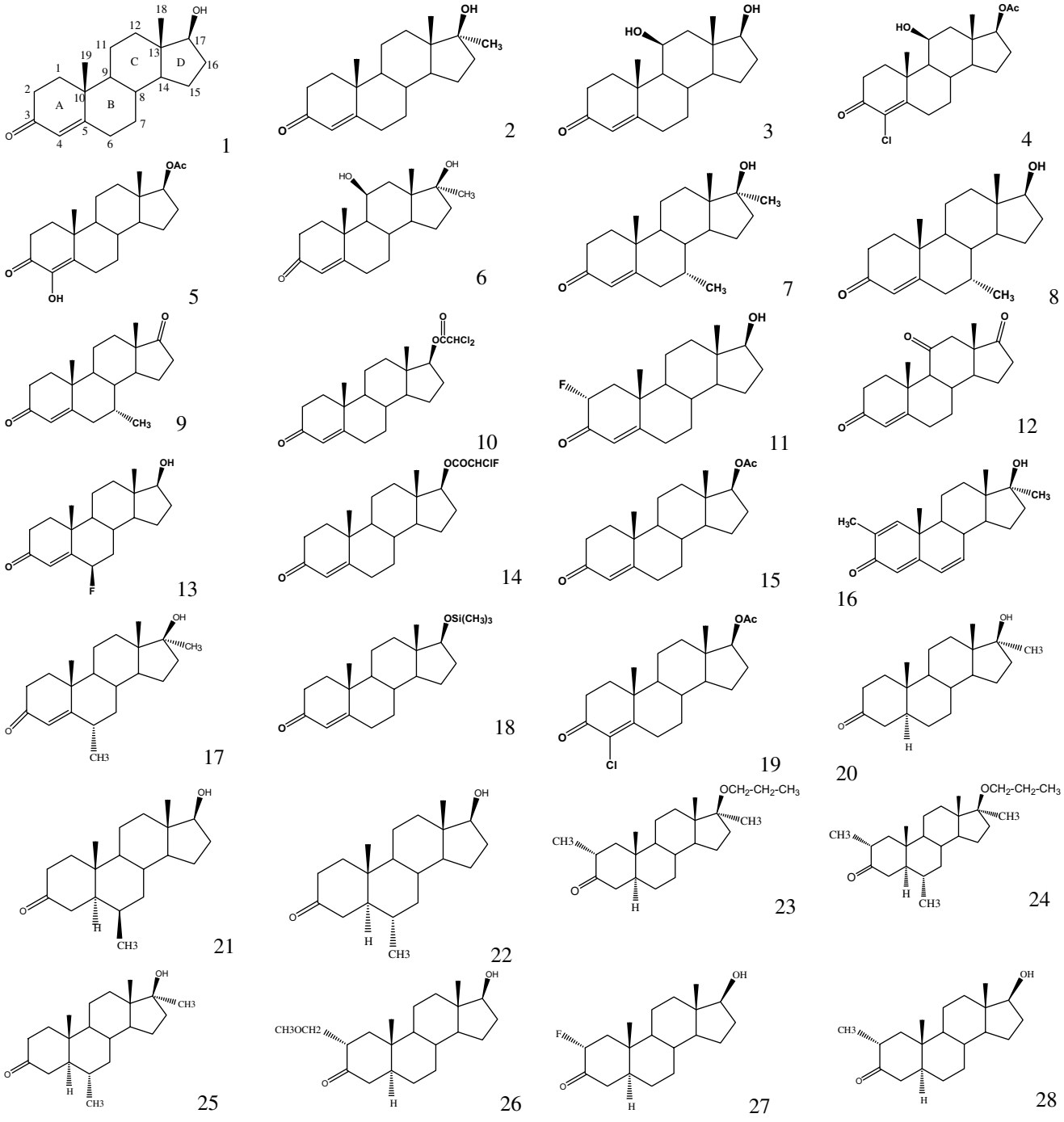

Figure 1. Testosterone and dihydrotestosterone derivatives.

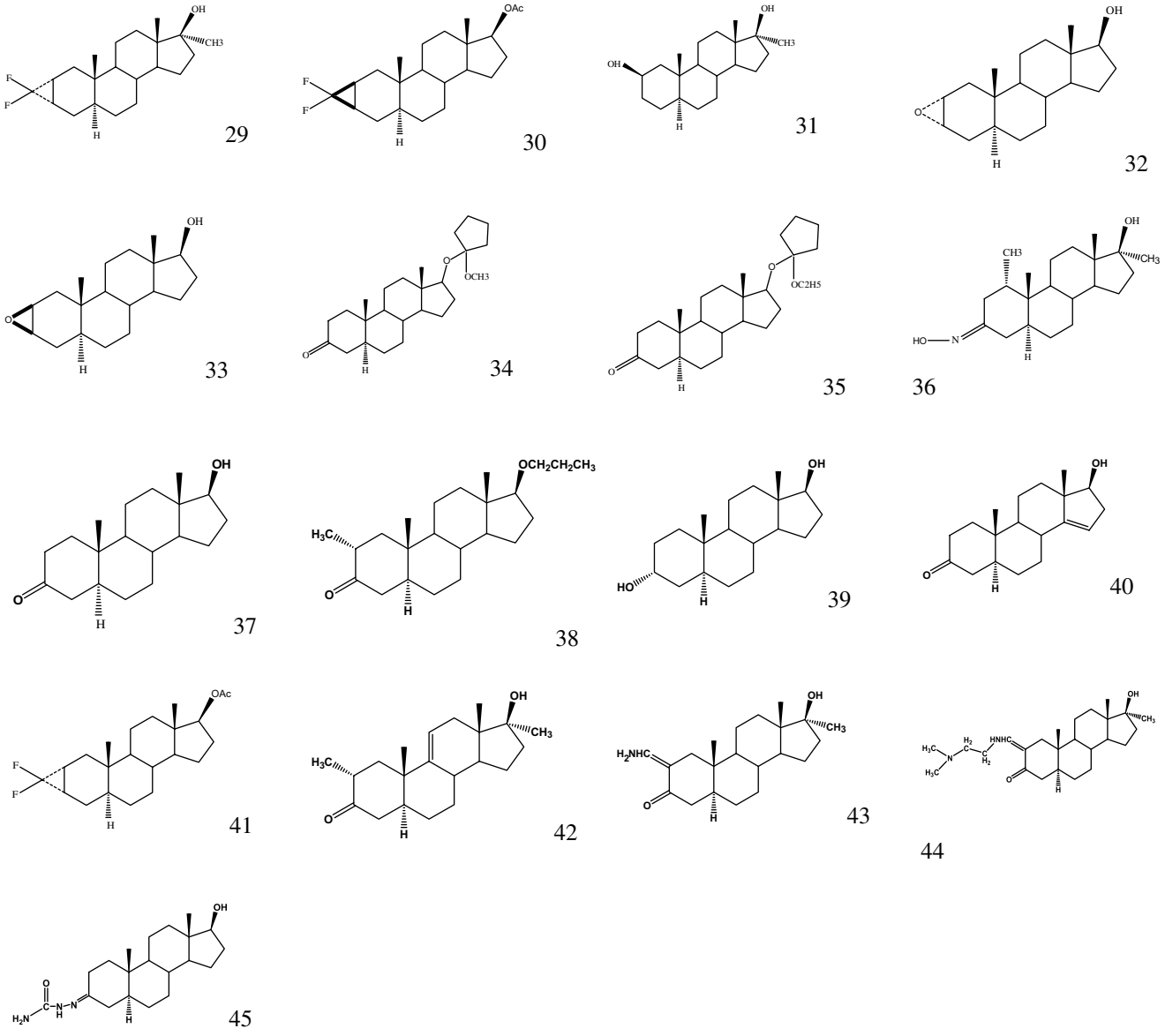

Figure 1 (*continued*)

genic activity. In integrating both measures the anabolic index is used, which relates the ratio of anabolic to androgenic response for a given steroid. If an anabolic index is greater than one it indicates a higher tendency for anabolic effect and, therefore, the drug is classified as an anabolic steroid. A measure lower than one, in turn, assesses the steroid as androgenic.[7] At present, the commercially available anabolic compounds were synthesized during the 30 years of anabolic steroid research. Some authors have said that it is not possible even to generalize which chemical modifications will reinforce the anabolic activity with a simultaneous decrease in the androgenic activity.[8,9]

Vida in 1969 collected a database of steroids with anabolic and androgenic activities (AASs) evaluated in vivo.[10] At present, there is no other standardized reference, where the values of the anabolic and androgenic activities be reported for this kind of molecules. In the database of Vida the AASs appear contained in different steroids families: 17β-hydroxy-5α-androstane, 4,5α-dihydrotestosterone, testosterone, and 19-nor-testosterone derivatives.

Recently, we report QSAR models for congeneric series of AASs: 17β-hydroxy-5α-androstane,[11] and 4,5α-dihydrotestosterone.[12] The predictive approach reported for the dihydrotestosterone derivatives was improving with regard to the article for the 17β-hydroxy-5α-androstane derivatives. In the present report the approach is similar to the one described in dihydrotestosterone derivatives; but the interpretation of the model QSAR obtained for this family is different to the one obtained in the previously reported families including some structural considerations in another series of compounds. We report too a robust *biosilico* model of linear discriminant analysis (LDA).[13] This model was used to analyze the anabolic/androgenic activity of structurally diverse steroids and to discover novel AASs, as well as to give a structural interpretation of their anabolic–androgenic ratio (AAR). We selected a group of 366 steroids[10] having as much structural variability as possible and containing the four families of compounds included in the Vida databases. The LDA technique allowed us to generate general models, capable of discriminating between steroids with high and moderate-low AAR.

The general idea of our work is to develop general models of classification for steroids with high and moderate-low AAR and subsequently quantify their anabolic and/or androgenic activities in the models of multi-linear regression (MLR) according to the family that belongs to each molecule selected. Finally, our

approach could help with the future successful identification of 'real' or 'virtual' AAS steroids.

The main aim of this report was to develop QSAR models in a testosterone and dihydrotestosterone steroid families by using quantum and physicochemical molecular descriptors (MDs) as well as a genetic algorithm as method for the selection of the best set of variables.

2. Results and discussion

2.1. Construction of training and test sets using hierarchical cluster analyses

It is well known that the quality of a regression model is highly dependent on the quality of the selected data set. The most critical aspect for constructing the training set is to warrant molecular diversity enough on it. Taking this into account, we selected a data set of 45 steroids (with both anabolic and androgenic activities) having a great structural variability. In order to demonstrate the structural diversity of this data set, we performed a hierarchical CA of these chemicals.[49–51] The hierarchical clustering approach finds a hierarchy of objects represented by a number of MDs. The dendrogram given in Figure 2, using the Euclidean distance (*X*-axis) and the complete linkage (*Y*-axis), illustrates the results of the *k*-NNCA developed in this set. As it can be seen in the dendrogram, there are a great number of different subsets, which prove the molecular variability of the selected chemicals in these databases.

Furthermore, this procedure permits selecting compounds for the training and test sets, in a representative way, in all levels of the linking distance. The main idea of this procedure consists in making a partition of chemicals in several statistically representative classes of compounds. This procedure ensures that any chemical class (as determined by the clusters) will be represented in both compound series. This 'rational' design of training and predicting series allowed us to design both sets that are representative of the whole 'experimental universe'. Moreover, the selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. From these 45 steroids, 36 (80% of the data) were chosen at random to form the training set. The great structural variability of the selected training set makes possible the discovery of lead compounds. The remaining subseries composed of 9 steroids (20% of the data) was prepared as a test set for the external cross-validation of the models. These chemicals were never used in the development of the QSAR models. Figure 3 graphically illustrates the above-described procedure, where cluster analysis was performed to select a representative sample for the training and test sets.

It should be remarked that, recently, several authors had developed a classification of steroids using CA,[14–16] but this analysis has been presented only for the benchmark steroids with the corresponding globulin affinity.

A complete discussion of the clustering is out of the context of the present study, but several interesting features should be noted. Furthermore, a few relative easy instances, where the similarity is relatively obvious, may be identified from direct inspection of the molecular structures. The dendrogram should reflect this high logic-visual similarity. A very obvious case lies in molecules **22** and **25**. These have very similar physicochemical and quantum chemical properties and, in fact, it is identified in the dendrogram. A similar case exists between molecules **29** and **30**, where the only difference is the difluoro-methylen group in the same position. The more interesting cases arise with molecules **32** and **33**. They only differ in the stereoisomerisms of the epoxy group in the same position. The dendrogram reflects this high similarity. Compounds **40** and **31** also showed high similarity. These four molecules can be taken as outliers. Therefore, another interesting observation is that some chemicals remain outliers of any cluster for a very long time along the clustering procedure. The most prominent example is four molecules, which are outliers for every cluster, except in the ultimate stage of the process, where by construction they have to

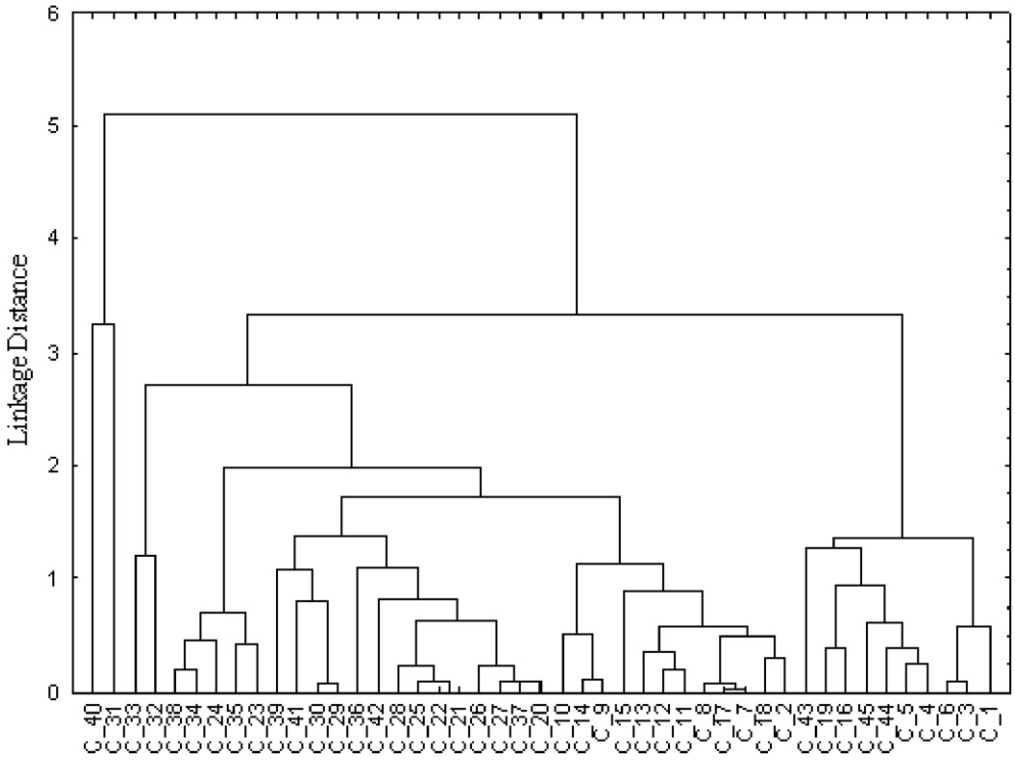

Figure 2. A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of 45 steroids used in the training and prediction sets of the present work.

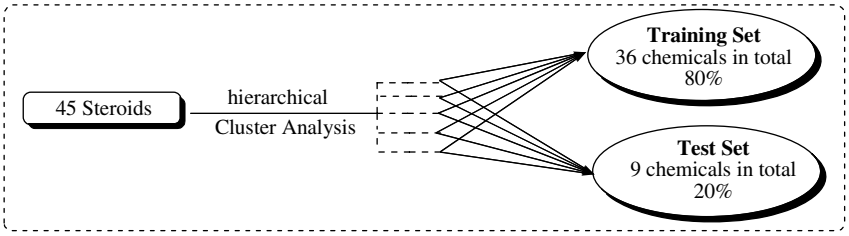

Figure 3. General algorithm used to design training and test sets throughout *k*-NNCA.

be taken in a conclusive cluster. This will be well illustrated in this report in the QSAR development of the different biological activities.

2.2. Development and validation of the QSAR models

2.2.1. QSAR models for anabolic activity (log(1/LA))

The training set of testosterone and dihydrotestosterone derivatives includes a set of compounds formed by the steroids: **3**, **4**, **6**, **8, 9, 11–21, 23, 24, 26–33**, and **36–45** (n = 36, see Fig. 1 and Table 1 for more details).

The variables selected by the genetic algorithm as the best models of anabolic activity are shown in Eq. 1. In order to compare the external predictions corresponding to Eq. 1 the steroids: **1, 2, 5, 7, 10, 22, 25, 34**, and **35** were chosen as test set (n = 9, see Fig. 1 and Table 1 for more details). The obtained QSAR model is given below together with the statistical parameters of both learning and prediction sets:

$$\log(1/\mathrm{LA}) = +0.52(\pm0.05)\log P - 0.87(\pm0.07)n + 4.30(\pm0.38) \quad (1)$$

n = 36, R^2 = 0.84, q^2 = 0.80, s = 0.25, F = 81.68, p < 0.001

Test set: n = 9, R^2 = 0.90, s = 0.12, F = 72.49.

The R^2 (R-square statistic or determination coefficient) indicates that the model explains 84% of the variance for the experimental values of log(1/LA). The model has a q^2 of 0.80. This value of $q^2 > 0.5$ can be considered as a proof of the high predictive ability of the model as well as the good prediction of the test set (R^2 = 0.90). Table 4 shows the correlation between the observed and predicted anabolic activities from Eq. 1.

2.2.2. QSAR models for androgenic activity: ventral prostate (log(1/VP)) and seminal vesicle (log(1/SV))

The VP and SV training set of testosterone and dihydrotestosterone derivatives consists of the following chemicals: **3**, **4**, **6**, **8, 9, 11–21, 23, 24, 26–33**, and **36–45** (n = 36, see Fig. 1 and Table 1 for more details). The variables selected by the genetic algorithm as the best models of androgenic activity are shown in Eqs. 2 and 3. In order to compare the predictive ability of VP and SV steroid-based models we used a test set composed of by nine compounds (**1, 2, 5, 7, 10, 22, 25, 34**, and **35**, see Fig. 1 and Table 1 for more details). The QSAR models obtained for description of VP and SV, as well as their statistical parameters of both training and test sets, are depicted below as Eqs. 2 and 3, respectively:

$$\log(1/\mathrm{VP}) = 0.42(\pm0.07)\log P - 0.90(\pm0.10)n + 4.58(\pm0.53) \quad (2)$$

n = 36, R^2 = 0.72, q^2 = 0.60, s = 0.36, F = 33.70, p < 0.001

Test set: n = 9, R^2 = 0.72, s = 0.23, F = 18.76.

$$\log(1/\mathrm{VS}) = 0.49(\pm0.08)\log P - 0.90(\pm0.11)n + 4.30(\pm0.57) \quad (3)$$

n = 36, R^2 = 0.70, q^2 = 0.59, s = 0.37, F = 39.41, p < 0.001

Test set: n = 9, R^2 = 0.55, s = 0.32, F = 8.61.

The R^2 for Eqs. 2 and 3 were 0.72 and 0.70, correspondingly, so these models explained the 72% and 70% of the variance for the experimental values of logVP and logSV. These models, Eqs. 2 and 3, also showed high stability to data variation in the LOO cross-validation procedure (q^2 = 0.60 and q^2 = 0.59, respectively) and a good *predictive* square correlation coefficient of 0.72 and 0.55, correspondingly. Tables 3 and 4 show the correlation between observed and predicted values of androgenic activities for a Eqs. 2 and 3, respectively (Tables 5 and 6).

2.2.3. QSAR models of the anabolic/androgenic ratio: (log(1/LA)/log(1/VP)) and (log(1/LA)/log(1/SV))

In order to design compounds which retain a high degree of anabolic activity and a vastly diminished androgenic activity, anabolic/androgenic (A/A) ratios: (log(1/LA)/log(1/VP)) and (log(1/LA)/log(1/SV)) of testosterone and dihydrotestosterone steroids were estimated. The A/A ratios were quantified using the anabolic and androgenic activity values shown in Table 1. The training set includes a set of compounds formed by the steroids: **2–4, 6, 8–11, 13–22, 24, 25, 27, 29–31, 34, 36–39, and 41–45** (n = 33, see Fig. 1 and Table 1). The statistical outliers: 2α,3α-epoxy-17β-hydroxy-5α-androstane (**32**), 2β,3β-epoxy-17β-hydroxy-5α-androstane (**33**) and 4,5α-dihydro-Δ^{14} testosterone (**40**), (n = 3, see Fig. 1) were removed from the database. Outlier detection was carried out using the following standard statistical tests: residual, standardized residual, Studentized residual, and Cooks distance. The QSAR models obtained for the A/A ratio apparently do not describe the stereo-electronic effect of these molecules.

The MDs selected by the genetic algorithm are shown in Eqs. 4 and 5:

$$\log((1/\mathrm{LA})/\log(1/\mathrm{VP})) = 29.56(\pm5.28)q15 - 0.19(\pm0.09)\eta + 4.73(\pm0.60) \quad (4)$$

n = 33, R^2 = 0.57, q^2 = 0.25, s = 0.27, F = 20.70, p < 0.001

Test set: n = 9, R^2 = 0.53, s = 0.12, F = 7.00.

$$\log((1/\mathrm{LA})/\log(1/\mathrm{SV})) = 24.71(\pm3.89)q15 - 0.26(\pm0.06)\eta + 4.69(\pm0.44) \quad (5)$$

n = 33, R^2 = 0.68, q^2 = 0.51, s = 0.20, F = 33.99, p < 0.001

Test set: n = 9, R^2 = 0.74, s = 0.12, F = 20.14.

The best regression QSAR model for either A/A ratio was obtained by Eq. 5. The R^2 indicates that the model explains 68% of the variance for the experimental values of log((1/LA)/log(1/SV)) ratio and this model has an adequate q^2 of 0.51. This value of $q^2 > 0.5$ can be considered as a proof of the high predictive ability of the model, the same as the good prediction of the test set (R^2 = 0.74). On the other hand, Eq. 4 depicted an adequate fitness (R^2 = 0.57) but very low predictive power (q^2 = 0.25). This value of $q^2 < 0.5$ can be considered as a proof of the low predictive ability of the model, the same as the bad prediction of the test set (R^2 = 0.53).

Table 7 shows the observed and calculated logLA/logSV ratio values as well as residuals from the best regression model (Eq. 5). The model predicts that the 2-aminomethylene-17α-methyl-17β-hydroxy-5α-androstan-3-one (**43**) compound is the most

Table 1
The anabolic and androgenic activities for testosterone and dihydrotestosterone derivatives

Compound[a]	log(1/LA)	log(1/VP)	log(1/SV)
1. Testosterone	1.56	1.45	1.70
2. 17α-Methyl-testosterone	2.06	2.01	1.97
3. 11β-Hydroxy-testosterone	1.60	1.54	1.52
4. 4-Chloro-11β-hydroxy-testosterone acetate	1.83	1.15	1.15
5. 4-Hydroxy-testosterone acetate	1.72	1.45	1.38
6. 11β-Hydroxy-17α-methyl-testosterone	1.60	1.52	1.54
7. 7α,17α-Dimethyl-testosterone	2.16	2.23	2.32
8. 7α-Methyl-testosterone	2.04	1.83	1.56
9. 7α-Methyl-Δ^4-androsten-3,17-dione	2.35	1.79	1.89
10. Testosterone 17-dichloro-acetate	2.54	2.54	2.54
11. 2α-Fluoro-testosterone	1.70	1.30	1.30
12. Androst-4-ene-3,11,17-trione [adrenoterone]	1.85	1.68	1.68
13. 6β-Fluoro-testosterone	1.48	1.48	1.48
14. Testosterone 17-fluorochloro-acetate	2.49	2.37	2.56
15. Testosterone 17-acetate	1.87	1.99	1.97
16. 2,17α-Dimethyl 17β-hydroxy-androsta-1,4,6-trien-3-one	2.48	2.11	2.06
17. 6α,17α-Dimethyl-testosterone	1.76	1.81	1.81
18. Testosterone 17-trimethyl-silyl ether	2.11	2.12	2.32
19. 4-Chloro-testosterone acetate	2.10	1.74	1.68
20. 17α-Methyl-4,5α-dihydro-testosterone	1.41	1.81	1.72
21. 6β-Methyl-4,5α-dihydro-testosterone	1.86	1.90	1.90
22. 6α-Methyl-4,5α-dihydro-testosterone	1.60	1.36	1.54
23. 2α,17α-Dimethyl-4,5α-dihydro-testosterone	2.30	1.70	1.70
24. 2α,6α,17α-Trimethyl-4,5α-dihydro-testosterone propionate	2.30	1.70	1.70
25. 6α,17α-Dimethyl-17β-hydroxy-5α-androstane-3-one	1.70	1.48	1.40
26. 2α-Methoxymethyl-17β-hydroxy-5α-androstan-3-one	1.48	1.00	1.00
27. 2α-Fluoro-17β-hydroxy-5α-androstan-3-one	1.70	1.30	1.30
28. 2α-Methyl-17β-hydroxy-5α-androstan-3-one	1.79	1.38	1.41
29. 2α,3α-Difluoro-methylene-17α-methyl-5α-androstan-17β-ol	1.45	0.70	1.04
30. 2β,3β-Difluoro-methylene-5α-androstan-17β-ol acetate	1.92	1.53	1.62
31. 17α-Methyl-2β, 17β-Dihydroxy-5α-androstane	0.18	−0.30	−0.40
32. 2α,3α-Epoxy-17β-hydroxy-5α-androstane	−0.10	−0.30	−0.30
33. 2β,3β-Epoxy-17β-hydroxy-5α-androstane	0.83	0.11	0.11
34. 5α-Androstane-17β-ol-3-one (1′-methoxy)cyclo-pentyl ether	2.49	2.09	2.43
35. 5α-Androstane-17β-ol-3-one 17-(1′-ethoxy)cyclopentylether	2.48	2.11	2.43
36. 17β-Hydroxy-1α,17α-dimethyl-5α-androstan-3-one-oxime	2.23	1.85	1.81
37. 4,5 α-Dihydro-testosterone	2.03	2.08	2.09
38. 2α-Methyl-17β-propionoxy-5α-androstan-3-one	2.30	1.70	1.70
39. 3α,17β-Dihydroxy-5α-androstane	2.04	2.27	2.13
40. 4,5α-Dihydro-Δ^{14} testosterone	1.74	1.92	1.45
41. 2α,3α-Difluoro-methylene-5α-androstan-17β-ol-acetate	2.09	1.43	1.64
42. 2α,17β-Dimethyl-17β-hydroxy-5α-androst-9 (11)-en-3-one	2.18	1.79	1.81
43. 2-Aminomethylene-17α-methyl-17β-hydroxy-5α-androstan-3-one	2.20	1.30	1.30
44. 2[2′-(*N*,*N*-Dimethyl-amino)ethylamino-methylene]-17α-methyl-5α-androstan-17β-ol-3-one	2.18	1.48	1.48
45. 17β-Hydroxy-5α-androstan-3-one semicarbazone	1.88	1.81	1.72

[a] Structure of compound give in Figure 1.

potent anabolic steroid, by contrast, the 17α-methyl-2β,17β-dihydroxy-5α- androstane (**31**) compound is the least potent one of this series.

2.3. Driving forces for biological activities of testosterone and dihydrotestosterone derivatives

Interrelations of MDs make difficult the interpretation of the QSAR model. Therefore, it is well known that the interrelatedness among the different MDs results in highly unstable regression coefficients, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficient in a model.[17] However, in some cases strongly interrelated descriptors can enhance the quality of a model, because the small fraction of a descriptor that is not reproduced by its strongly interrelated pair can provide positive contributions to the modeling. On the other hand, the coefficients of the QSAR model based on orthogonal descriptors are stable to the inclusion of novel descriptors, which permit to interpret the regression coefficients and to evaluate the role of individual molecular fingerprints in the QSAR model. Calculated quantum and physicochemical molecular descriptors were subjected to an intercorrelation study (see Table 3). Correlation between variables included in each QSAR model was rather low, indicating the different information content of each term in these equations.

The log *P* is a hydrophobic descriptor related to the pharmacokinetics (mostly due to transfer features across biological membranes) and to the non-covalent interaction origins (van der Waals and hydrophobic effect) of the biological response.[18]

It was found that stability of molecules is related to hardness (η).[19] Harness is defined as

$$\eta = 1/2^{*}(E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}) \quad (6)$$

where, E_{HOMO} and E_{LUMO} are the energies of the highest occupied and lowest unoccupied molecular orbital, respectively.

Models for anabolic and androgenic activity description (Eqs. 1–3) explains the steroid transport and the steroid-receptor interaction. It is mostly due to the biological activities expressed by the log *P* hydrophobic descriptor (describing the pharmacokinetics of the series) and electronic descriptor (η). The log *P* of molecules has positive contribution to the anabolic and androgenic activities and negative η term indicates those high values of η producing unfavorable anabolic and androgenic effects.

Table 2
Quantum and physicochemical parameters values included in the QSAR models of 45 steroids in database

Compound[a]	log*P*	η	q_{15}
1	2.53	5.03	−0.09
2	3.91	5.03	−0.09
3	2.85	5.05	−0.09
4	2.88	4.60	−0.09
5	3.06	4.67	−0.09
6	2.93	5.05	−0.09
7	4.24	5.02	−0.09
8	4.17	5.02	−0.09
9	4.69	5.02	−0.09
10	5.04	4.96	−0.09
11	3.86	5.02	−0.09
12	3.75	5.05	−0.09
13	3.53	5.11	−0.09
14	4.70	4.97	−0.09
15	4.20	4.94	−0.10
16	4.01	4.39	−0.09
17	4.24	5.03	−0.09
18	3.95	4.88	−0.09
19	3.87	4.58	−0.09
20	4.01	5.65	−0.08
21	4.26	5.64	−0.08
22	4.26	5.64	−0.08
23	5.66	5.62	−0.08
24	5.99	5.62	−0.09
25	4.34	5.64	−0.08
26	3.77	5.64	−0.08
27	3.96	5.60	−0.08
28	4.50	5.62	−0.08
29	4.72	6.01	−0.08
30	4.77	5.99	−0.08
31	3.80	6.77	−0.13
32	3.50	6.58	−0.08
33	4.50	6.59	−0.08
34	5.68	5.65	−0.09
35	6.02	5.65	−0.08
36	5.00	5.41	−0.08
37	3.93	5.65	−0.08
38	5.59	5.63	−0.09
39	5.49	2.27	−0.09
40	3.50	5.20	−0.15
41	4.77	5.90	−0.09
42	4.14	5.22	−0.08
43	2.80	4.51	−0.08
44	3.19	4.50	−0.09
45	3.45	4.79	−0.09

[a] Number of compound give in Table 1.

Table 3
Correlation between quantum and physicochemical molecular descriptors included in the QSAR models

	log*P*	η	q_{15}
log*P*	1	0.15	0.37
η		1	0.02
q_{15}			1

Finally, Eq. (5) shows that the anabolic/androgenic ratio increases with the increase in the positive charge of atom 15 (ring C). Again, the negative η suggests that the selectivity lowed with the increase in the values of this descriptor. The importance of high values in the positive charge on atom C-15 in the steroid molecule, as evidenced from this study, corroborates that the 2-aminomethylene-17α-methyl-17β-hydroxy-5α-androstan-3-one (**43**) compound is the most potent anabolic steroid, and the 17α-methyl-2β, 17β-dihydroxy-5α-androstane (**31**) compound is the least potent one of this series.

In general, Eqs.1–3 and 5 signify the importance of log *P* showed positive contribution in all QSAR models, which implies that the

Table 4
Experimental and calculated values for the anabolic potency of the compounds given in Table 1 and Figure 1

Steroid[a]	log(1/LA)[b] Obsd.	log(1/LA)[c] Calcd.	Residual[d]
1*	1.56	1.24	0.32
2*	2.06	1.96	0.10
3	1.60	1.39	0.22
4	1.83	1.79	0.04
5*	1.72	1.82	−0.11
6	1.60	1.43	0.17
7*	2.16	2.14	0.02
8	2.04	2.10	−0.06
9	2.35	2.37	−0.02
10*	2.54	2.60	−0.06
11	1.70	1.94	−0.24
12	1.85	1.86	−0.01
13	1.48	1.69	−0.21
14	2.49	2.42	0.07
15	1.87	2.19	−0.32
16	2.48	2.56	−0.09
17	1.76	2.13	−0.37
18	2.11	2.11	0.00
19	2.10	2.32	−0.22
20	1.41	1.47	−0.06
21	1.86	1.61	0.25
22*	1.60	1.61	−0.01
23	2.30	2.35	−0.05
24	2.30	2.53	−0.23
25*	1.70	1.65	0.04
26	1.48	1.36	0.12
27	1.70	1.49	0.21
28	1.79	1.75	0.04
29	1.45	1.53	−0.08
30	1.92	1.57	0.35
31	0.18	0.39	−0.21
32	−0.10	0.40	−0.49
33	0.83	0.91	−0.08
34*	2.49	2.34	0.15
35*	2.48	2.52	−0.04
36	2.23	2.19	0.04
37	2.03	1.43	0.61
38	2.30	2.31	−0.01
39	0.90	1.23	−0.33
40	1.74	1.60	0.15
41	2.09	1.65	0.43
42	2.18	1.91	0.27
43	2.20	1.84	0.37
44	2.18	2.05	0.13
45	1.88	1.93	−0.05

[a] Number of compounds given in Table 1 and Figure 1. Chemicals marked with asterisk in test set.
[b] Experimental values of the effective dose in levator ani muscle test.
[c] Values calculated by Eq. 1.
[d] Observed minus calculated values.

binding affinity increases with the increase in hydrophobic features of compounds until reaching a critical value after which the affinity decreases. On the other hand, the negative η in the models suggests that the anabolic and androgenic activities are lowed with the increase in the values of this descriptor. The importance of the positive charge of atom 15 (ring C) implies a possible involvement of steroid electronic properties with the binding site in biological membranes.

2.3.1. Interpretation with a bit wider scope: brief note on the domain of applicability of the model

A crucial problem in chemometric and QSAR studies is the definition of the applicability domain (AD) of a classification or regression model. 'Not even a robust, significant, and validated QSAR model can be expected to reliably predict the modeled property for the entire universe of chemicals. In fact, only the predictions for chemicals falling within this domain can be considered reliable and not model extrapolations'.[20] The AD is a theoretical region in

Table 5
Experimental and calculated values for the androgenic potency of the compounds given in Table 1 and Figure 1

Steroid[a]	log(1/VP)[b] Obsd.	log(1/VP)[c] Calcd.	Residual[d]
1*	1.45	1.12	0.33
2*	2.01	1.70	0.32
3	1.54	1.23	0.31
4	1.15	1.65	−0.50
5*	1.45	1.66	−0.21
6	1.52	1.26	0.25
7*	2.23	1.85	0.39
8	1.83	1.82	0.02
9	1.79	2.03	−0.24
10*	2.54	2.23	0.31
11	1.30	1.68	−0.38
12	1.68	1.61	0.07
13	1.48	1.46	0.02
14	2.37	2.08	0.28
15	1.99	1.90	0.09
16	2.11	2.31	−0.20
17	1.81	1.84	−0.02
18	2.12	1.85	0.28
19	1.74	2.08	−0.34
20	1.81	1.18	0.62
21	1.90	1.30	0.61
22*	1.36	1.30	0.06
23	1.70	1.90	−0.20
24	1.70	2.04	−0.34
25*	1.48	1.33	0.15
26	1.00	1.09	−0.09
27	1.30	1.21	0.09
28	1.38	1.41	−0.03
29	0.70	1.15	−0.45
30	1.53	1.20	0.33
31	−0.30	0.08	−0.38
32	−0.30	0.13	−0.43
33	0.11	0.54	−0.43
34*	2.09	1.89	0.20
35*	2.11	2.03	0.09
36	1.85	1.81	0.04
37	2.08	1.15	0.93
38	1.70	1.87	−0.17
39	1.11	0.94	0.18
40	1.92	1.37	0.55
41	1.43	1.28	0.15
42	1.79	1.62	0.17
42	1.30	1.70	−0.40
44	1.48	1.87	−0.40
45	1.81	1.72	0.08

[a] Number of compounds given in Table 1 and Figure 1. Chemicals marked with asterisk in test set.
[b] Experimental values of the effective dose in ventral prostate test.
[c] Values calculated by Eq. 2.
[d] Observed minus calculated values.

Table 6
Experimental and calculated values for the androgenic potency of the compounds given in Table 1 and Figure 1

Steroid[a]	log(1/SV)[b] Obsd.	log(1/SV)[c] Calcd.	Residual[d]
1*	1.70	1.01	0.69
2*	1.97	1.69	0.28
3	1.52	1.15	0.37
4	1.15	1.57	−0.42
5*	1.38	1.59	−0.21
6	1.54	1.19	0.35
7*	2.32	1.86	0.46
8	1.56	1.83	−0.27
9	1.89	2.08	−0.19
10*	2.54	2.30	0.24
11	1.30	1.67	−0.37
12	1.68	1.59	0.09
13	1.48	1.43	0.05
14	2.56	2.13	0.42
15	1.97	1.91	0.06
16	2.06	2.31	−0.25
17	1.81	1.85	−0.04
18	2.32	1.85	0.47
19	1.68	2.07	−0.39
20	1.72	1.18	0.53
21	1.90	1.32	0.59
22*	1.54	1.32	0.23
23	1.70	2.02	−0.32
24	1.70	2.18	−0.48
25*	1.40	1.36	0.04
26	1.00	1.08	−0.08
27	1.30	1.20	0.10
28	1.41	1.45	−0.03
29	1.04	1.20	−0.16
30	1.62	1.25	0.37
31	−0.40	0.07	−0.47
32	−0.30	0.09	−0.39
33	0.11	0.58	−0.46
34*	2.43	2.00	0.43
35*	2.43	2.17	0.26
36	1.81	1.88	−0.07
37	2.09	1.14	0.95
38	1.70	1.98	−0.28
39	1.11	0.94	0.18
40	1.45	1.33	0.12
41	1.64	1.33	0.31
42	1.81	1.63	0.18
42	1.30	1.62	−0.32
44	1.48	1.82	−0.34
45	1.72	1.68	0.04

[a] Number of compounds given in Table 1 and Figure 1. Chemicals marked with asterisk in test set.
[b] Experimental values of the effective dose in seminal vesicle test.
[c] Values calculated by Eq. 3.
[d] Observed minus calculated values.

chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors. That is to say, AD of the QSAR model is 'the range within which it *tolerates* a new molecule'.[21]

It is generally acknowledged that QSARs are valid only within the same domain for which they were developed. In fact, even if the models are developed on the same chemicals, the AD for new chemicals can differ from model to model, depending on the specific descriptors. However, model validation is sometimes neglected, and the application domain is not always well defined.[22]

The purpose of this section is to outline how validation and domain definition determines in which situation it is correct to use the model. The aim of the present work was to develop a model for predicting *A*/*A* ratios of steroids at early stages of drug discovery and development. In consonance, we selected only testosterone and dihydrotestosterone analogues. Consequently, one may not pretend to extrapolate the use of these models to other kinds of class-steroids making uncertain predictions in conditions very different to those fixed to derive the model.[23] It is important to note that in multiple predictor models, simple single-variable range checks are not sufficient to verify AD. At present, there are several approaches to evaluate the DA of QSAR models. For RLM, a multiple predictor problem with normally distributed data, the distance-based measures, like *leverage* is one of most used. Through the leverage approach[24] it is possible to verify whether a new chemical will lie within the structural model domain. The leverage h[25] of a compound measures its influence on the model. That is, leverage used as a *quantitative measure* of the model AD is suitable for evaluating the degree of extrapolation, which represents a sort of compound 'distance' from the model experimental space. Leverage values can be calculated for both training compounds and new compounds. In the first case, they are useful for finding training compounds that influence model parameters to a marked extent, resulting in an unstable model. In the second case, they are useful for checking the applicability domain of the model.[20,21] The *warn-*

Table 7
log(1/LA)/log(1/SV) ratio: observed (Obs.), predicted (Pred.) values and residual from Eq. 5

Compound[a]	log(1/LA)/log(1/SV) Exp.[b]	log(1/LA)/log(1/SV) Calcd.[c]	Residual[d]
1*	0.92	1.09	−0.17
2	1.04	1.08	−0.03
3	1.06	1.21	−0.15
4	1.60	1.29	0.30
5*	1.24	1.27	−0.03
6	1.04	1.19	−0.15
7*	0.93	1.19	−0.26
8	1.31	1.20	0.11
9	1.24	1.12	0.13
10	1.00	1.21	−0.21
11	1.31	1.21	0.09
12*	1.10	1.16	−0.06
13	1.00	1.16	−0.16
14	0.98	1.11	−0.13
15	0.95	0.98	−0.03
16	1.20	1.38	−0.18
17	0.97	1.19	−0.22
18	0.91	1.16	−0.25
19	1.25	1.29	−0.05
20	0.82	1.23	−0.40
21	0.98	1.23	−0.25
22	1.04	1.23	−0.19
23*	1.35	1.23	0.12
24	1.35	1.07	0.29
25	1.22	1.23	−0.01
26*	1.48	1.23	0.25
27	1.31	1.24	0.07
28*	1.27	1.23	0.03
29	1.39	1.12	0.27
30	1.19	1.13	0.05
31	−0.44	−0.28	−0.16
32[e]	0.32	0.96	−0.64
33[e]	7.25	0.96	6.29
34	1.02	0.99	0.04
35*	1.02	1.23	−0.21
36	1.23	1.29	−0.06
37	0.97	1.23	−0.25
38	1.35	1.04	0.31
39	0.81	0.91	−0.10
40[e]	1.20	−0.32	1.52
41	1.27	0.92	0.35
42	1.21	1.35	−0.14
43	1.69	1.55	0.15
44	1.47	1.31	0.16
45	1.09	1.27	−0.18

[a] Number of compounds given in Table 1 and Figure 1. Chemicals marked with asterisk in test set.
[b] Experimental values of the effective log(1/LA)/log(1/SV) ratio.
[c] Values calculated by Eq. 5.
[d] Observed minus calculated values.
[e] Statistical outliers.

ing leverage, h^*, is a *critical value* or *cut-off* to consider the prediction made for the model for specific compounds in data set. The leverage h^* can be defined as $3x\ p'/n$, where n is the number of training chemicals and p' is the number of model parameters plus one.[20,21] Prediction should be considered unreliable for compounds of high leverage value ($h > h^*$). A leverage greater than the warning leverage h^* means that the compound-predicted response can be extrapolated from the model, and therefore, the predicted value must be used with great care. Only predicted data for chemicals belonging to the chemical domain of the training set should be proposed. However, this fact can be see for two points of view taking into consideration the set of compounds evaluated. For example, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals (good leverage). Conversely, a high leverage chemical in the test set is structurally distant from the training chemicals (bad leverage), thus it can be considered outside the AD of the model.

To visualize the AD of a QSAR model, a double ordinate Cartesian plot of cross-validated residuals (first ordinate), standard residuals (second ordinate), and leverage (Hat diagonal: abscissa) values (h) defined the domain of applicability of the model as a squared area within ±2 band for residuals and a leverage threshold of $h = 0.16$ for androgenic activity (Eq. 2). This plot, the so-called Williams scheme can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with CV standardized residuals greater than two standard deviation units, $>2\sigma$) and structurally influential chemicals in a model ($h > h^*$). For instance, Figure 4 shows the Williams plot of Eq. (2) (for describing androgenic activity of steroids included in this study) as an example. As can be noted in Figure. 4, almost all steroids used lie within this area. Actually, some chemicals like **40** and **41** have leverage higher than the threshold but show jack-knifed residuals and standard residuals within the limits. That is to say, newer steroids were wrongly predicted ($>2\sigma$); it is any chemical completely outside the AD of the model, as defined by the Hat vertical line (high h leverage value). Thus, there do not exist any compounds that are both a response outlier and a high leverage chemical. Two other chemicals, **42** and **44**, (squares at 0.16 h) slightly exceed the critical hat value (vertical line) but are very close to other chemicals of the training set, slightly influential in the model development: the predictions for new compounds in this tense situation (for instance, included in a external test set) can be considered as reliable as those of the training chemicals and the possible erroneous prediction could probably be attributed to wrong experimental data rather than to molecular structure. In closing, no apparent outliers were detected and the model can be used with high accuracy in this applicability domain.[23,24]

3. Conclusions

In the present report, predictive QSAR models for biological activity of the testosterone and dihydrotestosterone steroid family were obtained by a multiple linear regression analysis. The employed MDs were quantum-calculated, as well as physicochemical properties, in relation to anabolic and androgenic activities. Genetic algorithms were used as a variable selection method. The developed QSAR models allow the identification, selection and future design of new steroid molecules with increased anabolic activities. MDs included in the reported models allow the structural interpretation of the biological process, evidencing the main role of the shape of molecules, its hydrophobicity and its electronic properties too.

The selected QSAR equation for anabolic and androgenic activities explains the steroid transport and the steroid-receptor interaction. It is mostly due to the biological activities expressed by the log P hydrophobic descriptor (describing the pharmacokinetics of the series) and electronic descriptor (η). The log P of molecules has positive contribution to the anabolic and androgenic activities and negative η term indicates those high values of η producing unfavorable anabolic and androgenic effects.

The model of anabolic/androgenic ratio (expressed by weight of levator ani muscle, LA, and seminal vesicle, SV, in mice) predicts that the 2-aminomethylene-17α-methyl-17β-hydroxy-5α-androstan-3-one (**43**) compound is the most potent anabolic steroid, and the 17α-methyl-2β,17β-dihydroxy-5α-androstane (**31**) compound is the least potent one of this series.

The CA was also applied to this set of steroid molecules, and was found to give much extra information on the clustering of the compounds. This information is not readily visible from the original MDs, enhancing the information contained in the dendrogram. The models described in this study are an alternative for the discovery and optimization of leading anabolic compounds.

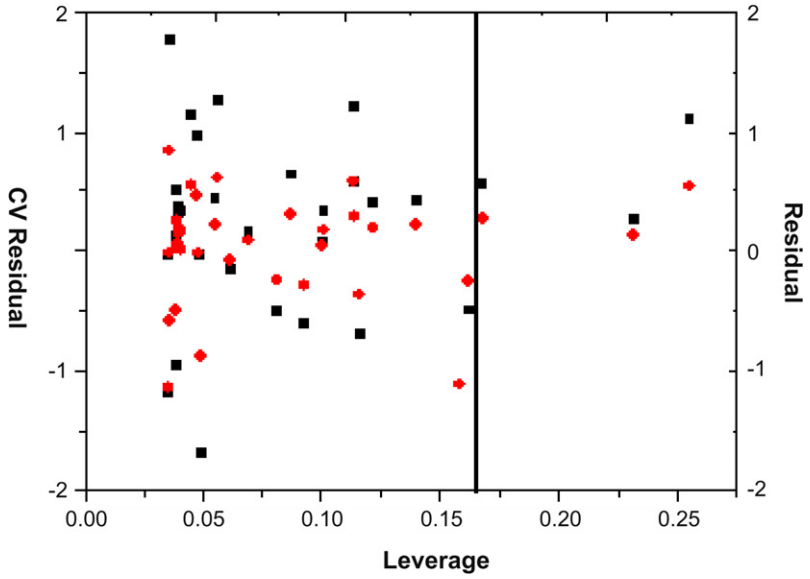

Figure 4. William plot of Eq. 2: outlier will be chemicals are points with jack-knifed (CV standardized) residuals greater than two standard deviation units; influential chemicals are points with high leverage values higher than the *threshold* or *cut-off* value $h^* = 0.16$.

4. Methods

4.1. Data set for QSAR studies

In the present work, the inverse logarithm of the biological activity was used in order to establish classical QSAR correlation equations. A data set of 45 steroids with anabolic and androgenic activities determined in vivo was taken from the literature.[10] At present, there is no other standardized reference where the values of the anabolic and androgenic activities are reported for these kinds of molecules. In the determination of the anabolic and androgenic activities, researchers isolated three organs from each rat: the seminal vesicle (SV), the ventral prostate (VP), and the levator ani muscle (LA).

These organs were all weighted and a comparison between the active groups and the placebo groups was made. The differences in weight of the seminal vesicles and the ventral prostates represent the androgenic activity, while the difference in weight of the levator ani muscle in the control and active group represents the anabolic activity. The experimental values of these biological activities and molecular structures for all steroids are shown in Table 1 and Figure 1, respectively.

The data set was divided randomly into three training sets ($n = 36$, 80% of the data): (1) steroids with anabolic activity, expressed by log(1/LA), (2) steroids with androgenic activity, expressed by log(1/VP) and by log(1/SV), and (3) anabolic/androgenic (A/A) ratios, expressed by [log(1/LA)/log(1/VP)] and [log(1/LA)/log(1/SV)]. The predictive ability of each model was then evaluated by test sets including the remaining steroids ($n = 9$, 20% of data): (1) test set for anabolic activity, (2) test set for androgenic activity, and (3) test set for A/A ratios. The A/A experimental ratios were quantified using the anabolic and both androgenic activity values shown in Table 1. A group of statistical outliers was removed from the data set as it will be discussed below.

4.2. Molecular descriptors for QSAR analysis

A large number of MDs are usually used in QSAR methods.[26,27] The specific biological action of drugs is frequently described by hydrophobic, electronic, and steric properties. The hydrophobic properties express the ability of a molecule to be transported through the organism in order to interact with biological membranes and to be bound to the receptor by van der Waals forces.[28–30] We considered as hydrophobic descriptor the logarithm of the octanol–water partition coefficient (log P).[31,32] Electronic and steric properties characterize the pharmacodynamic properties in the ligand–receptor interaction. They define the ability of the drug to join the receptor.[29] Calculated electronic descriptors by quantum mechanical procedures were: (1) hydration energy (E_{H_2O}),[33] (2) polarizability (P),[34] (3) dipole moment (μ), (4) electronic energy (E), (5) total energy (E_T), (6) HOMO (highest occupied molecular orbital) eigenvalue, (7) LUMO (lowest unoccupied molecular orbital) eigenvalue, (8) net atomic charges of C atoms 1–17 in the steroid backbone (q_1 to q_{19}),[35] electrophilicity index (ω), chemical hardness (η), and softness (S).[36] Electronic descriptors were calculated with MOPAC 6 software[37] using the parametric method 3 (PM3) semi-empirical Hamiltonian[37] after the full geometrical optimization of each molecule. Steric properties were: (1) approximate surface area (ASA), (2) grid surface area (GSA) (calculated by two methods: a fast approximate method and a slower grid-based method), (3) molar volume (VM) (calculated by bounded van der Waals or solvent-accessible surfaces, using a grid method),[37,38] and (4) molar refractivity (MR).[39]

The MDs calculated in the present work and that were included in QSAR models are given in Table 2. Correlations among physicochemical parameters are listed in Table 3.

4.3. Chemometric analysis

All hydrophobic, electronic, and steric properties were used as MDs for derived QSARs. One of the difficulties with the large number of MDs is deciding which ones will provide the best regressions. Furthermore, as testing a large number of all possible combinations of variables would be a tedious task and time-consuming procedure, we have used an input selection by genetic algorithm (GA).[40,41] GA is a class of algorithms inspired by the process of natural evolution in which species having a high fitness under some conditions can prevail and survive the next generation; the best species can be adapted by cross-over and/or mutation in the search for better individuals. Therefore, a GA is a metaheuristic

method for the optimization of functions.[42] A population of potential solutions is refined iteratively by employing a strategy inspired by the Darwinist method (natural selection). The selection method from a population of potential solutions, with preference to the 'fittest' individuals, has given this type of algorithm the name 'genetic', or some times 'evolutionary'. The individuals in a population are often called 'chromosomes', which one built out of 'genes' that represent the properties of the individual, and the function to optimize is referred to as a 'fitness' function. Each iteration is called a 'generation'. In the case of feature selection, for instance, a chromosome is made by a very high number of genes (as many as the variables) each of them being just 1 bit long (0, variable absent; 1, variable present).[43]

The BuildQSAR[44] software was employed to perform variable selection and QSAR modeling. The mutation probability was specified as 35%. The length of the equations was set for three or four terms (according to the models sought-after) and a constant. The population size was established as 100. The GA with an initial population size of 100 rapidly converged (200 generations) and reached an optimal QSAR model in a reasonable number of GA generations. The search for the best model can be processed in terms of the highest correlation coefficient (R) or F-test (Fisher-ratio's p-level $p(F)$) equations, and the lowest standard deviation (s) equations.[45] The quality of models was also determined by examining the leave-one-out (LOO) cross-validation (CV) (q^2). Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of a particular QSAR model.[46,47] Nevertheless, in a recent paper Golbraikh and Tropsha demonstrated that high values of q^2 appear to be a necessary but not sufficient condition for the model to have a high predictive power.[48] Therefore, in addition to this statistical value, we also used an external prediction test set. This type of model validation is very important, if we take into consideration that the predictive ability of a QSAR model can be estimated using only an external test set of compounds (in the model range), which was not used for building the model itself.[49]

4.4. Clustering

Cluster analysis (CA) encompasses a number of different classification algorithms and it permits to organize the observed data into meaningful structures. Conceptually, the approach used by CA in order to address this problem can be described well by the saying 'birds of a feather flock together'.[50] Many CA algorithms have been invented and they belong to two categories: hierarchical clustering and partitional (non-hierarchical) clustering. Hierarchical clustering rearranges objects in a binary tree-structure (joining clustering) and these methods are implemented in an either agglomerative (bottom-up) or divisive (top-down) procedure. On the other hand, the partitional clustering assumes that the objects have non-hierarchical characters.[51,52]

Most popular partitional cluster algorithms are k-mean cluster algorithms (k-MCA) and Jarvis–Patrick (also known as k-nearest neighbor cluster algorithm; k-NNCA) algorithms. k-Mean clustering algorithms use an interchange (or switching) method to divide n data points into k groups (clusters) so that the sum of distances/dissimilarities among the objects within the same cluster is minimized. The k-mean approach requires that k (the number of clusters) is known before clustering. The Jarvis–Patrick method requires the user specifies the number of nearest neighbors, and the number of neighbors in common to merge two objects. Jarvis–Patrick method is a deterministic algorithm; it does not require iterations for computations.[51,50]

In order to design training and test series and to demonstrate the structural diversity of the present database, we carried out one of these kinds of cluster analyses (k-NNCA) for steroid series. The STATISTICA VER. 5.5, software package[53] was used to develop these CA.

In this study, we used the 'average linkage' metric as the method to merge objects into clusters. The average linkage distance between two clusters is defined as the average (squared Euclidean) distance between pairs of objects, one in each cluster. Average linkage tends to join clusters with small variances and produces clusters with roughly the same variance.

Acknowledgments

This research was supported by the Center for Pharmaceutical Chemistry (CQF), Cuba and the Faculty of Chemistry, Universidad de La Habana, and computational facilities were provided by Deutscher Akademischer Austauschdienst (DAAD) in Bonn, Germany. The Universidad Autónoma de Madrid—Universidad de La Habana program under the auspices of CajaMadrid, Spain, also supported part of this work. One of the authors (M.-P.Y.) thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University (2008). Finally, but very importantly, M.-P.Y. thanks the Flemish Interuniversity Council (VLIR) of Belgium for partial support of this research through a part of the fund of the project '*Strengthening postgraduate education and research in Pharmaceutical Sciences*'. Anonymous reviewers are gratefully acknowledged for their helpful suggestions that have led to improving the paper.

References and notes

1. Llewellyn, W. *Anabolics 2004*. # 22-308 Júpiter, FL 33458, 2004; pp 3–10.
2. Hengge, U. R.; Baumann, M.; Maleba, R.; Brockmeyer, N. H.; Goos, M. *Br. J. Nutr.* **1996**, *75*, 129–138.
3. Bowers, M. Anabolic steroids in the treatment of HIV-related wasting. *Bull. Exp. Treat. AIDS*. No. 30, 1996.
4. Morales-Polanco, M. R.; Sánchez-Valle, E.; Guerrero-Rivera, S.; Gutiérrez-Alamillo, L.; Delgado-Márquez, B. *Arch. Med. Res. Spring* **1997**, *28*, 85–90.
5. Dunn, J. M. *HIV Hotline* **1998**, *8*, 4–5.
6. Phillips, W. N. *Anabolic Reference Guide*, 1991.
7. Llewellyn, W. *Anabolic 2004*. # 22-308 Júpiter, FL 33458, 2004, pp 17–18.
8. Murad, F.; Haynes, R. C., Jr. *Las Bases Farmacológicas de Terapéutica* **1984**, *3*, 1413–1429.
9. Anabolic Steroids. www.saludhoy.com/html/depor/articulo/esteroi2.html.
10. Vida, J. A. *Androgens and Anabolic Agents*. New York and London, 1969.
11. Alvarez-Ginarte, Y. M.; Crespo Otero, R.; Montero Cabrera, L. A.; Ruiz García, J. A.; Marrero Ponce, Y.; Santana, R.; Pardillo Fontdevila, E.; Alonso Becerra, E. *QSAR Comb. Sci.* **2005**, *24*, 218–226.
12. Alvarez-Ginarte, Y. M.; Crespo Otero, R.; Marrero Ponce, Y.; Montero Cabrera, L. A.; Ruiz García, J. A.; Padrón García, A.; Torrens Zaragoza, F. *QSAR Comb. Sci.* **2006**, *25*, 881–894.
13. Alvarez-Ginarte, Y. M.; Marrero Ponce, Y.; Ruiz García, J. A.; Montero Cabrera, L. A.; García de la Vega, J. M.; Noheda Marin, P.; Crespo Otero, R.; Torrens Zaragoza, F.; García Domenech, R. *J. Comput. Chem.* **2008**, *29*, 317–333.
14. Klein, C. T.; Kaiser, D.; Ecker, G. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 200–209.
15. Bultinck, P.; Carbó-Dorca, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170–177.
16. Restrepo, G.; Villaveces, J. L. *Croat. Chem. Acta* **2005**, *78*, 275–281.
17. Randić, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45–59.
18. Yazdanian, M.; Briggs, K.; Jankovsky, C.; Hawi, A. *J. Pharm. Res.* **1998**, *15*, 1490–1494.
19. Parr, R. G.; Szentpaly, L. V.; Liu, S. *J. Am. Chem. Soc.* **1999**, *1922*, 1922.
20. Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
21. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
22. González-Díaz, H.; Vilar, S.; Santana, L.; Podda, G.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15*, 2544–2550.
23. Papa, E.; Villa, F.; Gramatica, P. *J. Chem. Inf. Model.* **2005**, *45*, 1256.
24. Atkinson, A. C. *Plots, Transformations and Regression*; Clarendon Press: Oxford, 1985.
25. Liu, H.; Papa, E.; Gramatica, P. *Chem. Res. Toxicol.* **2006**, *19*, 1540.
26. Hansch, C. *Comprehensive Medicinal Chemistry*. Oxford, Vol. 3, 1990.
27. Wermuth, C. G. *The Practice of Medicinal Chemistry*. London, 1996.
28. Kubinyi, H. *QSAR: Hansch Analysis and Related Aproaches*. Weinheim (Germany), 1993.
29. Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Application in Chemistry and Biology*. Washington, DC, 1995.
30. Todeschini, R.; Consonin, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H. *Handbook of Molecular Descriptors*. Germany, 2000.
31. Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80–90.
32. Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

33. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.
34. Millar, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.
35. Stewart, J. J. P. *Program MOPAC*, Tokio, 1993–1997.
36. Parthasarathi, R.; Subramanian, V.; Roy, D. R.; Chattaraj, P. K. *Bioorg. Med. Chem.* **2004**, *12*, 5533–5543.
37. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221–264.
38. Bodor, N.; Gabanyi, Z.; Wong, C. J. *J. Am. Chem. Soc.* **1989**, *111*, 3783–3786.
39. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1987**, *27*(1), 21–23.
40. Hasegawa, K.; Kimura, T.; Fumatsu, K. *Quant. Struct.-Act. Relat.* **1999**, *18*, 262–272.
41. So, S. S. K. *J. Med. Chem.* **1996**, *39*, 1521–1530.
42. Mitchell, M., An Introduction to Genetic Algorithms. Cambridge, 1996.
43. Coley, D. A. *Introduction to genetic Algorithms for Scientists and Engineers* **1999**.
44. Barbosa de Oliveira, D.; Gaudio, A. C. *QSAR Comb. Sci.* **2003**, *19*, 599–601.
45. Ford, M. C.; Salt, D. C. *Chemometric Methods Mol. Des.* **1995**, *2*, 283–292.
46. Golbraikh, A.; Tropsha, A. *J. Comp. Aided Mol. Des.* **2002**, *16*, 357.
47. Wold, S.; Sjöström, M.; Eriksson, L. *Statistical Validation of QSAR Results. Validation Tools*. In *Chemometric Methods in Molecular Design*. New York, 1995; pp 309–318.
48. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
49. Marrero, P. Y. *Molecules* **2003**, *8*, 687–726.
50. Xu, J.; Hagler, A. *Molecules* **2002**, *7*, 566–600.
51. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*(4), 269–276.
52. Farland, J. W.; Gans, D. J. *Chemometric Methods Mol. Des.* **1995**, 295–307.
53. Statsoft *STATISTICA*, 1999.